University of Lausanne
Faculty of Social and Political Sciences
MA in Public Opinion and Survey Methodology

Split questionnaire design: can a question battery be split and still produce the same measurements? Evaluating context effects in the study of moral values

- Master thesis -

Presented by: Jimena Sobrino Piazza

Supervisor: Prof. Dr. Caroline Roberts

Expert: Dr. Kenneth Horvath

Winter session 2019

Acknowledgements

This project would not have been possible without the support from the International Survey Team of FORS, to whom I am most grateful. I would like to thank in particular Michèle Ernst-Stähli for giving me the opportunity to join this team, and to Alexandre Pollien for welcoming me into his project on question order effects. A special thanks to Michael Ochsner for the guidance and the great conversations throughout the summer. Thanks also to Patricia Milbert, Jessica Herzing and Marlène Sapin for their inputs and precious time.

A very special thanks to Caroline Roberts for her patience and infinite support. Our long discussions ended always with a boost of motivation.

Thanks to Kenneth Horvath for his help with my analysis.

Thanks to Pierre Gallaz, who gave me so many tools to enjoy the process.

An enormous thanks to my beloved friends and family, who suffered from my monothematic conversations. Thank you, peruchas, for being that pillar, when family is so far away. Thanks Joëlle, Marianne, Emilia for the early messages and the company in the library.

Abstract

Previous questions may impact answers given to later questions. Question order effects and context effects have been studied in past decades. In a moment in which the split questionnaire design (questionnaire modularization) becomes increasingly popular, this study aims to contribute with the refining of splitting strategies, in order to prevent the introduction of context effects. A particular question setting is studied: question batteries in which items related to different constructs are intermixed, sharing a common rating scale. Can a question battery be split and still produce the same measurements, both at the item level and scale level? The question battery on moral beliefs of the Swiss EVS 2017 is studied. This question battery was administered in two versions, as a result of a split questionnaire design. Measurements of whole and split versions are compared. Out of the fifteen items embedded in the battery, four presented significantly different estimates of the mean. When analyzed by subgroups of respondents, differences increased/decreased depending on variables such as the level of religiosity, age, education level, and in a lesser extent, political orientation. At the multi-item level, a multi-group confirmatory factor analysis (MCFA) revealed differences in the factor structures comprising the scalar invariance for one construct and the strict invariance for the other construct. Scale means remained equivalent across question battery versions.

Table of Contents

1. INTRODUCTION

1.1 Problem Formulation

Survey questions are never asked in a vacuum. Among the many factors that may influence the way respondents answer to a given question, the questionnaire content may play an important role. Earlier questions in the questionnaire generate a context in which questions are embedded. So, the meaning of questions, the ideas respondents consider to answer them, the standards of comparison, are influenced by questions previously answered in the questionnaire. Also, previous questions may generate, independently of their content, fatigue to respondents, so that depending on the position a question occupies in the questionnaire, the likelihood of bad response quality may increase. These assertions are far from new in survey methodology literature. Question order effects have been deeply studied in the late 70s, 80s and 90s (e.g. Schuman & Presser, 1981; Sudman, Bradburn, & Schwarz, 1996; Tourangeau, Rips, & Rasinski, 2000), their possible impact on measurements being well-known by methodologists.

Recent developments in the field of survey methodology make necessary to bring the topic of question order and context effects back to discussion. Concretely, the increasing pressure to reduce questionnaire length in order to adapt to new data collection technologies make the topic of context effects of great relevance again. Online surveys are not new. Indeed, surveys in market research have been transitioning to web for almost two decades now, to a great extent motivated by the enormous reduction of costs it represents. More academic surveys, and particularly so the international comparative projects, have kept however a safe distance from online questionnaires. Face-to-face interviews have continued to be the golden standard for the most ambitious general population surveys. However, conducting these traditional one-hour-long interviews is becoming increasingly difficult.

The present epoch has been described by Dillman, Smyth, and Christian (2009) as a turbulent time for survey methodology, these authors explaining how the technological developments from the last decades have resulted in changes in the cultural norms, and how this constitutes a challenge for survey research. As they explain, with the new technology for facilitating communication, came also the means for ignoring these new massive amounts of correspondence. Whereas before people found difficult or even rude to ignore the request from a surveyor asking for collaboration, today's norm is to be suspicious about such requests coming from a

stranger. Another consequence of the new technologies is that interactions are faster, shorter and spontaneous. In consequence, survey questions are now, more than ever, required to be short, fast and interesting, in order to be reconcilable with today's people's rhythm.

The transformation of society, such as described by Dillman et al. (2009), makes understandable why one-hour-long face-to-face interviews conducted by a stranger that knocks at the door, start to look from another time. Unsurprisingly, response rates have been falling, which demands constantly to increase fieldwork efforts, and makes face-to-face interviews more and more expensive. In order to ensure the continuity of projects, reducing costs has become unavoidable. The solution would seem precisely to transition to online mode of data collection. Yet, there is a major methodological challenge for accomplishing this transition: the questionnaire length. The traditional one-hour-long questionnaires must to be adapted to online format, in which a twenty-minutes-long questionnaire is already considered too long. Methodologists have found a solution to implement these surveys, and still collect the same information necessary for comparisons over time: the split questionnaire design or modular questionnaire design. In short, this consists in splitting the original questionnaire into different modules, so that many different and complementary, short versions of the questionnaire are created. Although the split questionnaire design is not a novelty, it has regained interest in the field out of the need of adapting long questionnaires into web surveys.

The use of the split questionnaire design for the methodological transition to web entails that a same question is simultaneously embedded in different questionnaire versions for different respondents, and thus preceded by different questions. If this becomes common practice, a recall of the literature on context effects is more than ever important. It becomes crucial to understand the risks of context effects and the situations in which they are likely to occur, in order to decide how best to split questionnaires. The way questions are allocated to the different modules determines whether the context in which questions are asked may change or not.

The aim of this project is to contribute in the understanding of the consequences of altering questionnaire content on the measurements. The focus of this master thesis is on the functioning of question batteries that share a common rating scale. This question setting is particularly prone to present context effects, because of the intrinsic comparative framework that it constitutes. Furthermore, in this project the aim is to study the implications of splitting a question battery in a specific situation: when items that measure different multi-item constructs are

presented intermixed in the same battery. By studying this particular question setting, the idea is to shed light on the possible context effects that may be introduced if the question battery items are changed, both at the item- and multi-item level of analysis, and to provide insights about how best to split this type of question batteries in the context of questionnaire modularization designs, or if they should be split at all. Beyond the split questionnaire design, the contribution of this project could be helpful for questionnaire design in general.

To study this, a particular question battery about moral values will be analyzed, borrowed from the Swiss European Values Study (EVS) 2017 data. This survey conducted a methodological experiment precisely to shed light on how best to transition from face-to-face interviews to online surveys. In parallel to a traditional face-to-face survey, a self-administered push-to-web survey was conducted, implemented both by means of a split questionnaire design, as well as by full-length questionnaires. As part of the split questionnaire design, the question battery on moral values was split into two sections. In consequence, this battery was administered to respondents in two different versions.

1.2 Research Questions

The general question guiding this study is whether a question battery can be split and still produce equivalent measurements. In order to evaluate the possible effects of splitting a question battery, three research questions will be addressed:

- **RQ1.** What are the effects of splitting a question battery on measurements at the item level?
- **RQ2.** To what extent are any observed effects of splitting at the item level moderated by respondent characteristics?
- **RQ3.** To what extent do effects of splitting a battery at the item level affect multi-item measures embedded in the battery?

2. LITERATURE REVIEW

2.1 Context effects and question-order effects – definition

Dillman, Smyth, and Christian (2014) define *question-order effects* as "measurement error that results from early questions unintentionally influencing answers to later questions" (p. 230). The encyclopedia of survey research methods (Lavrakas, 2008) defines *context effects* as "a process in which prior questions affect responses to later questions in surveys" (p.142). The terms of context effects and of question-order effects might seem at first view as interchangeable, and have indeed been mobilized in the literature as virtual synonymous (e.g. Tourangeau et al., 2000: 200).

In their classification of question-order effects, however, Schuman and Presser (1981) distinguish two type of effects: *context effects* and *sequence effects*. According to them, one may speak of context effects when a transfer of meaning occurs from previous questions into subsequent questions. On the other hand, sequence effects refer to mechanical types of artifacts, as it would be the case of a question asked at the end of the questionnaire, suffering from the consequences of the respondent's fatigue. According to this typology, context effects would be thus a specific kind of question-order effect (Schuman & Presser, 1981: 23).

Schuman (1992), nonetheless, nuances this terminology. He points out to the fact that one may speak of question-order effects only when the order of the questions is controlled. Effects of context, however, have also been identified in self-administered paper questionnaires, in which an item is influenced by questions appearing later in the questionnaire (Schwarz & Hippler, 1995). Indeed, self-administered paper questionnaires give respondents the freedom to look the entire questionnaire before answering the questions in whatever order they prefer. From this perspective, not all context effects would necessarily be at the same time order effects.

The term of context effects, as mobilized in this paper, could be seen as a type of measurement error resulting from the *thematic content* of other questions in the questionnaire. Dillman et al. (2014) explain that this type of effect is likely to occur by the presence of topic-related questions presented close to one another in the questionnaire. The more topic-related, and the closer the position of these questions in the questionnaire, the higher the likelihood of these effects to appear (Dillman et al., 2014: 234). Paradoxically, the way questionnaires are traditionally

built – by regrouping topic-related questions together to imitate the logic of a conversation (Dillman et al., 2014) – is prone to introduce this type of effects.

2.2 Classification of context effects

2.2.1 Part-whole/part-part – assimilation/contrast effects

In this section the classification proposed by Schuman and Presser (1981) on context effects is presented. This classification was provided as a result of a series of split-ballot experiments conducted by Schuman and colleagues during the 70s. These authors propose a typology based on two dimensions: the direction of the effect, and the type of relation between the previous and subsequent questions that originated it. Table 2.1 summarizes this typology. Examples for each scenario are provided thereafter.

Direction of the effect. Schuman and Presser (1981) catalogued context effects depending of the *direction of the effect*. When answers to questions end up being *more similar* than they would have been if the order of questions would have been different, the term of *assimilation effect*[1] is used. And if answers are *more different* than they would have been, one may speak of *contrast effect* (Schuman & Presser, 1981: 27-8).

Type of relation between questions. Schuman and Presser (1981) identified two different scenarios, depending on the *type of relation* between the preceding and the subsequent questions. The first scenario occurs when two or more related questions appear together in a questionnaire, but one of the questions is more general than the others, in such way that it contains or implies the other questions (Schuman & Presser, 1981: 27). Scenarios of this type are named "*part-whole combinations*". Literature on question order effects has found this type of combination of question to be particularly susceptible of presenting effects of context (e.g. Willits & Saltiel, 1995). The other type of situation distinguished by these authors occurs when two or more related questions are asked together, but all questions have a similar level of specificity (Schuman & Presser, 1981: 28). This scenario was referred by them as "*part-part combinations*".

[1] In their book, Schuman and Presser (1981) use the term 'consistency effect' rather than 'assimilation effect'. The term of 'assimilation effect', has been widely mobilized in the literature as an equivalent term to the concept of 'consistency' (e.g. Sudman et al., 1996; Dillman et al., 2014). Here, the term of 'assimilation' is used in order to distinguish this concept from the 'norm of consistency' mentioned in later sections of this paper.

Table 2.1: *Classification of context effects*

Type of relation	Direction of the effect	
	Contrast	**Assimilation**
Part-whole combination	e.g. General/specific abortion legalization	e.g. Life/marriage satisfaction
Part-part combination	e.g. Parents/own political identification	e.g. Communist/ American reporters

Part-whole contrast effects. The effect between the general and specific abortion questions might serve to illustrate this scenario. This question-order effect was first detected when divergent results from two consecutive national surveys for the United States were found on the item of general support for the legalization of abortion. The same question had been used in both surveys. The difference, however, was that in one of the surveys, the general item on abortion was preceded by a more specific question about support of abortion in the case of a defect in the unborn child (Schuman & Presser, 1981: 36). The order effect was then tested by means of split-ballot experiments, in which the two items – specific and general – were presented together in alternating orders. Results were consistent: agreement with the general item was higher when the general item was asked *before* the specific item (Schuman & Presser, 1981: 37-8). Such an effect has also been referred as a "subtraction effect" (Schuman, Presser, & Ludwig, 1981). The interpretation is that, when asked after a topic-related, more specific question, respondents may understand a general question as referring to all other cases *aside from* the specific case previously asked.

Part-whole assimilation effects. However part-whole combinations do not necessarily conduce to such subtractions. For example, Smith (1979) studied another part-whole combination consisting of a question about general life satisfaction, asked together with another, more specific question about satisfaction with marriage. He found that the reported general life satisfaction was higher when the marital satisfaction item was asked first. This order effect was then tested by Schuman and Presser (1981) in a split-ballot experiment in 1980, their findings

supporting the ones of Smith (1979).[2] The interpretation was that the reminder of a happy marriage lead respondents to report higher general life satisfaction.

Part-part contrast effects. The example of the effect reported by Willick and Ashley (1971) about political party identification illustrates a case of a part-part contrast effect. In a survey among college students, respondents were asked to report both their own, as well as their parents' political party identification. It was found that when the question about the parents' political party came first, students were *less likely* to report having the same political party identification themselves. Researchers interpreted this effect as resulting from a necessity of students of reporting their independence from their parents. Schuman and Presser (1981) only found this example of part-part contrast effects, pointing out how rarely this type of scenario may occur.

Part-part assimilation effects. A classic example to illustrate this scenario is the case of the questions about Communists and American reporters, which was first detected by Hyman and Sheatsley (1950). The context effect was found within a split-sample experiment carried out in 1948 on a national survey for the United States. Two questions were asked. One was whether a Communist country like Russia should allow American journalists to go there and report news back to their American newspapers from within the country. The other question, very similar, asked whether the United States should let newspaper reporters from Communist countries enter America. The experiment was then replicated by Schuman and colleagues in 1980. Results were consistent (Schuman & Presser, 1981: 28-31). Depending of which question came first, answers to the items were strikingly different. Indeed, according to Tourangeau, Rips, and Rasinski (2000) this may be the largest context effect ever discovered in survey literature (p. 212). If the Communist reporter question came first, approval for both questions was lower; whereas if the American reporter question was asked first, the approval for both questions was much higher (see Schuman & Presser, 1981: 29). The interpretation of this order-effect was that respondents were applying a norm of reciprocity or evenhandedness (Schuman & Presser, 1981; Tourangeau et al., 2000; Schuman, 2009). So, having answered the first question based on their attitudes towards the topic, when confronted with the second question, respondents would compute an answer on the basis of what they had already responded in the previous question.

[2] Nevertheless, this life/marriage satisfaction combination has been repeatedly studied, and findings have been rather diverse, some authors reporting also subtraction effects (see Schwarz, Strack, & Mai, 1991).

2.2.2 Comparative vs. noncomparative contexts – framework dimension effects

The contribution of Moore (2002) to the study of context effects is presented in this section. This author detected context effects of a different nature than the ones reviewed in the previous section. Given his discovery, he proposes a classification of context effects according to two different dimensions: the item dimension and the framework dimension.

Comparative and noncomparative contexts. In his article, Moore (2002) distinguishes two possible scenarios in which a question can be asked: either within a comparative context, or within a noncomparative context. The noncomparative context implies that a question is answered without influence of any topic-related question. In the comparative context, in contrast, a question is asked after a related question, and thus the answer to it can be influenced by the presence of previous related items. Context effects would arise thus in comparative contexts.

Item dimension effects. According to Moore (2002), the comparative context could induce respondents to evaluate an item in relation to the previous related item. This is what he calls context effects at the item dimension. Assimilation and contrast effects such as presented by Schuman and Presser (1981) – see previous section 2.2.1 – are effects of this type.

Framework dimension effects. Moore (2002) detected however that context effects could also occur as a consequence of the comparative context per se. Effects of this type are referred by this author as framework effects. In his own words, they occur: "when respondents look beyond the two items being evaluated to the larger framework within which the questions are posed" (Moore, 2002: 84).

The example presented by Moore (2002) about people's perceptions of racial hostility may help to understand how a framework effect works. The effect was detected when two questions on perceptions of racial hostility were asked together in alternating order. These questions intended to measure separately, on the one hand, the degree of racial hostility existing among white people towards black people, and, on the other hand, the degree of racial hostility among black people towards white people. Results showed that both items received higher scores when they were asked second. In other words, when asked in the comparative context, items got higher scores than when asked in the noncomparative context. The effect was neither contrast nor assimilation effect, given that the gap between the scores of the two questions remained the same, independently of the order in which the items were presented. The effect was thus referred as an "additive" effect, in which the comparative context increased the score an item would get. The interpretation was

that, after being reminded that racial hostility existed among people of both races, respondents answers to the second question were reflecting their perception of the overall existing level of racial hostility.

The comparative context may thus evoke respondents an underlying concept common to the different items, and make respondents answer according to that underlying concept, rather than to the concrete case each item presents. In the example of the perceptions of racial hostility, the comparative context led to an overall rise of the scores, referred by Moore (2002) as "additive effect". The comparative context may also lead however to an overall decrease of the scores. In that case, Moore (2002) speaks of "subtractive effect".

2.3 Explanations for context effects in attitudinal questions

2.3.1 The answering process of survey questions

Together with efforts to classify question-order effects, an important volume of the literature would follow in the decade of 1980s and 1990s, focused on the understanding of the psychological processes underlying context effects (e.g. Strack & Martin, 1987; Tourangeau & Rasinski, 1988; Strack, 1992). Conceptual models of the processes involved in answering a survey question were originally an extension of the general information processing models developed in cognitive psychology. However, these models would be completed in order to capture specificities of the survey situation (Sudman et al., 1996: 55). As pointed out by Schwarz and Strack (1991), individuals in the survey situation should not be seen as isolated information processors. Surveys are a social interaction, thus the social context had to be taken into account. As a result, authors proposed models that combine both communicative and individual thought processes (Schwarz & Strack, 1991).

Cognitive steps. From the cognitive perspective, the process of answering a survey question is perceived as a number of tasks to be accomplished by the respondent. Tourangeau & Rasinski (1988) summarize the respondent's tasks in four steps: question comprehension, information retrieval, computation of a judgment and selection of a response. Similarly, Strack and Martin (1987) also describe a four-steps model of the response process, but identify the following four steps: first, interpreting the question; second, generating an opinion; third, fitting it into the response format that has been provided; and forth, editing the response. Although not all researchers use the same labels, there is a wide agreement about the different steps involved in the act of answering a question. The tasks required for

answering an attitudinal question[3] would be: interpreting the question, retrieving information about the issue, generating an opinion, fitting it into the response format that has been provided, and editing the response (Sudman et al., 1996).

Communicative logic. The other component of the model takes into account the communicative process in which a survey is inserted. Surveys have been understood as a particular sort of conversation. Because of this, authors have argued, respondents rely on conversational rules when confronted with a questionnaire (e.g. Schwarz, 1999). The rules of conversation described by Grice (1975) have been mobilized in survey methodology to conceptualize how respondents behave in a survey situation. According to Grice (1975), conversations would be guided by fours maxims or tacit assumptions. Schwarz (1999) reviews these four rules and translates them into the survey situation. A *maxim of relation* enjoins individuals to make contributions judged to be relevant to the aims of the ongoing conversation. Translated to the survey situation, this maxim leads respondents to interpret questions within the context demarcated by previous questions. A *maxim of quantity* encourages speakers to make contributions as informative as required, but no more informative than required. This maxim discourages respondents to reiterate themselves, and thus enjoins them to give new information at every question. A *maxim of manner* guide speakers in that contributions should be clear rather than obscure, ambiguous, or wordy. Respondents in a survey will thus assume the most obvious meaning of a question. Finally, a *maxim of quality* enjoins individuals to avoid statements they believe to be false (Schwarz, 1999). And as pointed out by Sudman et al. (1996), these conversational rules that make speakers to be relevant, informative, clear, and truthful enjoins respondents not only to behave that way, but also to assume that the questionnaire is in itself relevant, informative, clear, and truthful too.

The cognitive process and the communicative logic described above have been mobilized to identify the different mechanisms by which previous questions may influence the way respondents answer a subsequent question (Strack & Martin, 1987; Tourangeau & Rasinski, 1988; Schwarz & Strack, 1991; Strack, 1992; Sudman et al., 1996).

[3] Tasks are described slightly different, depending on whether it is an attitudinal question or if it is a behavioral question (Sudman et al., 1996: 57). In this project the focus is only on attitudinal questions, so the description above refers to them.

2.3.2 Interpretation stage: Influencing the literal and practical meaning of questions

Previous questions may influence respondents in their first cognitive step: understanding or interpreting the question (see Strack, 1992 for a detailed discussion). In order to understand the question meaning, respondent require not only to understand the question literally, but also to understand the pragmatic meaning of the question (Strack, 1992; Schwarz, 1999). The understanding of the literal meaning of a question "involves the identification of words, the recall of lexical information from semantic memory, and the construction of a meaning of the utterance, which is constrained by its context" (Sudman et al., 1996: 62). Strack (1992) explains how answering preceding questions may have an activation function that occurs even without the respondent awareness; it is what social psychologists call "priming effect". According to this author, this activation might come either from the literal content of previous questions, or from the cognitive process that those previous questions generated in respondents in order to answer them. Having been activated, this information is more accessible for respondents, which is what influences them at the moment of interpreting the literal meaning of a subsequent question. As stressed by Sudman et al. (1996) the influence may occur in particular when the subsequent question contains ambiguous terms. But this can also occur even if question wording is not ambiguous.

Understanding the literal meaning of a question is often not sufficient for answering it. For example, as explained by Schwarz (1999), if the question asks "What have you done today?", respondents may hesitate whether they should report them taking a shower. For this, according to him, inferring the intended meaning, or pragmatic meaning is also required. This is where the conversational rules may play a role, in particular the maxim of quantity (Tourangeau & Rasinski, 1988).

Going back to the example of the abortion questions, the context effect could be explained thus by the fact that respondents, once having answered the specific item on abortion in the case of a defective unborn child, and relying in the assumption that the questionnaire would not be redundant (maxim of quantity), might interpret the intended meaning of the general question more or less as following: "aside from the circumstances of a defective unborn child, are you supportive of the legalization of abortion?".

2.3.3 Judgment stage: making information more accessible

Influence on the judgment stage. Information activated by previous questions may also influence respondents at the moment of retrieving information

and generating an opinion. Schwarz and Bless (1992) propose a model to conceptualize the underlying processes explaining how previous questions influence the judgment stage of the response answering process, giving rise to assimilation or contrast effects: the inclusion/exclusion model[4]. Once having interpreted the meaning of the question, respondents need to form a mental representation of the *target*, this is of the issue the question is about. In addition, the model postulates, respondents need to form a mental representation of some *standard of comparison*. It is by confronting their representation of the target against the standard of comparison that respondents generate an evaluative judgment for answering the question. Both the representation of the target, as well as the representation of the standard of comparison, can be influenced by previous questions. According to the model, in order to form these mental representations, a quick process of information retrieval takes place. The information that is more accessible to the respondent at the moment of judgment, determines the representation that is formed. Chances are that the first information that comes to mind for the respondent will be chronically accessible information - that is, information that always comes to mind, whenever the respondent thinks about the given topic. It is also possible that the information coming to mind had be activated by the context, this is temporarily accessible information. So, previous questions may influence the mental representations respondents form from the question topic, and/or from the standard of comparison, by making some information more accessible in their minds. Moreover, the inclusion/exclusion model predicts the direction of the context effect – assimilation or contrast – as a function of whether previous questions influenced the mental representation of the target stimulus (question topic), or of the standard of comparison. If information activated by previous questions is used for constructing a representation of the standard of comparison, the model predicts a *contrast effect*. Inversely, if the activated information played a role when forming the representation of the target stimulus, this would result in an *assimilation effect*.

2.3.4 Selection stage: Influence of rating scales and consistency

Influence on the selection stage. Context effects may also occur at the response selection stage, when respondents format their opinion into the response categories available, and edit their responses. Tourangeau and Rasinski (1988) identify two main mechanisms on how previous questions may influence these tasks.

[4] Tourangeau, Rasinski and Bradburn (1992) propose a similar model, the belief sampling model, but which focuses primarily in the emergence of assimilation effects, while the conditions giving rise to contrast effects are not explicitly identified.

The first mechanism they describe consists in previous questions acting as anchors for the response scale, this is, as reference points, or standards for comparisons (p. 311). When asked to rate a series of items on a scale, it might seem that respondents anchor the rating scale on the extremes of the stimulus continuum. This perspective effect is particularly pronounced when stimuli are rated along the same scale (Sudman et al., 1996: 93). Moreover, according to Schwarz and Wyer (1985), context effects at this stage may reduce if all the response categories are labeled.

Schuman and Presser (1981) speak of initial frame of reference effects. They presented a set of experiments that suggested that respondents may react differently to the first item of a series of rating questions, with respect to how they may react to later items. The set of experiments they present are the ones conducted by Carpenter and Blackwood (1979), who varied the starting point of rating lists concerning attitudes towards animals (see Schuman & Presser, 1981: 51-2). Their findings showed that items usually received either their highest or their lowest values when appearing first in the list. Schuman and Presser (1981) explained that rating a series of items was a setting susceptible to introduce order effects, because of shifting frames of reference.

The second mechanism in which previous questions may interfere in the process of selecting a response category is by heightening the relations among items. In this sense, McGuire (1960), demonstrated that asking questions about related beliefs could make the relation between items more salient, and induce respondents to reduce the inconsistencies among their beliefs. At the moment of selecting a response category, respondents would thus try to appear consistent, and edit their response to make it similar to answers given to previous items (Tourangeau & Rasinski, 1988).

2.4 Moderators of context effects: attitude strength

Since the first studies of context effects, the idea that context would not influence all respondents equally has been present. As explained by Krosnick and Schuman (1988), authors had shared a sort of assumption that individuals whose attitudes are intense, held with great certainty, or taken as personally important, would be less prone to be affected by context. On the contrary, individuals with uncrystallized, weakly held attitudes, could be much easily influenced by the context in which the question is asked – or by the form of the question, or the wording of it –.

This assumption has been explained by a series of arguments. One theoretical explanation to this hypothesis is that stronger attitudes should be

associated with more chronically accessible beliefs. Thus, given that chronically accessible beliefs are context independent, stronger attitudes should be less affected by context (Tourangeau, & Rasinski, 1988; Krosnick, & Schuman, 1988; Lavine, Huff, Wagner, & Sweeney, 1998). In addition, strong attitudes should be retrieved in memory quicker than weaker attitudes (Lavine et al., 1998). Moreover, the assumption has been explained by the fact that these attitudes are likely to be more extreme, and more resistant to change (Krosnick, & Schuman, 1988).

Krosnick and Schuman (1988) tested the hypothesis of attitude strength as a moderator of response effects in 27 experiments. Attitude strength being a complex concept, in order to operationalize it, they measured the related concepts of attitude intensity, attitude importance, and attitude certainty, as indicators of crystallized attitudes. These strength-related characteristics were measured by means of a follow-up question. Among the effects they tested, they studied the abortion effect (see previous sections) in four experiments, evaluating whether attitude certainty and attitude importance were moderators of this context effect. Although the first experiment confirmed the hypothesis that respondents with stronger attitudes (both attitude certainty and attitude importance) were less affected by question order, later replications of the experiments did not present significant results. The authors concluded that the significant results obtained in the first experiment were likely due to chance.

Later work on the topic of attitude strength insisted in the multidimensionality of the concept (see Krosnick, Boninger, Chuang, Berent, & Camot, 1993; Krosnick & Petty, 1995). A four-group categorization of strength-related attitude properties was proposed: first, properties of the attitude itself (e.g., extremity); second, characteristics of the cognitive structure in which the attitude is embedded (e.g., ambivalence); third, subjective perceptions of attitude strength (e.g., importance); and fourth, processes underlying the formation of attitudes (e.g., elaboration) (Krosnick & Petty, 1995).

Taking into account the new theoretical developments on the topic, Lavine et al. (1998) readdressed the question of whether strong attitudes were less susceptible to context effects, by conducting three experiments. They operationalized attitude strength in up to six different dimensions: importance, certainty, intensity, frequency of thought, extremity, and ambivalence. These attitude properties were measured a priori, in a separate session, before testing the context effects. Contrary to previous findings on the topic, Lavine and colleagues found that the strength-related dimension of attitudinal embeddedness did moderate context effects. Moreover,

context effects were also moderated by attitude strength, when multiple strength-related dimensions were taken into account simultaneously in the form of an index.

Based on the new evidence presented by Lavine et al. (1998), Bassili and Krosnick (2000) retested the effects previously studied by Krosnick and Schuman (1988). This time, more strength-related dimensions were included in the analysis, and were evaluated both independently, as well as simultaneously in the form of an index. Their findings showed attitude extremity as a moderator in the abortion context effect. However, no single attitude strength dimension was found to be a consistent moderator of *all* the studied effects, nor was the aggregate of these dimensions a consistent moderator. They concluded that the fact that attitude strength failed to interact consistently with context effects only reinforced the idea that context effects are caused by multiple psychological mechanisms. Attitude strength would have an impact on only some of these processes.

2.5 Other question order effects: the impact of sequence

In the previous sections we have deepened in how the *content* of preceding questions may influence the way respondents answer to subsequence questions. However, question order can also play a role, independently of question content. Schuman and Presser (1981) use the term of “sequence effects” to refer to what they call as more mechanical effects. The central idea is that as the respondent progresses in the questionnaire, either by cognitive fatigue, or loss of interest or motivation, the quality of responses may decrease. Thus, questions appearing later in the questionnaire would be more likely to suffer from fatigue, and have lower response quality, than those appearing earlier.

The theory of satisficing (Krosnick, 1991) provides a framework for understanding how respondents may cope with questionnaire burden or fatigue. Krosnick postulates an extension of the cognitive models of the answering process of attitude questions (comprehension, information retrieval, judgment and selection of response). He postulates that performing these four tasks carefully and comprehensively constitutes what he calls “optimizing”. However, this requires a great cognitive effort that respondents are not always capable or motivated to undertake. In such situations respondents may employ a series of strategies for simplifying these tasks. This is what he calls “satisficing”. A “weak” form of satisficing would consist in performing these same four cognitive steps, only that more superficially. For example, respondents would do this by choosing the first response alternative that seems reasonable, instead of first considering all the alternatives

and, only then, choosing the one that fits the best. The impact of this response strategy is studied in the literature on response order effects. Another form of weak satisficing is acquiescence, defined by Krosnick as "the tendency to agree with or accept any assertion, regardless of its content" (1991: 217).

As fatigue continues to increase, however, respondents would simplify the answering process even further, omitting the retrieval and judgment steps altogether, and limiting themselves to interpret questions only superficially, and to select an answer. This is what he denominates as "strong" satisficing. Krosnick identified four of such strategies: saying "don't know", selecting responses at random (or 'mental coin-flipping'), endorsing always the *status quo* instead of social change, and 'non-differentiation', this is failing to differentiate items presented in a question battery, selecting for all items the same point in the rating scale (Krosnick, 1991: 215). Moreover, Krosnick predicts the appearance of satisficing strategies as a function of three factors: the difficulty of the task the respondent confronts, the ability of the respondent and his motivation. As he points out: "The greater the task difficulty, and the lower the respondent's ability and motivation to optimize, the more likely satisficing is to occur" (1991: 221).

2.6 Item-level vs. scale-level effects

The question of how much do context effects matter has been debated in the field of survey methodology for quite a long time. Within this debate, the distinction between effects at the item level and effects on the correlations between items has been of central importance.

Stouffer and DeVinney (1949) stated that, whereas differences in question order could have an impact on results at the item level, correlations between variables were unaffected by them. Thus, these authors proposed that the way of solving the problem of effects of context was to use multi-item scales, instead of relying on univariate distributions. This idea was later referred to as the assumption of "form-resistant correlations" (Schuman & Presser, 1981). As explained by Sudman et al. (1996), at the time this assumption was proposed by Stouffer and DeVinney, most of the analysis on survey data was focused on associations between variables. That is why, if correlations were indeed unaffected, the problem posed by context effects was perceived as less severe. It was no coincidence that studies on order effects declined considerably in the 1950s, not to be revitalized until the late 1970s, when univariate distributions regained importance, as the interest of analyzing social trends over time increased (Sudman et al., 1996).

Independently of whether correlations suffer of context effects or not, effects at the item-level should not be underestimated. Schuman (1992) develops a number of examples of the pervasive implications context effects at the univariate or marginal results may have. Context effects can have practical political implications. For example, if surveys are treated as referenda to guide policies, even small differences due to context may determine if an issue exceeds or not the 50% barrier that symbolizes the majority. Another example is the one of surveys predicting voter choices. Crespi and Morris (1984) found context effects in such surveys, the support to candidates having been affected by the order in which questions were asked. Since polls have been found to influence voter choices (e.g. Rothschild & Malhotra, 2014), context effects within those polls may have an impact on actual election results.

Context effects at the item-level would be however particularly problematic for the study of social change over time. Schuman (1992) explains that, in order to make analysis of trends over time of a specific item, researchers should replicate the measurement including in the questionnaire all previous topic-related items. Nevertheless, even if the same questions are asked in the same order, it is still possible that a context effects lead to different conclusions. This author presents the example of the US and communist reporters context effect (see section 2.2.1 on part-part assimilation effects). This effect was first identified in data from 1948, and then replicated in 1980. Schuman (1992) explains that when the two questions were asked with the item on American reporters appearing first, followed by the item on Communist reporters, results from 1948 and 1980 led to the conclusion that attitudes towards Communists countries had remained unchanged over the curse of three decades. When the same questions were asked in the reversed order, however, a substantial difference in attitudes towards Communists countries was observable between the two points in time.

Moreover, research found evidence against the form-resistant correlations assumption (e.g. Schuman & Duncan, 1974). Work on context effects from the 1970s on considers both effects on distributions and on associations between variables of great importance. Schuman and Presser (1981) distinguish between effects on marginals (distributions) and on correlations. The distinction made by Tourangeau et al. (2000) is somewhat different. They speak of directional effects and correlational effects. Directional effects would be those effects in which previous items cause responses to a subsequent item to shift towards a given direction. This shift occurs always towards the same direction, independently of how respondents answered the previous items. On the contrary, correlation effects imply a respondent answering

two items consistently to one another. Tourangeau et al. (2000) point out to the fact that most of the context effects reported in the literature are directional effects, which can be explained by the fact that they are much easier to detect.

Feldman and Lynch (1988) draw attention of the consequences of context effects on correlations between items. If respondents use answers to previous questions as inputs for generating later questions, this may increase artificially the correlations between items. These authors speak thus of a 'self-generated validity'. Feldman and Lynch create awareness of how the instrument by itself may contribute to the validity of the construct it is meant to measure.

Knowles (1988) studied personality scales, changing the serial position of items in the instruments. His findings were that as respondents move from the beginning to the end of a 30-item test, their responses become increasingly consistent with one another, and the correlations between items and the scale construct become stronger. Validity among the later items being higher.

Harrison and McLaughlin (1993) produced three different arrangements of the items of the multi-item instrument designed to measure work attitudes (the Job Descriptive Index), The purpose of their study was to test whether item context effect had an effect on the multi-item scale. Effects of the changing order were found at the item-level only. Total scores, variances and reliabilities of the scales were compared. All these scale-level indicators were similar across the different instrument versions.

Another study that changed the content of a scale was the one conducted by Desai and Braitman (2005). Studying properties of instruments assessing violence, these researchers tested the effect of scale carving (i.e. the act of administering only a selection of items, instead of entire instruments). Their findings were that scale means changed as a consequence of reducing the items of the instrument. The reliability of the scale did not present significant differences.

Gehlbach and Barge (2012) studied also differences at the scale level due differences in the question battery content. They found that regrouping together in the questionnaire items measuring a same construct is likely to increase correlations between items, and thus increase artificially the construct consistency. In order to avoid this, these authors recommend to intermix items from different (but still related) constructs in question batteries.

2.7 The split questionnaire design

In this section, the split questionnaire design (Raghunathan & Grizzle, 1995), also referred as modular questionnaire design is presented. This design is being currently

explored in the field of survey methodology as a solution for the conduction of long surveys online.

Researchers have looked for alternative ways for collecting more information, without extending questionnaires. This, mainly in order to reduce respondent burden and survey costs. A proposed solution to this has been to split questionnaires into subsets of questions, and to administer different subsets of questions to different subsamples of individuals, in such way that *every question* is administered to at least a group of individuals. Shoemaker (1973) proposed such a design and called it "multiple matrix sampling design". The name of matrix sampling aimed to reflect "the idea that respondents (rows) and items (columns) are both "sampled" from a conceptual complete population data matrix" (Thomas, Raghunathan, Schenker, Katzoff, & Johnson, 2006: 217). Missing data is inherent to this type of survey design. Imputation methods have been used to create complete data sets. It is important to stress the importance of the imputation techniques in the use of this type of design.

The split questionnaire design as developed by Raghunathan and Grizzle (1995) proposes a specific splitting strategy, in order to ensure that every *pair of questions* is administered to at least a group of individuals. This specific splitting strategy goes one step further with respect to the multiple matrix sampling design, in the extent that it allows to estimate all two-way associations between variables of the entire data set, making the conduction of multivariate analysis possible.

The splitting design Raghunathan and Grizzle (1995) present has been carefully developed with a clear goal in mind: to minimize information loss, so that imputation results are improved. The design consists first in selecting a number of "core items" that are administered to all individuals. These items should be those that predict the best all other variables from the survey, or those that are central for the survey analysis (e.g. sociodemographics). Once the core items have been selected, the next step is to strategically allocate all other items into different modules or blocks. According to these authors, the way of doing this should be to first identify those variables that explain the best each other, and to allocate them into different modules.[5] In this way, it is avoided that items explaining one another very well are jointly missing in an observation.[6] Once the modules are formed, the last step is to

[5] As pointed out by Rässler, Koller, and Mäenpää (2002), it is necessary to have a complete dataset to know what variables correlate with one another. For surveys that are conducted in a regular basis, this is not a problem, because this information can be provided by data from previous waves. If this is not the case, the entire questionnaire should be conducted on a small subsample, as a previous step to the matrix design construction.

[6] As explained by Thomas et al. (2006): "[a] good matrix sampling design allocates split items to blocks in such a way that for each split item excluded from a block, there are split items included in the block

combine them in such ways that every pair of items is administered to at least a subsample of individuals. Table 2.2 illustrates how the final matrix design would look like, if items were allocated into a total of four modules (and the "core" module). In such a case, a total of six different questionnaire versions are administered, each of them to a different subsample of individuals.

Table 2.2: *Example of split questionnaire design with four modules*

		Split variables			
Questionnaire version	Core module	Module 1	Module 2	Module 3	Module 4
1	■	■	■		
2	■	■		■	
3	■	■			■
4	■		■	■	
5	■		■		■
6	■			■	■

■ asked ; □ not asked.

Although most of the strategy for allocating items into modules described by Raghunathan and Grizzle (1995) is determined by the correlation between items, these authors mention that this strategy should also care of maintaining the contextual placement of certain items. They conclude declaring the necessity of refining the splitting strategy.

Adigüzel and Wedel (2008) provide some insights about how best to split the questionnaire into modules. These authors compared two different methods for optimizing the splitting strategy in terms of minimizing information loss. The first, the between-block design, consists in allocating entire blocks of questions – i.e. several questions presented in block in order to measure together a particular trait – into the same module. The second, the within-block design, allocates questions from a same block to different modules. They concluded that the between-block design produced closer estimates to those from the complete data set, while at the same time reduced the completion time and the respondent fatigue. The within-block design, on the contrary, lead to less boredom with the questions. It is important to note that in their study, Adigüzel and Wedel (2008) compared points estimates of means and

that, together with the core items, are predictive of the excluded item; this facilitates the recovery of information about the excluded item during analyses of the data" (p. 219).

variances, but that they did not study the estimation of covariances or of latent construct.

In a moment in which the split questionnaire design becomes increasingly popular in the field of survey methodology, the present study aims to shed light on the risks of introducing context effects as a consequence of the questionnaire split. Concretely, this study wishes to determine whether a question battery can be split without introducing context effects on its measurements.

3. METHODS

3.1 Data: Swiss EVS 2017

An Experimental Design. This project analyzes data from the Swiss European Values Study (EVS) of 2017.[7] Launched in 1981, the EVS is a cross-national survey that has been carried out every nine years in an increasing number of European countries, in order to provide insights about values, attitudes, beliefs and ideas that Europeans have about life, work, family, politics, religion and society.[8] The four first waves of the EVS were all conducted through face-to-face interviews, often considered to be the "golden standard" for international comparative studies. However, in the 5th wave that started in 2017, the option of a self-administered online survey was proposed, to be conducted in parallel to the usual face-to-face interviews. This parallel mixed-mode was thought as a methodological experiment, aiming to evaluate the plausibility for the EVS to transition from face-to-face to web. Countries willing to participate in the experiment could choose between two alternative online survey designs. One was simply to implement the original source questionnaire in a self-administered web survey. The other possibility was to implement it following a matrix questionnaire design. Switzerland undertook both experiments.[9]

The Swiss EVS 2017, directed by the Swiss Centre of Expertise in the Social Sciences (FORS), can be seen as the product of two independent but parallel surveys. On the one hand, a face-to-face survey was conducted as usual, the only difference with the previous wave of 2008 being that the net sample size was reduced by half. On the other hand, a self-administered web/mail survey, following a complex experimental design, was carried out. The aim of this second survey was clear: to answer the question of how best to implement a long online survey. To this end, the design combined 1 hour long online surveys and the matrix design. In addition two other conditions were tested: first, the effects of changing the thematic structure of a questionnaire; and second, what happens when different announcements of survey length are given.

[7] The entire analysis uses data files in their version on date 7th September 2018. I was given access to the data as part of an internship, during which the preparation of the data files was still ongoing. It is possible that still some changes were introduced to the datasets thereafter.

[8] For more information about this survey project, visit the webpage of the European Values Study, https://europeanvaluesstudy.eu/

[9] Other countries participating in the general methodological experiment are Finland, Iceland, the Netherlands, Denmark and Germany. It is still possible that other countries decide to participate as well.

In the following pages the design and implementation of this survey will be described in detail. All the information was obtained from the documentation of the survey field (Ernst Stähli, M., Joye, D., Pollien, A., Ochsner, M., Milbert, P., Nisple, K., & Sapin, M., to be published).

3.1.2 Matrix design experiment

A matrix design was used to shorten the original 1 hour long questionnaire into several 30 minutes long questionnaires, the short versions thought to be better suited for the web mode. In order to build the matrix, the original source questionnaire was split into 5 blocks of comparable length: 4 thematic modules and 1 core block, containing socio-demographics and some transversal questions. Then, based on these thematic modules, the new short questionnaires were created. All possible 2-module combinations were formed, making a total of 6 new short questionnaires. The core block was added to every version. These questionnaire versions, to which I will refer from now on as *main* questionnaires, cover over 50% of the original source questionnaire.

In addition, all persons having completed the main questionnaires, were invited to the answer the *follow-up* questionnaires. These new questionnaires were meant to complete the main questionnaires, offering respondents the remaining 2 thematic modules. In order to ensure comparability, a few questions from the core were asked again.[10] Table 3.1 summarizes the questionnaire matrix design.

Table 3.1: Matrix questionnaire design

Respondent Groups	Main questionnaires					Follow-up questionnaires				
	Core	A	B	C	D	Repet. Quest.	A	B	C	D
M1	■	■	■			■			■	■
M2	■	■		■		■		■		■
M3	■	■			■	■		■	■	
M4	■		■	■		■	■			■
M5	■		■		■	■	■		■	
M6	■			■	■	■	■	■		

■ asked ; □ not asked.

[10] Repeated questions: sex, gender, importance in life and personal health.

The most important criterion when building the new questionnaire versions was to produce questionnaires that were meaningful for respondents. This is why the thematic split, or “between-block” design, was chosen over the purely random split. Indeed, it was argued that a complete random content would demand a higher cognitive burden to respondents, who would have to transition constantly from one topic to the other and got the feeling that the questionnaire were clumsily structured, increasing thus the risk of drop-out. The thematic split was in a way conceived to increase likelihood of respondents participating in the follow-up.

Apart from the thematic split, in order to optimize comparability, the question order of the original source questionnaire was maintained. In other words, questions were skipped, but never moved. For example, question 22 would always appear after question 10; the difference between questionnaire versions was that in some of them some other questions would be asked in between, whereas in other versions question 10 would be followed directly by question 22.

Moreover, special attention was put into potential context effects. Questions related with one another, that could be interpreted differently if appearing separately, were kept in the same block. Likewise, variables usually analyzed together were kept in most cases in the same module.

Question batteries were split only when the new short batteries make still sense to respondents and the meaning of items were not changed. In cases where question batteries were split into different modules, special attention was put into distributing equally positive and negative items. Also, items usually analyzed together were kept together. However, items having very high correlations could be split, in order to increase quality of future imputation.

Finally, the core block contained the most used socio-demographic variables, as well as the most broadly used variables, that are usually used to create subsamples or filter questions. Some of them were in addition repeated in the follow-up questionnaires. Table 3.2 summarizes the thematic content of the questionnaire modules.

Table 3.2: *Presentation of the thematic distribution into five blocks*

Modules	Topics
Core	Socio-demographics, questions often used as controls or correlating highly with other items of the thematic blocks A-D
A	Family, work, sociodemographic questions about parents and partner
B	Religion, morality and social identity
C	Society
D	Politics

3.1.3 Full-length questionnaires experiment

In addition to the matrix design, a 1 hour long online survey experiment was also conducted. The main idea was to test whether a 1-hour long questionnaire can be conducted online, without introducing a too strong bias of selection. For example, if not only highly educated people would be willing to participate in such a long online survey. Furthermore, full length questionnaires were administered in four different variants, aiming to test two more experimental conditions: first, the effects of the thematic structure, and second, the effect of the announced duration.

In order to test the effect of the thematic structure of a questionnaire, two different full length questionnaire versions were administered. On the one hand there was an *original* full-length questionnaire, this is a questionnaire following the same question order as the source questionnaire. On the other hand, there was a *reversed* full-length questionnaire, containing the same questions, but following an inverted thematic order[11]. The idea behind this experimental condition was to evaluate if a less logic and harmonized thematic order could increase the burden for respondents, a condition inherent to the implementation of a matrix design.

The second experimental condition no longer concerned the questionnaire content, but rather the information offered in the invitation letter. Each full-length questionnaire version was administered in two different modalities. To half of them an "*honest*" duration of the survey of 45 minutes was announced in the invitation letter, whereas the other half got a *dishonest* or short time indication of 25 minutes. This experiment wished to test whether cheating on the time announcement of the

[11] This questionnaire followed the same structure as the 5th questionnaire version from the matrix design, first presenting the blocks from the main questionnaire (thus blocks B and D) and then the blocks from the follow-up questionnaire (thus blocks A and C), plus the core. The 5th version was chosen because it was seen as the less harmonized of all new short questionnaire versions.

survey duration could have an impact on participation rates and/or on selection bias. This last experimental condition will not be included in the analysis of this project.

Swiss EVS 2017: Sample

The **target population** was the Swiss resident population aged 18 years or older[12] that lives within private households[13], independently of their nationality, citizenship or language.

The **sampling frame** used was the Swiss population register (Stichprobenrahmen für Personen- und Haushaltserhebungen - SRPF), which is managed by the Swiss Federal Statistical Office by collecting and combining population registers from all municipalities. The SRPF is a highly reliable database that is updated every three months, and which contains, in addition to the persons' name and address, a number of socio-demographic variables, such as sex, year of birth, marital status, nationality, country of birth and type of residence permit (Roberts, Lipps, & Kissau, 2013: 3).

The **sampling design** consisted in one stage stratified random sampling of individuals, who were proportionally allocated according to the seven big regions of Switzerland.[14] As a result, each drawn individual represented the same number of persons in the target population, and thus no weight for correcting probability of selection was needed. Two separate samples were drawn, one for the face-to-face and another for the self-administered survey.

Concerning **sample sizes**, the first one consisted of 1400 individuals, and the second, of 6800 individuals. The self-administered sample was then divided into different subsamples: 4800 individuals were randomly assigned to the matrix experiment – 800 for each questionnaire version – and 2000 to the full length questionnaire experiment – 500 to each variant.

Concerning the **mode**, the face-to-face survey was carried out through Computer Assisted Personal Interviews (CAPI) by a survey agency. The self-administered survey, on the other hand, was conducted integrally by FORS and followed a push to web design. Indeed, all target persons were encouraged to respond online – they all received the web link and a login –, but they were given the possibility to order a paper questionnaire if desired. Later in the fieldwork, non-

[12] On date 1st September 2017.
[13] Excluding thus all persons living in institutions, such as jail and hospitals.
[14] Lake Geneva region, Espace Mittelland, Northwestern Switzerland, Zurich, Eastern Switzerland, Central Switzerland and Ticino (variable NUTS2).

respondents received a printed questionnaire and a prepaid envelope to send it back.

The **field** started in both cases the second week of September 2017[15] and finished between January and February 2018[16]. All persons were first contacted by post mail, and received with the invitation letter an unconditional incentive in form of postal check of 10.- CHF. Invitation letters for the matrix subsample announced already the possibility of being contacted again for a follow-up. After this initial contact, the way individuals were further contacted differs depending on the sample. Individuals from the CAPI sample got up to 5 personal visits from interviewers at home, and in case of non-contact, a second letter was sent, followed by a telephone call.[17] On the contrary, persons from the self-administered survey were only contacted by post mail. Individuals got up to 3 reminders letters for the main questionnaires. Those having completed the main questionnaire received soon after an invitation letter to the follow-up, this time offering a conditional non-monetary incentive: the possibility of winning one of the 3 iPads that would be raffled among participants. There were up to 2 reminder letters for the follow-up. Moreover, a telephone hotline was made available for all target persons in both CAPI and self-administered surveys. Additionally, for the self-administered survey only, an e-mail hotline was also provided.

Interviews and questionnaires were completed in the three national **languages** of Switzerland, german, french and italian. A total of 673 valid interviews were conducted face-to-face (response rate: 52.3%)[18], whereas 2934 complete self-administered questionnaires were collected (response rate: 43.8%)[19]. Table 3.3 presents the response rates of the self-administered survey by questionnaire version.

[15] The first face-to-face interviews were conducted the 11th September 2017, whereas the fieldwork for the self-administered survey started some days later, the 14th September 2017.
[16] The last face-to-face interview took place the 21st January 2018. The fieldwork of the web/mail survey lasted some weeks longer : the last paper questionnaire was registered the 22nd of February 2018.
[17] Only individuals for which a telephone number was available, were called. The others received only a second letter (mentioned above), in which they were asked to take contact by telephone with the survey agency.
[18] To calculate the response rate, the following categories were taken out from the gross sample: R deceased, Language barrier, R moved, still in country, R moved to unknown destination, R moved out of country, Address not traceable, Address not residential: institution.
[19] To calculate the response rate, the following categories were taken out from the gross sample: R deceased, Language barrier, Address not traceable, Address not residential: institution.

Table 3.3: *Response rates in the self-administered survey by questionnaire version*

	Initial Gross sample	Gross sample size[a]	Net sample 1st session	Response rate 1st session	Net sample 2nd session	Response rate 1st + 2nd session
Self-administered survey	**6800**	**6699**	**2934**	**43.8**	-	-
Full-length original (FL1)	1000	979	403	41.2	-	-
Full-length reversed (FL2)	1000	987	439	44.5	-	-
Matrix design	**4800**	**4733**	**2092**	**44.2**	**1661**	**35.1**
M1	800	787	349	44.3	277	35.2
M2	800	787	378	48.0	298	37.9
M3	800	789	325	41.2	250	31.7
M4	800	784	357	45.5	292	37.2
M5	800	792	344	43.4	277	35.0
M6	800	794	339	42.7	267	33.6

Notes:
a: To calculate the gross sample size the following categories were taken out: R deceased, Language barrier, Address not traceable, Address not residential: institution.

3.2 Question Battery on Moral Beliefs

Although questionnaire modularization can potentially introduce context effects of a diverse nature, the present study focuses very concretely on the effects of splitting a question battery. The aim is to evaluate whether a question battery can be split without introducing context effects on its measurements. To this end, a specific question battery will be analyzed, which was split in two modules in the matrix experiment: the question battery on moral beliefs.

Since its first wave in 1981, the EVS has always included a question battery about moral beliefs that can be seen as a classic in analyses with EVS data[20]. The question battery presents always a list of different morally debatable issues/behaviors, such as 'suicide', or 'someone accepting a bribe in the course of their duties', and asks respondents to indicate in a 1-10 scale, whether these issues/behaviors can never be justified (score 1), or if they can always be justified (score 10) – rating scale with end-point labels –. All scores in between allow to indicate intermediate levels of justification. The question wording and the full list of items included in the 2017 EVS are shown in table 3.4 below:

[20] This question battery appears also in the World Values Survey (WVS).

Table 3.4: *Question battery on moral beliefs from the 5th wave of the EVS*

Question text:		
Please tell me for each of the following whether you think it can always be justified, never be justified, or something in between, using this card.		
Items		**Variable names**
1	Claiming state benefits which you are not entitled to	v149
2	Cheating on tax if you have the chance	v150
3	Taking the drug marijuana or hashish	v151
4	Someone accepting a bribe in the course of their duties	v152
5	Homosexuality	v153
6	Abortion	v154
7	Divorce	v155
8	Euthanasia (terminating the life of the incurably sick)	v156
9	Suicide	v157
10	Having casual sex	v158
11	Avoiding a fare on public transport	v159
12	Prostitution	v160
13	Artificial insemination or in-vitro fertilization	v161
14	Political violence	v162
15	Death penalty	v163
Rating scale	Never justified 1 2 3 4 5 6 7 8 9 10 Always justified	

A theory that has inspired studies with these items – and with many other variables of the EVS data – is the modernization theory (Draulans & Halman, 2005; Halman, 2009). These studies have tested the hypothesis of the de-traditionalization or individualization/privatization of moral values, the idea behind being that we would be witnessing a social and historical process in which moral beliefs – and values in general – would be becoming less dependent on traditional institutions, such as the Church, and would increasingly be legitimated by personal choice (Halman, 2009: 36).

The factor structure of the question battery has been repeatedly studied (Phillips & Harding, 1985; Harding et al., 1986; Moors & Wennekers, 2003; Draulans & Halman, 2005; Halman, 2009; Vauclair & Fischer, 2011), researchers having identified between two and three different moral dimensions. There is a general consensus among researchers with respect to one of these dimensions: the self-determination morality, also labeled by authors as personal/sexual morality. Items such as “homosexuality”, “abortion”, “divorce” or “euthanasia” are typical examples of this dimension, which refers to issues/behaviors that mainly concern the private life

sphere. As pointed out by Phillips and Harding (1985), these items have in common that they all represent actions or conducts traditionally condemned as 'sinful' by the Church.

On the other hand, the remaining items denote behaviors and issues related to the public good. Some authors have grouped all these items together under the label of civic morality (Moors & Wennekers, 2003; Halman 2009; Vauclair & Fischer, 2011). Others researchers have further distinguished them into two different categories (Phillips & Harding, 1985; Harding et al., 1986; Draulans & Halman, 2005). So, items such as "cheating on tax if you have the chance" or "avoiding a fare on public transport" would be part of a self-interest or personal interest morality, which indicates tolerance towards actions that strictly speaking, as explained by Phillips and Harding (1985), "contravene the law, but their rightness or wrongness is largely made by the individual, rather than determined by the state" (p. 97). The third dimension is labeled by these same authors as legal morality, and contains items such as "political assassination" or "someone accepting a bribe in the course of their duties". In opposition to the former category, these items would be considered more openly as law-breaking.

Since the first wave of the EVS these items have shown a very strong asymmetry, distributions greatly oriented toward the 'never justified' pole (score 1). However, there has been an evolution in the moral values of Europeans ever since. Indeed, Halman (2009) compared EVS waves from 1981, 1990 and 1999 and found that sexual permissiveness had greatly increased in European societies, whereas civic permissiveness had decreased. Translated into the variables' distribution, this means that responses to items of the self-determination morality have become with time less concentrated in the left pole of the scale – shifting progressively to the right pole –, while items from the civic morality dimension have become even more asymmetric, distributions tending increasingly to the left pole.

Moreover, the variables that interact most strongly and consistently with the moral beliefs items are the level of education, the religious belief, age, and the political affinity. In their own words, "[i]n general terms, the groups which show greatest tolerance in moral outlook are the young, the more highly educated, those who are more left-wing, and those describing themselves as non-religious or atheist" (Harding et al., 1986: 15).

Split of question battery

In the context of the EVS questionnaire modularization experiment, the question battery on morality beliefs was split in two. The division was done thematically. So, items related to the self-determination morality were allocated to the module B, which concerned religion, morality and social identity (see table 3.2). On the other hand, items from the civic moral dimension were allocated to the module C, whose main topic was society. In addition to the thematic criterion, a simultaneous effort was made in order to distribute the items evenly between the two modules. So, given that there were more items on the self-determination morality, the one item that was found to correlate the least strongly with the others, Euthanasia (item 8), was allocated to module C (society). The result was a module containing all the self-determination morality items, except from Euthanasia, and the other module containing all the civic morality items, plus Euthanasia. Table 3.5 presents the original an the resulting two split question batteries.

Due to the questionnaire matrix design, only questionnaire versions that presented together modules B and C got the entire question battery as in the original questionnaire. This was the case of the *main* questionnaire M4, and the *follow-up* questionnaire M3 (see Table 3.1), plus of course the full-length *original* questionnaire (FL1) that was not part of the matrix experiment. All the other questionnaire versions presented the question battery in its split version. Questionnaire versions from the matrix experiment presented either the items of module B only, or the items of module C only. This means in addition that the two parts of the battery on moral beliefs were completed in different survey sessions, a first module on the main survey session and a second module in the follow-up session. The full-length *reversed* questionnaire (FL2) is the exception, because even if the battery was split, both parts were nonetheless administered in the same questionnaire – though separated by other questions. Table 3.6 summarizes how this question battery was administered across the different questionnaire versions of the self-administered survey.

Table 3.5: *Question battery on moral beliefs: whole and split versions*

Whole question battery		Split question battery			
		Module B		Module C	
1	Claiming state benefits	~~Claiming state benefits~~		1	Claiming state benefits
2	Cheating on tax	~~Cheating on tax~~		2	Cheating on tax
3	Taking soft drugs	1	Taking soft drugs		~~Taking soft drugs~~
4	Accepting a bribe	~~Accepting a bribe~~		3	Accepting a bribe
5	Homosexuality	2	Homosexuality		~~Homosexuality~~
6	Abortion	3	Abortion		~~Abortion~~
7	Divorce	4	Divorce		~~Divorce~~
8	Euthanasia	~~Euthanasia~~		4	Euthanasia
9	Suicide	5	Suicide		~~Suicide~~
10	Having casual sex	6	Having casual sex		~~Having casual sex~~
11	Avoiding a fare	~~Avoiding a fare~~		5	Avoiding a fare
12	Prostitution	7	Prostitution		~~Prostitution~~
13	Artificial insemination	8	Artificial insemination		~~Artificial insemination~~
14	Political violence	~~Political violence~~		6	Political violence
15	Death penalty	~~Death penalty~~		7	Death penalty

Table 3.6: *Administration of the question battery on moral beliefs in the Swiss EVS 2017 self-administered survey*

Questionnaire version	Whole Battery	Split Battery	Modules Main session	Modules Follow-up session
Matrix design experiment				
M1		X	B	C
M2		X	C	B
M3	X		-	BC
M4	X		BC	-
M5		X	B	C
M6		X	C	B
Full-length experiment				
FL1 Full-length original	X		BC	
FL2 Full-length reversed		X	B + C	

3.3 Hypotheses

Once overviewed how the question battery on moral beliefs was split in the Swiss EVS 2017, and based on the theoretical literature on context effects reviewed in earlier sections, it is now possible to compute a number of predictions about what the results of this study would be. In this section the three research questions will be recapitulated, and hypotheses to each one of them will be formulated.

RQ1: Effects of splitting at the item level. The first research question is "What are the effects of splitting a question battery on measurements at the item level?". A first prediction is of finding an initial frame of reference effect (Schuman & Presser, 1981). As previously seen, items usually received either their highest or their lowest values when appearing first in the list (Carpenter & Blackwood,1979). As explained by Moore (2002), when appearing first in the list, an item is presented in a noncomparative context, whereas items appearing further down in the question battery are asked in a comparative context. When splitting a question battery, there is one particular item that passes from a comparative context in the original question battery, to a noncomparative context in the split version. This is the case of the item of taking soft drugs (see table 3.5), which passes from the third position in the whole battery, to the first position of module B list of items. So, the first hypothesis states the following:

> **H1a:** *Questionnaire versions in which the question battery was split produced significantly different estimates of V151 (taking the drugs marijuana or hashish) than questionnaire versions in which the whole question battery appeared together.*

A second prediction concerns the item of euthanasia. Whereas in the whole version this item is preceded by a number of items from the self-determination morality (see table 3.5), in the split version this item was allocated to a separate module, away from the other items of the self-determination dimension. It is likely, that the presence of a number of topic-related preceding questions in the whole version introduces a part-part assimilation effect (Schuman & Presser, 1981) on the item of euthanasia. In the split version, this effect would be absent. So the second hypothesis is:

> **H1b:** *Questionnaire versions in which the question battery was split produced significantly different estimates of V156 (euthanasia) than questionnaire versions in which the whole question battery appeared together.*

In the split version, self-determination items in module B were presented alone, without any item from the other morality dimension intermixed in the battery (see table 3.5). In such scenario, it is possible that respondents might have looked beyond the specific items and inferred more easily the underlying moral dimension that was being measured (self-determination morality or permissiveness), introducing thus a context effect due to the framework dimension (Moore, 2002). Furthermore, we could

expect this effect to be an additive effect, because probably only people adhering to those moral beliefs would be capable of reaching the level of abstraction required for the task. So, the third hypothesis states the following:

> **H1c:** *Questionnaire versions in which the question battery was split presented higher values for the variables V153 (homosexuality), V154 (abortion), V155 (divorce), V157 (suicide), V158 (having casual sex), V160 (prostitution) and V161 (artificial insemination), than questionnaire versions in which the whole question battery appeared together.*

RQ2: Moderators of item-level effects. The second research question is "To what extent are any observed effects of splitting at the item level moderated by respondent characteristics?". As previously explained, individuals whose attitudes are intense, held with great certainty, or taken as personally important, would be less prone to be affected by context (Krosnick & Schuman, 1988). In the case of self-determination morality measurements, given that these items represent conducts traditionally condemned as 'sinful' by the Church (Phillips & Harding, 1985), it is expected that highly religious people have more extreme attitudes on these issues and consequently be less affected by context. It is thus predicted that:

> **H2a:** *Context effects found among self-determination items are moderated by the level of religiousness of individuals.*

Moreover, highly educated, young and left-wing people have been found to show greatest tolerance in moral outlook (Harding et al., 1986). It is thus expected that:

> **H2b:** *Context effects found among morality items are moderated by the level of education of individuals.*

> **H2c:** *Context effects found among morality items are moderated by the age group of individuals.*

> **H2d:** *Context effects found among morality items are moderated by the political orientation of individuals.*

RQ3: Effects of splitting on multi-item measurements. The third research question was "To what extent do effects of splitting a battery at the item level affect

multi-item measures embedded in the battery?". Such as postulated by the form-resistant correlations hypothesis (Stouffer & DeVinney, 1949; Schuman & Presser, 1981), it is expected that effects at the item level will not be pervasive enough to alter multi-item measurements. So, three hypothesis are postulated:

> **H3a:** *The reliability of the scales of self-determination morality and civic morality is comparable across the two question battery versions.*
>
> **H3b***: Equal factor structure, factor loadings, and factor intercepts and factor error terms will be found across the whole and split question battery versions, i.e. configural invariance, metric invariance, scalar invariance and strict invariance will hold.*
>
> **H3c:** *Questionnaire versions in which the question battery was split produce comparable scores for the scale of self-determination morality and civic morality than questionnaire versions in which the whole question battery appeared together.*

3.4 Analytical Approach

The analytical approach consisted in comparing the measurements produced by two groups: those who answered to the question battery on moral beliefs in its split version, and those who responded to it in its whole original version.

Included in the analysis were all cases from the *self-administered* survey that had answered to the *entire* question battery (to all 15 items). Concretely, this means that cases from both the matrix and the full-length experiments were included. Also, this means that, from the matrix experiment, the analysis included almost exclusively persons having participated in the main *and* follow-up survey sessions. This, due to the fact that only they had answered all 15 items. The sole exception were respondents from the group M4, to who the entire question battery was already administered in the first survey session (see table 3.6). Thus, for that group, all respondents from the *main* survey were analyzed, independently of whether they took part in the follow-up or not. For all groups, both online and paper questionnaires were included in the analysis.[21] Together, 2568 cases were analyzed. Out of them, 1010 cases conformed the *whole* group and 1558 cases, the *split* group.

[21] The decision whether data from paper questionnaire should be included in the analysis or not was difficult. On the one hand, strictly speaking, the effect of question-order can only be controlled in

3.4.1 Preliminary analysis: sample comparison

The first step consisted in making sure that the composition of the two subsamples was the same. Although individuals were randomly assigned to the different experimental groups (see section 3.1), it is important to note that most of the *split* sample is composed by individuals that agreed to participate in the follow-up. A self-selection bias could have been introduced, such as for example that the follow-up respondents could show a relatively higher level of education. Thus, in order to exclude that differences in measurements between the *split* question battery and the *whole* question battery were not in reality due to differences in the characteristics of persons answering the survey, a comparison of their sample composition was undertaken. First, differences in their socio-demographics were assessed, using data from the sampling frame. Their sex, age, marital status, nationality, household size, urbanization and region of residence were compared.

In addition to the variables from the register, responses given to other particular survey questions were compared. These were variables that previous studies had found to be related to the morality items: the level of education, the religious belief and the political affinity (Harding et al., 1986). So, the highest level of education attained (v243 coded in 9 categories) was compared. The political left-right orientation (v102) was also contrasted. Finally, to compare the religious beliefs of respondents, first their religious denomination (v51 and v52) was contrasted, followed by the strength of their religious beliefs, which was operationalized in the variable of how important is God in your life (v63). In order to test for significant differences, Chi-square Tests of Independence were used for the categorical variables – differences in each category were tested separately –. In addition, to assess the statistical significance of differences among the few numerical variables, the non-parametric unpaired Two-Samples Wilcoxon Test was run, given their non-normal distribution.

Only after these preliminary analyses, the substantial questions of this project could be addressed.

---

situations where respondents have to answer question by question, without the possibility of going back to previous questions, nor knowing in advance the content of following items. In a paper questionnaire, this is not the case, because if desired, a respondent could even read the whole questionnaire before answering to it. Indeed, some effects from subsequent questions have been found in paper questionnaires (Schwarz & Hippler, 1995). The decision of including paper questionnaires in the analysis was taken, because the context effects hypothesized would not dependent on the order of the items, but rather on the overall presence or absence of particular items in the question battery.

3.4.2 Effects of splitting at the item level (RQ1)

First, effects of splitting the question battery on the measurements at the item level were tested. For doing this, the procedure consisted in first scrutinizing visually the data, in order to identify *which* were the items showing the most clear differences across the battery versions, and *how* do these differences looked like. Marked line graphs for each item were drawn, displaying simultaneously the response frequencies (in percentages) of each of the two groups. Graphs even included the category missing data (NA), as a response category like the others, in order to be able to identify at one glance the differences in the way respondents answered the questions.[22] After having identified the differences between the two groups by visual inspection, the statistical significance of these differences was tested. For that, Chi-square Tests of Independence were conducted pairwise over the frequencies of each one of the 10 points of the scale.

The next step was to evaluate what were the implications of splitting the battery on the estimates produced. The estimates of the mean were compared. Given the non-normality of the data, the non-parametric unpaired Two-Samples Wilcoxon Test was used to assess the statistical significance of the differences.

3.4.3 Moderators of item-level effects (RQ2)

The second research question was whether context effects are moderated by characteristics of respondents. For this, the four variables referred in hypothesis H2a, H2b, H2c and H2d (level of religiousness, education level, age and political orientation) were evaluated. In general, the groups that gave the most extreme answers were the youngest (18-29) and oldest (65+), the left-wing individuals (1-3 in variable v102) and the right-wing (8-10 in v102), the most religious (10 in v63) and the least religious (1 in v63), and the individuals with a relatively lower level of education (1-3 v243 recoded) as well as the ones with a relatively higher level of education (7-9 v243 recoded). Each of these four variables was thus divided into 3 categories, according to the direction of their ratings: those who tended to give comparatively lower scores, those who tended to give higher scores, and the ones in the middle. The Kruskal Wallis Equality of Populations Rank Test was computed to evaluate which subgroups responded significantly different to the particular items that had presented context effects. This analysis was conducted only on data from the

[22] The choice of marked lines graphs instead of barplots, more commonly used for categorical data, was because differences in the slopes of the lines connecting the points were a visual help for detecting differences. In addition, graphs presenting the cumulative frequencies were also drawn, again in the marked lines format.

full-length original questionnaire (FL1), in a sort of preliminary analysis. The final step was to actual compare the differences in means across the whole and split questionnaire versions, for the variables found to be significant in the Kruskal Wallis test. Differences in the mean across battery versions were tested, as done previously, with the non-parametric unpaired Two-Samples Wilcoxon Test.

3.4.4 Effects of splitting on multi-item measurements (RQ3)

To evaluate effects of splitting the question battery on the relations between variables, first the correlations between items were compared. For that, correlation matrixes were computed for each battery version. The objective was to identify if important shifts in the correlations between the variables had occurred. Next, a Principal Component Analysis (PCA) for each of the two battery versions was conducted, in order to see if, despite possible changes in the correlations, the factor structure of the battery remained the same. Only factors with eigenvalues above 1 were retained. The rotation method used was Varimax with Kaiser Normalization. Reliability tests were conducted for each moral dimension, separately for the whole and split question battery versions.

The last part of the analysis tested whether the latent concepts of self-determination morality and of civic morality are measured equivalently across the two battery versions. This is, whether the same factor structure holds across the two measurement contexts. This was done by means of a Multi-group Confirmatory Factor Analysis (MCFA).

Measurement invariance testing has been usually used to evaluate whether an instrument shows the same psychometric properties across heterogeneous groups (Chen, 2007). This type of analysis has been used above all to evaluate how a same instrument performs across different populations (for example different countries or cultures) (Freitag & Bauer, 2013)- , or to evaluate if the same instrument performs equally across different points in time (see for example Poznyak et al., 2014). In this context the test of measurement invariance is used to reveal whether different instruments have the same psychometric properties across an homogeneous group. This perspective has also been mobilized in analyses of mixed modes surveys, to assess if two different modes of data collection produce equivalent measurements (see for example Hox et al., 2015; Klausch et al., 2013).

The measurement invariance analysis was done separately for each latent concept. First, the baseline models were defined, and tested on all cases – *split* and *whole* question battery versions together. Then, the models were tested separately

on each of these two groups. Once verified that the model fitted both groups, the measurement invariance across the two groups could be tested. First, *configural invariance* between the two groups was tested. This is to see if the general baseline model fitted both groups when no cross-group constraints were added. Then, cross-group constraints were added to the model in order to test to what extent measurements from both groups were invariant. First, in order to test for *metric invariance*, we added the constraint that factor loadings of all items had to be equal in both groups. Thought graphically, this means that all manifest variables had to have an equal regression slope with respect to the latent concept. Then, we established that factor loadings and intercepts had to be equal in both groups in order to test for *scalar invariance*. After that, the next model to be tested was *strict invariance*, in which factor loadings, intercepts and residuals had to be equal. Finally, the last model included the constraint that latent means had to be equal. Adding model constraints inevitably decreased the goodness of fit of the model. However, when an additional constraint caused the CFI to decrease in more than .01, the new model was rejected (Chen, 2007). In such cases, partial invariant models were tested, by freeing some of the constraints imposed in the model. The estimation method used was Maximum Likelihood (ML).

4. RESULTS

4.1 Preliminary analysis: sample comparison

After a comparison between the composition of the two subsamples, *whole* and *split* groups, almost no significant differences were found (see tables 4.1 and 4.2). The only difference was found in the percentage of respondents having declared a score of 7 in the political left-right orientation variable. Results from this preliminary analysis support the idea that the two groups are very likely samples of the same population. Based on these results, the comparison in the measurements of the morality items across the two groups of interest – split and whole – could be done directly. There was no need of controlling for sample differences.

Table 4.1: *Sample comparison, socio-demographics from the sampling register*

	(1) Respondents to WHOLE QUESTION BATTERY n=1010		(2) Respondents to SPLIT QUESTION BATTERY n=1558		(3) Difference between respondents to the whole and the split question battery (2)-(1)	
	%	Std. Err.	%	Std. Err.	%	Sig.[a]
Male	48.8	1.6	45.3	1.3	-3.6	
Age group						
<30	16.4	1.2	15.3	0.9	-1.2	
30-39	17.1	1.2	16.0	0.9	-1.2	
40-49	15.9	1.2	18.6	1.0	2.6	
50-64	26.8	1.4	28.6	1.1	1.8	
65+	23.7	1.3	21.6	1.0	-2.1	
Marital Status						
Single	33.7	1.5	32.4	1.2	-1.2	
Married	52.0	1.6	54.0	1.3	2.0	
Divorced	9.9	0.9	8.6	0.7	-1.3	
Widowed	4.0	0.6	4.2	0.5	0.3	
Other (Legal Part. /dissolved)	0.5	0.2	0.8	0.2	0.3	
Nationality						
Swiss	84.4	1.1	85.6	0.9	1.3	
Bordering country	7.8	0.8	8.7	0.7	0.8	
Other European country	5.8	0.7	4.6	0.5	-1.2	
Other continent	2.0	0.8	1.1	0.6	-0.9	
Household size						
1	17.0	1.2	17.1	1.0	0.0	
2	35.7	1.5	35.8	1.2	0.1	
3	18.4	1.2	19.1	1.0	0.7	
4+	28.8	1.4	28.0	1.1	-0.8	
Urbanisation						
City/town center	27.3	1.4	29.6	1.2	2.4	
City/town suburbs	45.9	1.6	45.3	1.3	-0.6	
Isolated town	0.8	0.3	1.0	0.3	0.2	
Rural community	26.1	1.4	24.1	1.1	-2.0	
NUTS region						
Région lémanique	17.9	1.2	18.9	1.0	1.0	
Espace Mittelland	24.1	1.3	22.5	1.1	-1.6	
Nordwestschweiz	13.5	1.1	14.2	0.9	0.7	
Zürich	16.5	1.2	17.5	1.0	1.0	
Ostschweiz	12.9	1.1	12.0	0.8	-0.9	
Zentralschweiz	9.8	0.9	9.9	0.8	0.1	
Ticino	5.3	0.7	5.0	0.6	-0.3	

Notes:
a: Chi-square tests of independence: *** p<.001, **p<.01, *p<.05.

Table 4.2: *Sample comparison, variables from the EVS survey*

	(1) Respondents to WHOLE QUESTION BATTERY n=1010		(2) Respondents to SPLIT QUESTION BATTERY n=1558		(3) Difference between respondents to the whole and the split question battery (2)-(1)	
	%[a]	Std. Err.	%[a]	Std. Err.	%	Sig.[b]
Education (v243)						
Primary education (un)finished	2.5	0.5	1.9	0.3	-0.6	
Lower secondary education	9.5	0.9	8.4	0.7	-1.1	
Basic vocational training	4.5	0.7	4.6	0.5	0.1	
Apprenticeship 3-4 years	25.4	1.4	23.3	1.1	-2.1	
General education giving (partial) access to tertiary education	4.1	0.6	3.5	0.5	-0.6	
Post-secondary / First stage of tertiary education	26	1.4	26.6	1.1	0.6	
Tertiary education 1, Professional	12.2	1.0	14.4	0.9	2.2	
Tertiary education 1, Academic	13.5	1.1	14.5	0.9	1.0	
PhD	2.2	0.5	2.7	0.4	0.5	
Political orientation (mean)	5.3	0.1	5.3	0.1	0.1 [c]	
Political left-right orientation (v102)						
1 Left	3.7	0.6	3.0	0.4	-0.7	
2	4.8	0.7	5.3	0.6	0.6	
3	13.2	1.1	11.7	0.8	-1.5	
4	11.1	1.0	11.9	0.8	0.8	
5	28.5	1.4	27.1	1.1	-1.4	
6	12.1	1.0	11.6	0.8	-0.4	
7	10.3	1.0	13.7	0.9	3.3	*
8	10.1	0.9	9.7	0.8	-0.4	
9	2.8	0.5	3.1	0.4	0.2	
10 Right	3.4	0.6	3.0	0.4	-0.4	
Religious denomination (v51 + v52)						
Protestant reformed	26.1	1.4	27.0	1.1	0.9	
Free evangelical churches	3.5	0.6	2.8	0.4	-0.7	
Roman Catholic	26.2	1.4	27.7	1.1	1.5	
Christian Catholic	7.2	0.8	6.9	0.6	-0.3	
Islamic	2.6	0.5	2.4	0.4	-0.2	
Other	3.1	0.5	2.0	0.4	-1.1	
None	31.1	1.5	31.0	1.2	-0.1	
Importance of god (mean)	5.03	0.1	4.92	0.1	-0.1 [c]	
importance of God in your life (v63)						
not at all important	24.4	1.4	25.6	1.1	1.2	
2	9.3	0.9	7.6	0.7	-1.7	
3	6.1	0.8	7.6	0.7	1.5	
4	3.8	0.6	4.5	0.5	0.7	
5	12.1	1.0	13.0	0.9	0.9	
6	6.8	0.8	6.5	0.6	-0.3	
7	8.7	0.9	8.1	0.7	-0.6	
8	9.9	0.9	8.9	0.7	-1.0	
9	3.8	0.6	3.8	0.5	0.0	
very important	15.0	1.1	14.4	0.9	-0.6	

Notes:
a: Valid percentage.
b: Chi-square tests of independence: *** p<.001, **p<.01, *p<.05.
c: Unpaired Two-Samples Wilcoxon Test (non-parametric): *** p<.001, **p<.01, *p<.05.

4.2 Effects of splitting at the item level (RQ1)

Hypothesis H1a was confirmed. Compared to the whole question battery, the split version presents a .46 points *higher* mean for the justifiability of taking soft drugs (W = 691970, p<.001; see table 4.3). Indeed, whereas in the whole question battery this item has a mean of 3.71, in the split version the mean is 4.18. Although scores for this item tend in both versions towards "never justifiable" pole, this tendency is more pronounced in the whole question battery. The values of the 1st and 3rd quartiles attest this too: in the *whole* version, 50% of the respondents place themselves between the scores 1 (1st quartile) and 5 (3rd quartile); in contrast, in the *split* version they do it between the scores 1 and 7. The analysis of the response frequencies (see table 8.1 in Appendix) showed in what way respondents answered differently across the two battery versions. The split version contains a much lower proportion of people choosing the extreme value 1 (28.9% compared to 37.8%), a difference that is highly significant (p<.001). On the other hand, responses between the scale points 6-8 are in comparison significantly more frequent in the split version. The boxplots (figure 4.1) and marked line graph (figure 4.2) presented below allow to visualize the described differences.

Table 4.3: *Difference in the estimates of the mean of moral beliefs variables, by question battery version*

		(1) Whole question battery (N= 1010)						(2) Split question battery (N=1558)						(3) Difference between whole and split question battery (2)-(1)	
Variable	**Item**	**N (% valid)**	**Mean**	**SE[a]**	**Q1[a]**	**M[a]**	**Q3[a]**	**N (% valid)**	**Mean**	**SE[a]**	**Q1[a]**	**M[a]**	**Q3[a]**	**Δ Mean**	**Sig.[b]**
v149	State benefits	998 (98.8)	2.20	.06	1	1	3	1537 (98.7)	2.06	.05	1	1	2	-.14	
v150	Cheating on tax	996 (98.6)	2.05	.06	1	1	2	1543 (99.0)	2.00	.05	1	1	2	-.05	
v151	Taking soft drugs	999 (98.9)	3.71	.09	1	3	5	1544 (99.1)	4.18	.07	1	3.5	7	.46	***
v152	Accepting a bribe	999 (98.9)	1.63	.05	1	1	1	1539 (98.8)	1.54	.04	1	1	1	-.10	
v153	Homosexuality	988 (97.8)	7.78	.09	5	10	10	1537 (98.7)	7.87	.07	6	9	10	.09	
v154	Abortion	996 (98.6)	6.43	.09	5	7	9	1540 (98.8)	6.49	.07	5	7	9	.06	
v155	Divorce	996 (98.6)	7.54	.08	5	8	10	1542 (99.0)	7.60	.06	5	8	10	.06	
v156	Euthanasia	992 (98.2)	6.75	.09	5	8	9	1521 (97.6)	5.78	.07	3	6	8	-.97	***
v157	Suicide	990 (98.0)	4.45	.09	2	5	7	1529 (98.1)	4.78	.07	2	5	7	.33	**
v158	Having casual sex	989 (97.9)	5.21	.10	2	5	8	1538 (98.7)	5.36	.08	3	5	8	.16	
v159	Avoiding a fare	1003 (99.3)	2.47	.07	1	2	3	1545 (99.2)	2.39	.05	1	2	3	-.08	
v160	Prostitution	992 (98.2)	4.59	.09	2	5	7	1535 (98.5)	4.65	.07	2	5	7	.06	
v161	Artificial insemination	991 (98.1)	6.56	.09	5	7	9	1536 (98.6)	6.61	.07	5	7	9	.05	
v162	Political violence	992 (98.2)	2.03	.06	1	1	2	1535 (98.5)	1.73	.04	1	1	2	-.30	***
v163	Death penalty	998 (98.8)	3.11	.09	1	2	5	1536 (98.6)	2.95	.07	1	2	5	-.16	

Notes:

a: SE= Standard error of the mean; Q1= First quartile; M= Median; Q3= Third quartile.

b: Unpaired Two-Samples Wilcoxon Test (non-parametric): *** p<.001, **p<.01, *p<.05.

Figure 4.1: *Boxplots of* ***taking soft drugs****, by question battery version*

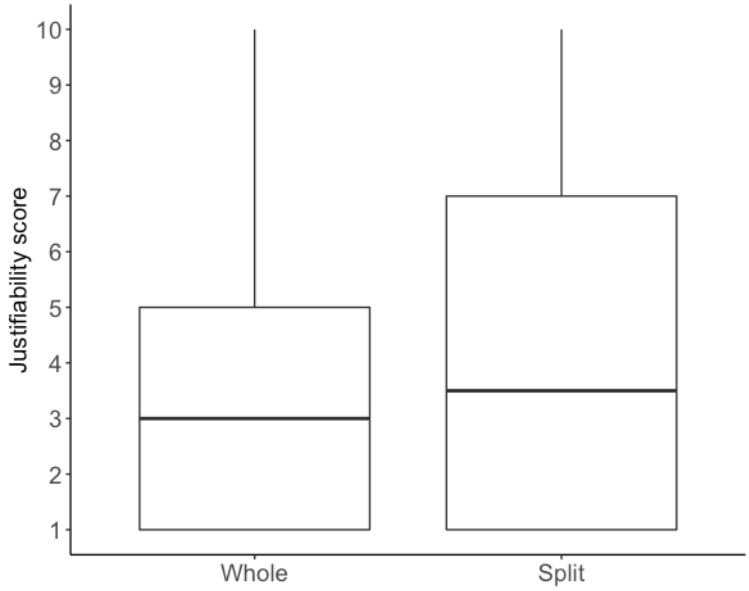
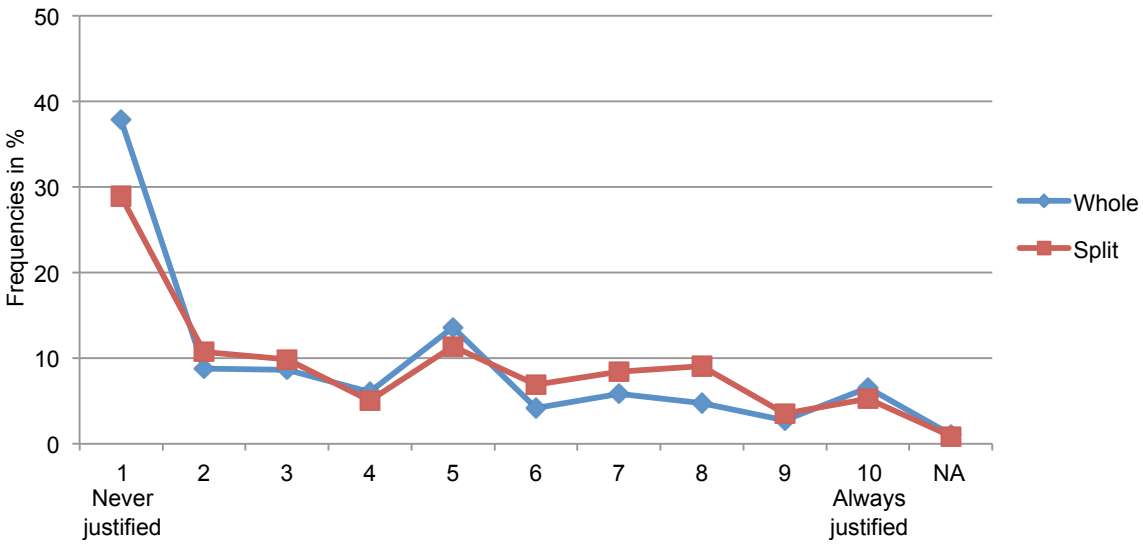


Figure 4.2: *Response frequencies of* ***taking soft drugs****, by question battery version*



Hypothesis H1b was also confirmed. The item of euthanasia presents differences across the two question battery versions. In the split battery version the mean of tolerance towards euthanasia is almost 1 point *lower* than in the whole version (W = 903080, p<.001, see table 4.3). Indeed, while the mean is 6.75 in the whole battery version, the split question battery presents a mean of 5.78. The difference between the two question battery versions is even more pronounced if the values of the median are compared: in the whole version, the median is the score 8, whereas in the split it is placed in the score 6. Analyzing the 1st and 3rd quartiles is also enlightening. In the whole version, 50% of the respondents place themselves between the values 5 and 9. In the split version, in contrast, they do so between the values 3 and 8. These results are presented graphically in the form of boxplots in figure 4.3.

The analysis of the response frequencies (see table 8.1 in Appendix) shows differences by scale point. The split version produced significantly higher frequencies in the scores 1-3, the difference in the scale point 1 being particularly remarkable (13.4%, compared to 8.5% in the whole version, p<.001). On the contrary, the frequency for the scale point 10 is in the split version much lower (11.9%, compared to 22.8% in the whole version, p<.001). The marked line graph in figure 4.4 present these differences visually. It is interesting to observe in those graphs that both curves have peaks of frequencies in the scale points 5, 8 and 10. However, whereas in the whole battery version these peaks have an *ascendant* trend – the scale point 10 presents the highest frequency –, in the split question battery those peaks have a *descendent* trend – the most representative answer category is 5 (17.6%). Moreover, it is also worthy to note the difference between the frequencies of the two extreme values. Whereas in the whole battery version the percentage answering 10 more than doubles the response frequency of 1 (22.8% compared to 8.5%), in the split version it is rather the answer category 1, which is more representative than the category 10 (13.4% vs. 11.9%).

Figure 4.3: *Boxplots of* ***euthanasia****, by question battery version*

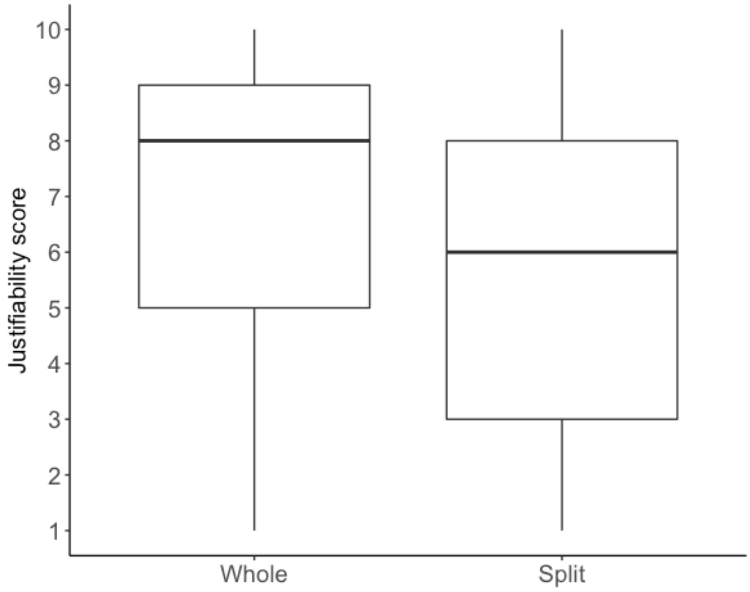
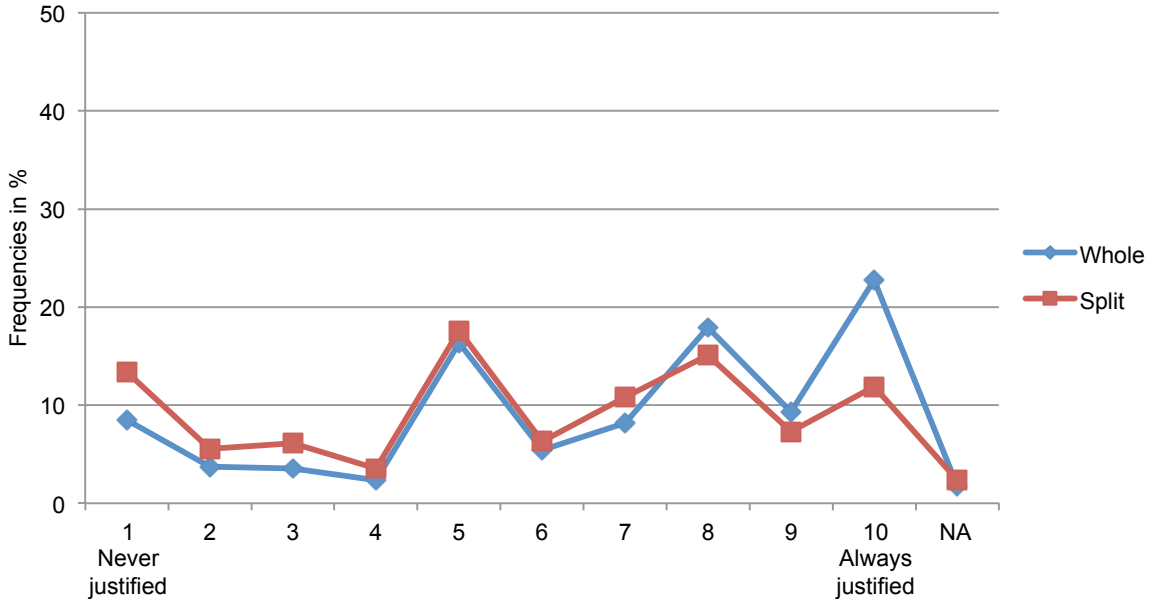


Figure 4.4: *Response frequencies of* ***euthanasia****, by question battery version*

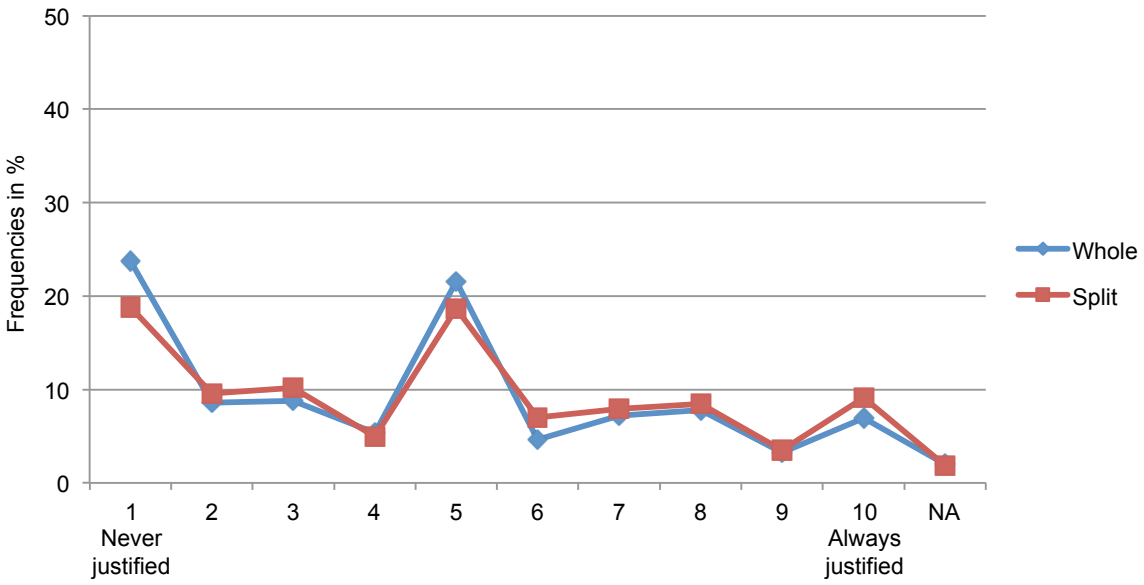


Another item presenting significantly different estimates at the mean across the two question battery versions is the one of suicide (see table 4.3), which is .33 points *higher* in the split version (W = 705390, p<.01). Indeed, while the estimate for the mean in the whole battery is 4.45, in the split battery it is of 4.78. The difference in

this item is less perceptible than in the other items. Indeed, the median and quartiles stay the same, the boxplots being identical. An analysis of the response frequencies by scale point revealed however significant differences in two values (see figure 4.5 and table 8.1 in appendix for details). The whole battery version contains significantly more persons selecting the scale point 1 (23.8%, in comparison to 18.8%, p<.01), and less respondents choosing the point 6 (4.7%, compared to 7.9%, p<.05).

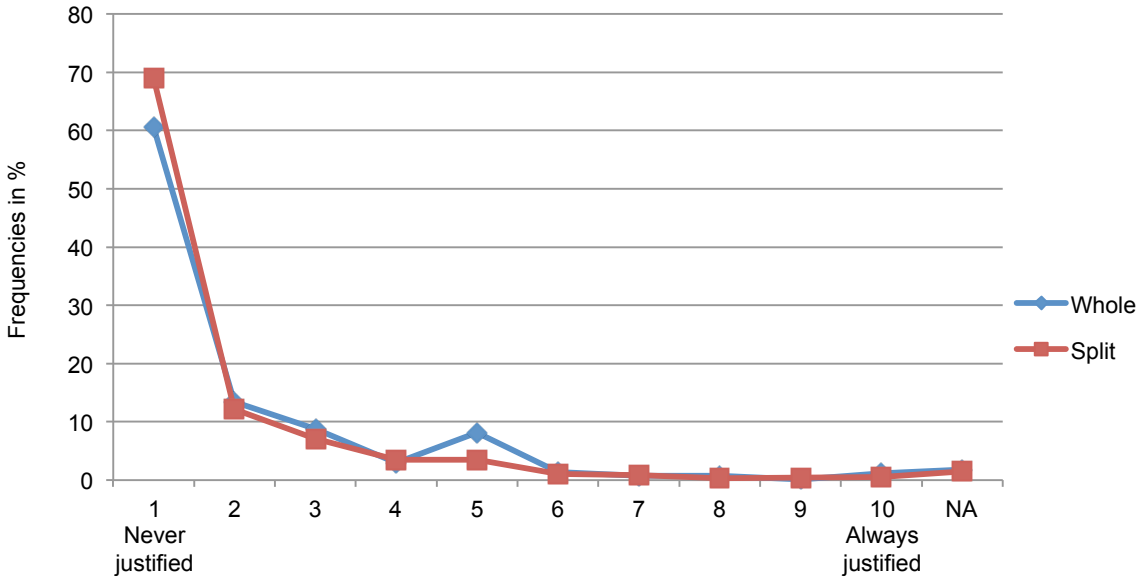
Hypothesis H1c could not be confirmed. Out of the 7 items hypothesized to present higher scores in the split version, only suicide presented significant differences. It is nonetheless worthy to note that although differences are not significant for the other 6 items, all them present higher estimates of the mean in the split version (see table 4.3).

Figure 4.5: *Response frequencies of* ***suicide****, by question battery version*



A forth item has significant different estimates of the mean, depending on the question battery version: political violence (see table 4.3). Whereas the whole question battery presents an estimate of the mean of 2.03, in the split version the estimate of the mean is 1.73. Thus, in the split version the mean is .30 points lower (W = 831970, p<.001). This item is extremely skewed, and thus no difference is perceptible when analyzing the median and quartiles. However, an evaluation of the response frequencies by scale points reveals significant differences in the frequencies of the values 1 and 5 (p<.001 for both cases, see table 8.1 in the appendix). The marked line graph in figure 4.6 shows those differences.

Figure 4.6: *Response frequencies of* ***political violence****, by question battery version*



There is a particular phenomenon that can be observed repeatedly among the items and that is worthy of attention. Independently of whether responses to an item tend to the left pole 'never justified' or towards the right 'always justified' pole, a peak in the distribution of frequencies at the scale point 5 is observed in almost all items. However, what catches attention, is that this peak is almost always higher for the whole battery version. Indeed, this is true for 12 out of the 15 items, this difference being statistically significant for 5 items: accepting a bribe, homosexuality, abortion, political violence and death penalty.

Moreover, some items present differences in the response frequencies across the rating scale, that did not translate into significant differences in the estimate of the mean. The study of the response frequencies (see table 8.1 in appendix) reveal that respondents answered rather differently to the items of abortion, prostitution and artificial fertilization.

It is very interesting to point out that, all the items concerning the civic morality obtained lower estimates of the mean in the split version, without exception (see table 4.3), even if differences were significant only for the item of political violence. Likewise, all the items covering the self-determination morality present higher estimates of the mean in the split version, even if not significant in all cases. The only exception, as already described, is the item of euthanasia.

4.3 Moderators of item-level effects (RQ2)

After having identified the items that presented significant differences in the estimate of the mean across the two battery versions, the next step was to evaluate if these differences were moderated by characteristics of respondents. Four variables were considered as potential moderators: religiousness (H2a), level of education (H2b), age (H2c), and political orientation (H2d). Each of these variables was divided into three categories.

First, as a preliminary analysis, it was tested whether these subgroups responded significantly different toward the items that had presented context effects, taking soft drugs, euthanasia, suicide and political violence. Results are presented in table 8.2 in the appendix. The three subgroups of distinct religiousness level responded significantly different to the items of taking soft drugs, euthanasia and suicide. The higher the level of religiosity, the lower the scores given to those items. The subgroups comparison according to level of education showed that people with the highest level of education answered significantly higher to the questions of taking soft drugs and suicide with respect to the other groups. The comparison according to age group showed that the oldest gave significantly lower scores to the item of taking soft drugs, compared to the categories of age. Moreover, the youngest rated the item of political violence significantly higher than their older. Finally, the comparison by subgroups of political orientation showed that the ones at the leftest of the political spectrum presented significantly higher means of justifiability of taking soft drugs, comparing to the other two categories. In addition, it was found that persons at the two extremes of the political spectrum (left and right) presented higher scores for the items of suicide and political violence than the ones at the center of the left-right political scale. Differences were significant between left and center for the item of suicide, and between right and center for the item of political violence.

Then, the variables that were found to interact with the items of taking soft drugs, euthanasia, suicide and political violence were retained for the next analysis: the one of moderation of context effects. A comparison of the means of these four morality items across whole and split battery versions was conducted, this time by subgroups of respondents. Results of these comparisons are summarized in table 4.4.

Table 4.4: *Difference in the estimate of the mean of moral beliefs items across question battery versions, by subgroups of religiousness level, education level, age, and political orientation*

		(1) Whole question battery			(2) Split question battery			(3) Difference between whole and split question battery (2)-(1)
Item	**Subgroup**	**N (% valid)**	**Mean**	**Std. Err.**	**N (% valid)**	**Mean**	**Std. Err.**	**Δ Mean Sig.[a]**
Taking soft drugs								
	Religiousness level							
	Highest	147 (99.3)	2.63	.22	216 (97.3)	2.80	.17	.18
	Medium	592 (98.8)	3.47	.11	924 (99.7)	4.07	.09	.60 ***
	Lowest	241 (100)	5.00	.21	395 (99.7)	5.17	.15	.17
	Education level							
	Lowest	160 (98.8)	3.00	.22	226 (98.7)	3.19	.19	.19
	Medium	542 (99.3)	3.48	.12	810 (99.3)	3.96	.10	.48 ***
	Highest	274 (99.6)	4.65	.18	479 (99.0)	5.04	.13	.40 *
	Age							
	Oldest	233 (97.5)	2.37	.15	331 (98.5)	2.67	.13	.29
	Middle	601 (98.8)	3.95	.12	977 (99.3)	4.34	.09	.39 **
	Youngest	165 (99.4)	4.75	.24	236 (99.2)	5.63	.20	.88 **
	Political orientation							
	Rightest	161 (99.4)	2.91	.21	237 (98.3)	3.23	.18	.31
	Center	607 (99.3)	3.51	.11	972 (99.4)	3.98	.09	.47 ***
	Leftest	212 (99.5)	5.00	.22	303 (99.7)	5.62	.16	.62 *
Euthanasia								
	Religiousness level							
	Highest	145 (98.0)	4.60	.27	211 (95.0)	4.10	.21	-.50
	Medium	590 (98.5)	6.93	.11	910 (98.2)	5.81	.09	-1.12 ***
	Lowest	240 (99.6)	7.70	.16	390 (98.5)	6.55	.14	-1.15 ***
Suicide								
	Religiousness level							
	Highest	146 (98.6)	2.89	.20	216 (97.3)	3.37	.18	.48 *
	Medium	589 (98.3)	4.37	.11	913 (98.5)	4.69	.09	.33 *
	Lowest	237 (98.3)	5.70	.19	391 (98.7)	5.77	.15	.07
	Education level							
	Lowest	157 (96.9)	3.74	.22	222 (96.9)	4.33	.19	.59 *
	Medium	541 (99.1)	4.16	.12	800 (98.0)	4.61	.10	.45 **
	Highest	269 (97.8)	5.52	.17	479 (99.0)	5.31	.13	-.22
	Age							
	Oldest	230 (96.2)	4.04	.19	331 (98.5)	4.29	.15	.24
	Middle	598 (98.8)	4.69	.12	964 (98.0)	4.94	.09	.25
	Youngest	162 (97.6)	4.15	.20	234 (98.3)	4.82	.19	.67 *
	Political orientation							
	Rightest	161 (99.4)	4.52	.23	238 (98.8)	4.33	.19	-.19
	Center	601 (98.4)	4.27	.11	960 (98.2)	4.65	.09	.38 **
	Leftest	209 (98.1)	5.04	.21	301 (99.0)	5.62	.17	.57 *
Political violence								
	Age							
	Oldest	233 (97.5)	1.82	.11	329 (97.9)	1.67	.09	-.15
	Middle	597 (98.8)	1.96	.07	974 (99.0)	1.66	.05	-.30 ***
	Youngest	162 (97.6)	2.57	.13	232 (97.5)	2.11	.10	-.46 **
	Political orientation							
	Rightest	160 (98.8)	2.23	.16	238 (98.8)	2.00	.12	-.23
	Center	602 (98.5)	1.97	.07	962 (98.4)	1.71	.05	-.26 *
	Leftest	213 (100)	2.07	.11	301 (99.0)	1.63	.07	-.43 ***

Notes:

a: Unpaired Two-Samples Wilcoxon Test (non-parametric): *** p<.001, **p<.01, *p<.05.

Religiousness level. Differences across battery versions for the items of taking soft drugs, euthanasia and suicide changed as a function of the level of religiousness of respondents. For the item of taking soft drugs, differences in the mean across battery versions were no longer significant among the most religious and the least religious (see table 4.4 and figure 4.9). Only the subgroup of medium level of religiousness presented significantly different responses to this item across battery versions.

For the item of euthanasia, differences across battery versions tended to increase, as the level of religiosity of respondents decreased (see figure 4.7). Among the most religious, no significant difference in the mean was found between split and whole battery versions.

For the item of suicide, on the contrary, the higher the level of religiousness of the respondents, the bigger the difference in the mean across battery versions (see figure 4.8). Among the least religious, no significant difference between the two battery versions was found.

These findings supported the ***hypothesis H2a***. Indeed, the level of religiousness was found to moderate the differences across battery versions of the three items of the self-determination morality dimension.

Figure 4.7*: Difference in the estimate of the mean of* ***euthanasia****, by battery version and level of religiousness*

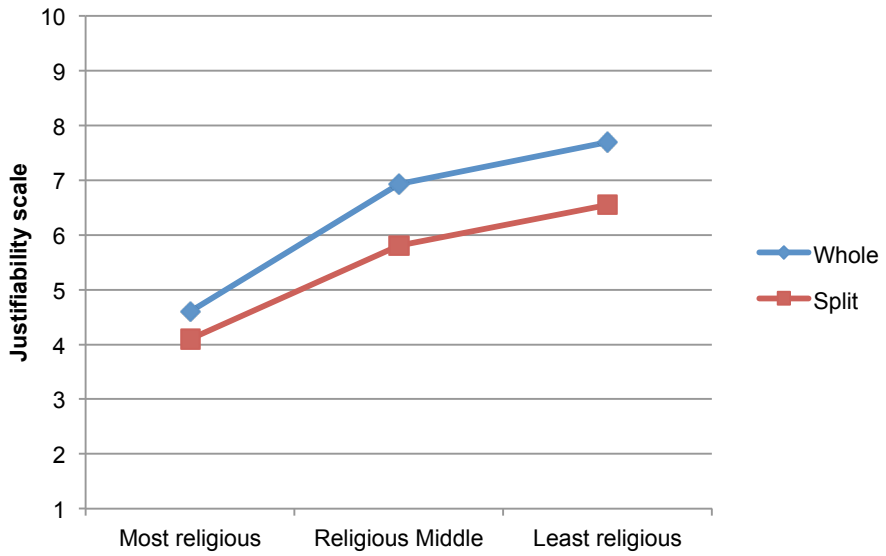
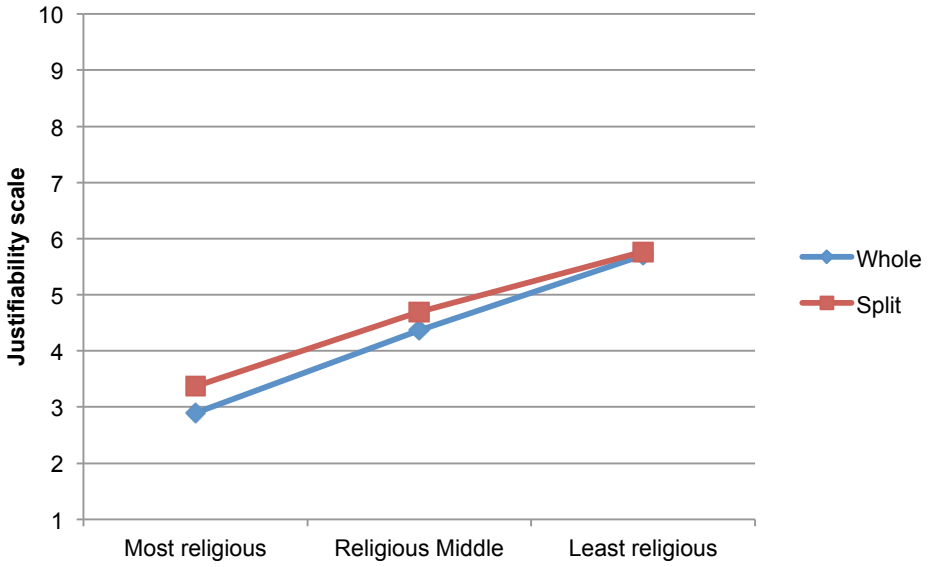


***Figure 4.8**: Difference in the estimate of the mean of **suicide**, by battery version and level of religiousness*



Education level. Differences across battery versions for the items of taking soft drugs and suicide changed as a function of the level of education of respondents. In the case of the item of taking soft drugs, respondents with the lowest level of education presented no significant differences in the mean of this item across battery versions.

For the item of suicide, on the contrary, it was the subgroup with the lowest level of education the one presenting the biggest differences in the mean across battery versions (see table 4.4). The higher the level of education, the lower the differences between split and whole battery versions for this item. Indeed, the difference was no longer significant among respondents with the highest level of education. These findings supported what was ***hypothesized in H2b***, this is that the level of education would moderate the differences of means across battery versions.

Age group. Differences across battery versions for the items of taking soft drugs, suicide and political violence changed as a function of the age of respondents. For the three items, differences in the mean across battery versions increased the younger the respondents were (see table 4.4 and figure 4.10 for the item of political violence). Also, In all three cases, the subgroup of oldest respondents did not present significant differences in the means across battery versions. Figure These findings supported what was ***hypothesized in H2c***, this is that age would moderate the differences of means across battery versions.

Political orientation. Differences across battery versions for the items of taking soft drugs, suicide and political violence changed as a function of the political orientation of respondents. For all these three items, differences between the split and the whole battery were higher, the more to the left respondents placed themselves along the political left-right spectrum. In all three cases, also, differences across battery versions were no longer significant among respondents at the rightest of the political scale. These findings supported what was ***hypothesized in H2d***, this is that the political orientation of respondents would moderate the differences of means across battery versions.

Figure 4.9 illustrates the complete analysis by subgroups of responses given to the item of taking soft drugs across the two battery versions.

Figure 4.9: *Differences in the estimate of the mean of* ***taking soft drugs****, by question battery version and subgroup of respondents*

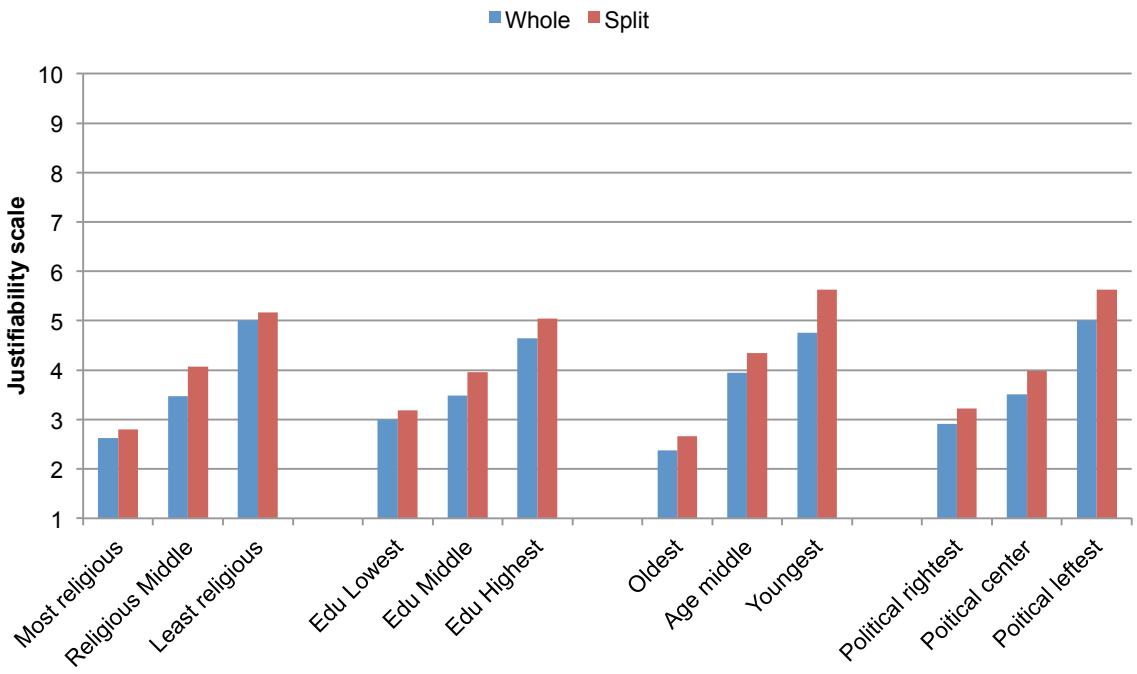
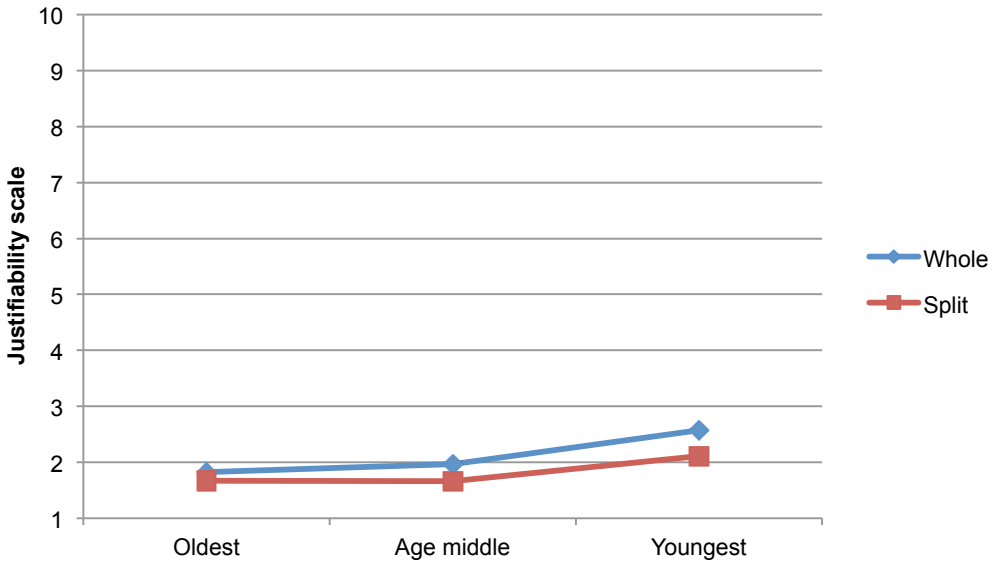


Figure 4.10*: Difference in the estimate of the mean of* ***political violence****, by battery version and age group*



4.4 Effects of splitting on multi-item measurements (RQ3)

Correlations between the items were analyzed by moral dimension (see tables 4.5 and 4.6). The item of euthanasia showed lower correlations with the other self-determination items, when presented in the split question battery version. Differences are presented in figure 4.11. In the split version, the item of suicide was also less correlated with the other items of this moral dimension, but differences were small. The correlation between the item of taking soft drugs and of homosexuality increased in the split version, in which both items were presented one after the other, whereas in the whole battery version another item was asked between the two (accepting a bribe). Interestingly, the split version shows also smaller correlations between the items of divorce and abortion, although in both battery versions the items appeared together.

Figure 4.11: *Correlations between* ***euthanasia*** *and other self-determination items, by battery version*

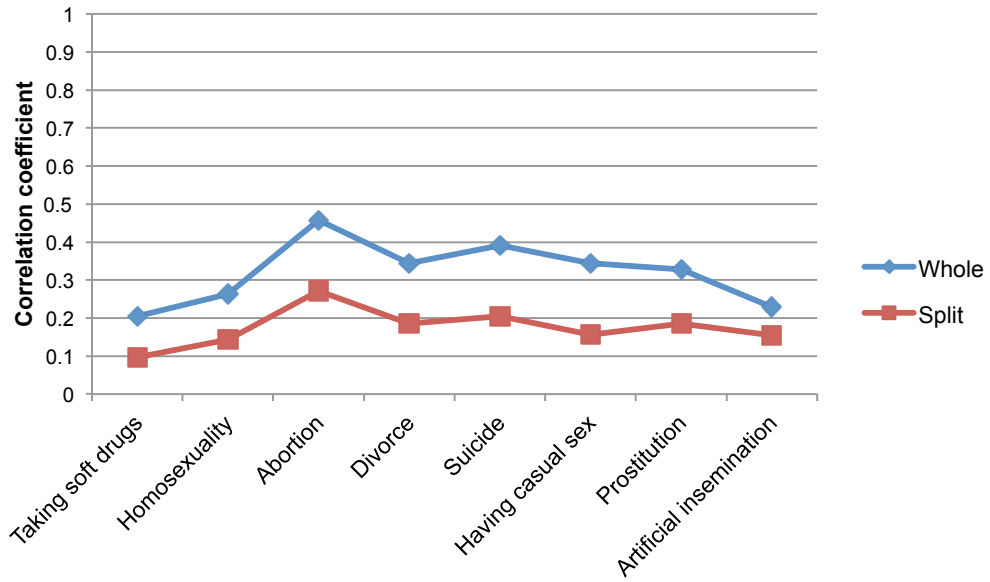


***Table 4.5**: Correlation Matrix by question battery version, 9 self-determination morality items*

Whole question battery								
Items	1	2	3	4	5	6	7	8
1- Taking soft drugs	1							
2- Homosexuality	.326	1						
3- Abortion	.320	.429	1					
4- Divorce	.334	.500	.614	1				
5- Euthanasia	.205	.263	.457	.344	1			
6- Suicide	.380	.310	.461	.434	.391	1		
7- Having casual sex	.479	.404	.449	.466	.344	.454	1	
8- Prostitution	.368	.276	.366	.374	.328	.372	.478	1
9- Artificial insemination	.199	.334	.364	.360	.229	.143	.284	.218
Split question battery								
Items	1	2	3	4	5	6	7	8
1- Taking soft drugs	1							
2- Homosexuality	.376	1						
3- Abortion	.320	.414	1					
4- Divorce	.341	.506	.540	1				
5- Euthanasia	.097	.144	.272	.186	1			
6- Suicide	.339	.347	.424	.405	.205	1		
7- Having casual sex	.473	.417	.470	.478	.156	.402	1	
8- Prostitution	.349	.239	.393	.354	.186	.406	.476	1
9- Artificial insemination	.205	.318	.362	.373	.155	.216	.316	.194

***Table 4.6**: Correlation Matrix by question battery version, 6 civic morality items*

Whole question battery					
Items	1	2	3	4	5
1- State benefits	1				
2- Cheating on tax	.492	1			
3- Accepting a bribe	.408	.471	1		
4- Avoiding a fare	.318	.388	.388	1	
5- Political violence	.256	.292	.414	.312	1
6- Death penalty	.160	.212	.218	.110	.247
Split question battery					
Items	1	2	3	4	5
1- State benefits	1				
2- Cheating on tax	.377	1			
3- Accepting a bribe	.388	.556	1		
4- Avoiding a fare	.313	.380	.375	1	
5- Political violence	.272	.342	.438	.332	1
6- Death penalty	.106	.230	.241	.137	.228

Among the items measuring civic morality, correlations between the items remained broadly similar across battery versions, although some differences were however observable. A higher correlation between the items of cheating on taxes and accepting a bribe could be observed in the split version. In the split version these items were asked together, whereas in the whole battery version the item of taking soft drugs was asked in between. Concerning the item of political violence, although the item showed slightly higher correlations with the other civic morality items in the split version, these differences were almost imperceptible.

Table 4.7: *Principal Components Analysis (PCA) by question battery version*

	Whole question battery				Split question battery			
		Factor loadings[a]				Factor loadings[a]		
Items	**Communalities**	**Factor 1**	**Factor 2**	**Factor 3**	**Communalities**	**Factor 1**	**Factor 2**	**Factor 3**
Abortion	.60	.8			.58	.7		
Divorce	.59	.8			.58	.8		
Having casual sex	.58	.7			.60	.8		
Homosexuality	.60	.7		-.3	.51	.7		
Suicide	.46	.7			.44	.6		
Prostitution	.43	.6			.40	.6		
Euthanasia	.55	.6		.5	.58			.7
Taking soft drugs	.59	.6	.4	-.3	.52	.6		
Artificial Insemination	.26	.5			.29	.5		
Accepting a bribe	.57		.8		.63		.8	
Cheating on tax	.57		.7		.60		.8	
Avoiding a fare	.58		.7		.55		.7	
Claiming state benefits	.48		.7		.41		.6	
Political violence	.40		.6		.44		.6	
Death penalty	.69			.8	.63			.7
Explained variance (Eigenvalue)		26.8% (4.2)	18.0% (2.7)	8.3% (1.1)		25.1% (3.9)	17.5% (2.6)	9.0% (1.2)
Total variance explained	53.1%				51.6%			

***Notes*:**

a: Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. Factor loadings below .3 are not displayed.

Principal Component Analysis. Next, results from the Principal Component Analysis, conducted separately for each question battery version, are explained (see table 4.7). The rotated solution for the split version presents each moral dimension with a better definition. Three clear components are identified, with each item loading only on one factor, contrarily to the whole battery version solution, where some cross-loadings are visible. Nevertheless, these factors have lower eigenvalues, and explain an inferior percentage of the total variance. Moreover, in the split version, the item of euthanasia no longer loaded in the same factor with others self-determination items. Although the split solution presents a higher communality for euthanasia, these would contribute mainly to factor 3.

Reliability tests were conducted for each moral dimension, separately for the whole and split question battery versions (see table 4.8). First, the 9 items of the self-determination morality dimension were analyzed. The Cronbach's Alpha for this dimension was higher in the whole battery version (.84 compared to .81 of the split version). The item of artificial fertilization does not contribute to the solution in the whole battery version, because if this item would have been deleted from the analysis, the Cronbach's Alpha would have remained the same. In the split battery solution, the item of euthanasia diminished the reliability of the scale. Indeed, the Cronbach's Alpha increased to .83 after removal of this particular item from the analysis. Nevertheless, it continued to be lower than the coefficient alpha from the whole version.

The reliability test showed similar solutions for the other morality dimension: the Cronbach's Alpha of the civic morality items was .70 in both cases, when solutions are rounded in two decimals. The test presented in both cases the item of death penalty as undermining the scale. Indeed, the Cronbach's Alpha increased to .74 in both battery versions, after removal from this item from the analysis.

Results about the consistency of the civic morality dimension coincide with what it was ***hypothesize in H3a***, this is that the reliability of the scale stay comparable across battery versions. Findings about the consistency of the self-determination morality did not support the hypothesis H3a, given that the split battery question presented lower reliability of this scale.

Multi-group confirmatory factor analysis. The last step was conducting a multi-group confirmatory factor analysis (MCFA), in order to test whether the factor structure of the latent concept of each morality dimension was equivalent across question battery versions. Each dimension was tested separately. First, the factor structure of the self-determination morality was tested. The model postulated the latent concept of self-determination morality as predictor of the 9 items. The tested model presented an overall CFI of .924 (see table 4.9). The model showed slightly better goodness of fit for the split battery version (CFI of .930 compared to .911). The scalar invariance did not hold. A partial scalar invariance was found, by freeing the constraint on the intercept of the item euthanasia. The final model was of strict invariance, with no constraint for the euthanasia intercept, but postulating equality of scale means. This final model hold (CFI: .913). So, ***hypothesis H3b*** was rejected for this morality dimension, since the model broke when the constraint for equality of factor intercepts was added. ***Hypothesis H3c*** was confirmed for this multi-item scale, this is that scale means stayed comparable across battery versions.

The model for the civic morality dimension only included 5 items. The item of death penalty was excluded from the analysis, given the results obtained in the PCA and reliability test, indicating that this item did not belong to the same dimension. The tested model presented an overall CFI of .981 (see table 4.9). Again, the model showed slightly better goodness of fit for the split battery version (CFI of .985 compared to .965). The Configural, metric and scalar invariance did hold (see table 4.9). However strict factorial invariance did not hold. This means that the residuals of the observed variables were not equal across groups. The error terms of both the item of political violence and of accepting a bribe had to be freed for the model to be invariant in both battery versions. The difference was that the variance for those items was larger in the whole question battery version. The final model was of strict invariance, with no constraint for the error terms of the variables political violence and accepting a bribe, but with constraint of equality of scale means. This model hold (CFI: .966). So, ***hypothesis H3b*** was rejected for this morality dimension, since the model broke when the constraint for equality of error terms was added. ***Hypothesis H3c*** was confirmed for this multi-item scale.

Table 4.8: *Reliability analysis of self-determination morality and civic morality, by question battery version*

	Whole		Split	
	Cronbach's Alpha	N (% valid)	Cronbach's Alpha	N (% valid)
Self-determination morality	.835	956 (94.7)	.813	1472 (94.5)
If item deleted:				
Taking soft drugs	.824		.797	
Homosexuality	.820		.791	
Abortion	.805		.779	
Divorce	.808		.782	
Euthanasia	.826		.825	
Suicide	.817		.790	
Having casual sex	.806		.777	
Prostitution	.820		.794	
Artificial insemination	.835		.806	
Civic morality	.703	979 (96.9)	.697	1512 (97.0)
If item deleted:				
State benefits	.651		.663	
Cheating on tax	.628		.615	
Accepting a bribe	.634		.619	
Avoiding a fare	.666		.654	
Political violence	.662		.649	
Death penalty	.744		.744	

Table 4.9: *Fit indices for models testing measurement equivalence of morality dimensions across question battery versions*

Latent concept	Model	X^2	df	CFI[a]	RMSEA[a]	SMRM[a]	Δ CFI[a b]
Self-determination morality	Overall Fit	481.489	27	.924	.083	.040	
	Fit by groups:						
	Whole question battery	254.150	27	.911	.094	.045	
	Split question battery	269.832	27	.930	.078	.039	
	Equivalence testing:						
	Configural	523.982	54	.922	.085	.041	
	Metric	556.561	62	.918	.081	.049	-.004
	Scalar	672.127	70	.900	.084	.056	-.018
	Partial Scalar (free: intercept of euthanasia)	574.423	69	.916	.078	.050	-.002
	Partial Strict (free: intercept of euthanasia)	599.918	78	.913	.074	.050	-.003
	Partial Strict, equal scale means (free: intercept of euthanasia)	603.570	79	.913	.074	.051	-.000
Civic morality	Overall Fit	53.848	5	.981	.063	.020	
	Fit by groups:						
	Whole question battery	40.68	5	.965	.085	.029	
	Split question battery	28.584	5	.985	.056	.019	
	Equivalence testing:						
	Configural	69.263	10	.978	.069	.023	
	Metric	80.393	14	.975	.062	.029	-.003
	Scalar	96.121	18	.970	.059	.032	-.005
	Strict	181.655	23	.940	.074	.049	-.030
	Partial strict (free: residuals of political violence & accepting a bribe)	106.742	21	.967	.057	.038	-.003
	Partial Strict, equal scale means (free: residuals of political violence & accepting a bribe)	112.183	22	.966	.057	.042	-.001

Notes:
a: CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.
b: Difference compared to the previous, more relaxed level of equivalence.

5. DISCUSSION

In this section the findings of this master thesis will be discussed. This project studied how changing the items presented in a question battery may affect the measurements produced by it. The aim was to help refining strategies of splitting a questionnaire in the implementation of modular questionnaire design.

RQ1: effects of splitting at the item level. The first objective was to identify changes in the measurements at the item-level, due to question battery version. For that, the estimates of the mean were compared across the two versions of the battery on moral beliefs. Out of the fifteen items, four of them were found to have significantly different estimates of the mean across battery versions: taking soft drugs, euthanasia, suicide and political violence.

As expected, the measurements of justifiability of taking soft drugs were different depending on the context in which this item was presented in the question battery. Whereas in the whole version this item was preceded by two items from the civic morality dimension, in the split version this item appeared first in the question battery. This result is consistent with previous studies having showed that, when presented first in a battery, items are answered differently (e.g. Carpenter & Blackwood, 1979). This is what Schuman and Presser (1981) would call an initial frame of reference effect. In this case, when appearing at the head of the list – i.e. in a noncomparative context (Moore, 2002) – the item of taking soft drugs obtained higher scores. The comparative context characterized by the presence of civic morality items previous to the question of taking soft drugs would make scores of this item to shift toward the “never justified” pole.

Also as expected, the measurements of justifiability of euthanasia were different across question battery versions. Whereas in the whole battery this item was preceded by a number of items from the same morality dimension (self-determination), in the split version the item of euthanasia was preceded only by items of the civic morality dimension. When preceded by items from the same moral dimension (self-determination), the mean of euthanasia was higher. This context effect was greater than the one found in the measurements of justifiability of taking soft drugs. Such big difference in the measurements of justifiability of euthanasia could be explained by two effects acting simultaneously. On the one hand, a part-part assimilation effect caused by the presence of previous topic-related items seems to be introduced in the whole battery version. On the other hand, it could be expected that, when asked after items from the civic morality dimension as in the split version,

a shift toward the “never justified” pole could be introduced, just as the effect observed in the item of taking soft drugs.

The item of suicide showed also different means depending on the battery version. In the whole version this item was asked after the item of euthanasia. On the contrary, in the split version the item of euthanasia was absent. It would not be too audacious to imagine that the combination of euthanasia and suicide questions asked one after the other could introduce something close to a subtraction effect; especially, if the question of “Euthanasia (terminating the life of the incurably sick)” was asked first. Similarly to the example of the abortion context effect (see sections 2.2.1 and 2.3.2), giving that these two questions address the common issue of ending life, the appearance of the question of euthanasia previous to the question of suicide could in a way reinforce the idea that the item of suicide refers to terminating life due to causes *other than* sickness. This would explain why, when asked after the item of euthanasia, the item of suicide got lower rating scores than when the item of euthanasia was absent.

The item of political violence presented also context effects. In the whole battery this item was asked after an item of the self-determination morality (artificial insemination). In the split version, in contrast, the question of political violence was preceded by another item of the same moral dimension (the item avoiding a faire on public transport, of the civic morality dimension). The mean of the item of political violence was lower when the previous question was of the civic morality dimension. This effect is interpreted as a result from a shift in the frame of reference.

RQ2: Moderators of item-level effects. The second question of this thesis was whether effects at the item level were moderated by characteristics of respondents. Findings suggest that the response is yes. The age group moderated the context effects of the items of taking soft drugs and of political violence in the same way. The youngest, who gave the highest scores to those items, were those who presented the biggest context effects. On the contrary, the oldest, who gave lower ratings to these items, did not present significant differences across battery versions.

The context effects found in the items of euthanasia and of suicide were moderated by the level of religiousness. The least religious gave the highest scores in those two items. However, whereas for the item of euthanasia they were the most affected by the context effect, in the case of the item of suicide this subgroup was not influenced by the context change. The most religious, on the contrary, were those giving the lowest scores to both items. However, while the context effect in the item of euthanasia was not significant for this subgroup, they were the ones that

presented the biggest differences across battery versions for the item of suicide. Such findings support the interpretation offered in previous paragraphs to these context effects. The fact that the most religious had presented the highest context effects in the item of suicide fits well the interpretation of that context effect as a subtraction effect, because it can be expected that highly religious people would be more sensible towards the issue of ending life that other subgroups. It is interesting that such interpretation implies that the context effect would arise among people with stronger attitudes (towards ending life in this case), precisely because of their strong attitudes. Such an explanation would go against the general assumption in the field that context effects would be less prone to occur among persons with strong attitudes (Krosnick & Schuman, 1988).

The analysis of context effects by subgroups showed results rather coherent with the interpretation given to the effects. This analysis provided thus with more insights to understand the mechanisms of the context effects. At the same time, it reinforced the evidence that effects due to context did arise.

RQ3: Effects of the splitting on multi-item measurements. The third question was whether context effects at the item-level were pervasive enough to impact measurements at the scale level. Although the scale means of both morality dimensions were found to be the same, the context effect on the euthanasia item undermined the internal consistency of the scale of self-determination morality.

The analysis of measurement invariance showed that this effect also affected the factor structure of the latent concept. The fact that the metric invariance hold, but that the model broke when testing scalar invariance suggests that the context effect found on the item of euthanasia was a directional effect in terms of Tourangeau et al. (2000). Indeed, the factor loading of this item on latent concept of self-determination morality was still similar across battery versions. It was the difference in the factor intercept what made the model broke. The factor structure of civic morality was also affected, but due to differences in the error terms of two variables.

Can a question battery be split and still produce the same measurements? The strategy used for splitting the questionnaire into modules in the case of the Swiss EVS 2017 followed a between-block design, such as proposed by Adigüzel and Wedel (2008). Items from each moral dimension were considered as two different blocks, because they were related to different constructs. This is how the question battery ended up being split in two. However, what are the consequences of splitting a question battery in which the items of different constructs are originally intermixed? Should not a question battery be rather considered as a block by itself?

Findings of this project point out to the impact the intermixing of items from different dimensions on a same question battery may have on the rating of items. Question batteries sharing a same rating scale introduce a comparative setting: previous items act as anchors, as standards of comparison at the moment of selecting a score on the rating scale (Sudman et al., 1996). The context effect found in the item of political violence show how responses to an item can shift depending on whether the preceding item belongs to the same construct, or to a different one.

The biggest context effect found in this project was however the one in the item of euthanasia. Both this effect and simultaneously the effect on the item of suicide could have been avoided, if the item of euthanasia had been allocated to the same block as the other items of the self-determination morality dimension, something that would have been consistent with the between-block design. If the context effects of the items of euthanasia and suicide are left aside, the split of a question battery containing fifteen items into two thematic blocks seems to have introduced context effects to only two items: the one of political violence and the one of taking soft drugs.

The administration of the question battery split by morality dimension did not increased the internal consistency of constructs. Gehlbach and Barge (2012) had draw attention of the risks of artificially increasing construct consistency when items measuring a same construct are regrouped together in a questionnaire. The battery split did not affect the scale means either. Moreover, apart from the effect of the item of euthanasia, the factor structures remained comparable in terms of structure, loadings and intercepts for both morality dimensions.

Based on the findings of this project, an answered is offered to the question of whether a question battery can be split and still produce the same measurements. Splitting a question battery is likely to introduce differences in the measurement of one particular item: the one that becomes first of the list, as was the case of the item of taking soft drugs. Other effects may be introduced as well, depending on the question battery content. Sizes of these effects at the item-level as well as at the multi-item level are to be relativized.

Limitations. The present study studied the impact of splitting a question battery on the measurements produced by it. The entire analysis was focused on identifying effects of context, i.e. changes in the answers given to questions because of the particular content of preceding questions. However, changes in the order of questions may also lead to other type of effects, that were not tested in this thesis. This other effects may be independent of the actual content of previous questions,

but rather a consequence of respondent fatigue. This is what Schuman and Presser (1981) call effects of sequence. The theory of satisficing of Krosnick (1991) postulates that as respondents progress in the questionnaire, they become increasingly prone to simplify the answering process and to give suboptimal answers. An example of this, particularly prone to arise in question battery settings is the case of respondents giving the same score to all items, independently of their content (non-differentiation). It is possible that measurements of the moral beliefs items were also affected by differences in the sequence effects the whole and the split battery versions may have introduced. In other words, it is possible that the measurement differences that were identified in this thesis, and that were interpreted as context effects, were in reality consequences of differences in the satisficing strategies employed by respondents.

When the frequencies distributions of items were compared across battery versions, it was revealed that the whole battery version presented higher frequencies in the rating point 5. This would suggest that respondents to the whole battery version incurred to satisficing strategies by using mid-scale points more frequently than respondents to the split battery version. This would not be surprising, given that the whole question battery contained twice as many items than the split battery. It is worthy to note that the item of political violence presented significant differences in the frequencies of the rating point 5, something that again is not surprising, given that this item appeared almost at the end of the battery. There could be however another explanation for respondents of the split battery showing a lower tendency to satisfice. A possible justification for these differences could be the fact that respondents to the split version were respondents that participated – voluntarily – to both the first session and the follow-up of the survey (see beginning of section 3.4). It can be expected that this group of respondents were particularly motivated by the survey, and thus less prone to incur in satisficing strategies. Anyway, It is not clear whether the measurement differences in the item of political violence found across battery versions were due to a change in context or to the fact that the battery was shorter.

In order to test for sequence effects a better control of the question order would have been necessary. In this thesis the method employed was to compare measurements depending of the question battery version. So, data were regrouped in only two groups. However, these data corresponded to actually eight different questionnaire versions. In each questionnaire version, the moral beliefs question battery was presented in different positions. So, the number of preceding items, prior to the appearance of this question battery, was not the same for all questionnaire

versions. Further analysis of sequence effects would need to take all these differences into consideration.

This simplification of reducing the eight subgroups into only two groups of comparison – whole and split – implicated also a limitation to the study of context effects. The influence of the items that preceded the question battery was not considered in my analysis. As pointed out by Schuman (1992), earlier items of the questionnaire may have an impact on subsequent items, even if questions appear in different sections of the questionnaire. Further research could analyze differences of context among the eight distinct questionnaire versions.

The analysis of moderators of context effects presented also important limitations. The method that was chosen of comparing effects by subgroups had the limitation that it did not control for other possible differences that respondents of the different categories may have. For example, when comparing responses to the item of euthanasia across the three subgroups of level of religiosity, it was not possible to know whether the category of most religious people answered differently only because of their level of religiosity or if it was also because that subgroup presented an overrepresentation of, for example, older people. An alternative method to conduct this analysis would have been by means of interaction effects within a regression analysis. Such method would have allowed to control for differences in more variables across the different categories of the moderator variables.

The analysis of the multi-item measurement was determined by the nature of the question battery. The battery on moral beliefs was not designed a priori for measuring the moral dimensions of self-determination morality and of civic morality. That factor structure has been identified a posteriori (see section 3.2). Moreover, these factor structures do not include the entire set of items of the question battery, but only a selection of items (e.g. Phillips & Harding, 1985; Moors & Wennekers, 2003; Draulans & Halman, 2005; Vauclair & Fischer, 2011). This is why the categorization of all the items into either self-determination or civic morality items may have seemed to a certain extent artificial. Effects of context in other type of instruments having been designed purposely for measuring a latent construct may have a different impact on the scales measurements.

6. CONCLUSION

The split questionnaire design is becoming increasingly popular in the field as a promising solution to reduce questionnaire length, without having to renounce to the amount of information that is collected. The implementation of surveys following a split questionnaire design is only likely to proliferate in the near future. This is why this thesis had the intention to contribute in the optimization of this design. Whereas recent studies on this topic have been particularly concentrated in improving the imputation techniques used in the split questionnaire design, relatively less attention has been given to the refining of the splitting strategies.

The purpose of this study was to draw attention toward an issue that has been widely studied in the field of survey methodology in past decades, but that may be important to bring back to discussion: the issue of context effects. Splitting strategies must take into consideration the impact that previous questions may have on the way subsequent questions are interpreted and answered. The between-block design (Adigüzel & Wedel, 2008) has been identified as the strategy for splitting a questionnaire. According to this design, block of questions presented together to measure a same concept should be kept together and allocated to the same module.

The present thesis has focused in the particular case of question batteries, in which items measuring more than one construct are presented together, intermixed, sharing a common rating scale. The general question guiding this study was if a question battery can be split and still produce equivalent measurements.

In order to study this, a particular question battery on moral beliefs was analyzed. Data was taken from the Swiss EVS 2017, which had implemented an experimental design to compare the feasibility of online surveys, both on a split questionnaire design and in full-length design. Given the experiment, the question battery of moral beliefs had been administered in two versions: one containing all items together (whole) and another presenting items in two short batteries, separated thematically (split), according to the two moral dimensions embedded in the battery, self-determination and civic morality.

Three research questions were addressed: first, what where the effects of splitting the question battery on item level measurements; second, whether these effects were moderated by characteristics of respondents; and third, what were the effects of the splitting on the multi-item measurements.

Out of the fifteen items embedded in the battery, four presented significantly different means across the two battery versions. A first difference was found on the item that, due to the split, passed from appearing in the third position of the question

battery to be presented in the first position of one of the new short lists. Introducing such an effects seems almost unavoidable when splitting a question battery, because at the moment of creating shorter lists, one item will pass from a comparative towards a non-comparative context. This was catalogued as a initial frame of reference effect. Two other effects were attributed to the fact of having allocated a particular item from one moral dimension into the wrong thematic module. Finally, a last effect was found attributed to a shift in the frame of reference, as consequence of changing the intermixed context of items of different dimensions to a single dimension context.

The effects were then analyzed by subgroups of respondents, in order to determine if effects were moderated by characteristics of respondents. Findings showed how effects on self-determination morality items were moderated by the level of religiousness of respondents. Findings suggest that effects are not necessarily diminished when attitudes of the respondent are stronger. An interesting effect arose from the combination of two questions asked together, on euthanasia and suicide. It would seem that the context effect for this question combination was introduced among those having the most religious people, who showed the most negative attitudes towards these items.

Context effects did not affect the scale means of both moral dimensions. Internal consistency of the self-determination morality was undermined by the effects presented in the one item that had been allocated to the wrong thematic block. This same effect also undermined the scalar invariance of the factor structure of self-determination morality. Effects among items of the civic morality dimension compromised the strict invariance of that factor.

In general terms, the findings of this thesis would suggest that splitting a question battery thematically does not increase artificially the constructs reliability. The take-home message is that, if such between-block design is implemented, special attention should be taken to allocate items to the most appropriate thematic block. Otherwise pervasive effects of context may be introduced.

In this thesis effects of context were studied in a very specific question battery. Further research should be done on this topic, in order to compare results from different question settings.

7. REFERENCES

Adigüzel, F., & Wedel, M. (2008). Split Questionnaire Design for Massive Surveys. *Journal of Marketing Research*, *45*(5), 608–617.

Bassili, J., & Krosnick, J. (2000). Do Strength-Related Attitude Properties Determine Susceptibility to Response Effects? New Evidence from Response Latency, Attitude Extremity, and Aggregate Indices. *Political Psychology*, 21(1), 107-132.

Carpenter, E. H., & Blackwood, L. G. (1979). The effect of question position on responses to attitudinal questions. *Rural Sociology*, 44, 46-72.

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464-504.

Crespi, I., & Morris, D. (1984). Question order effect and the measurement of candidate preference in the 1982 Connecticut elections. *Public Opinion Quarterly*, 48, 578-591.

Desai, S., & Braitman, K. A. (2005). The Effects of Scale Carving on Instruments Assessing Violence. *Journal of Family Violence*, 20(2), 101-107.

Dillman, D., Smyth, J., & Christian, L. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method* (3rd ed.). Hoboken: John Wiley & Sons.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method* (4th ed.). Hoboken, NJ: John Wiley & Sons.

Draulans, V., & Halman, L. (2005). Mapping Contemporary Europe's Moral and Religious Pluralist Landscape: An Analysis Based on the Most Recent European Values Study Data. *Journal of Contemporary Religion*, 20(2), 179-193.

Ernst Stähli, M., Joye, D., Pollien, A., Ochsner, M., Milbert, P., Nisple, K., & Sapin, M. (To be published). *European Values Study (EVS) 2017. Documentation of the survey field (CAPI and CAWI/MAIL).* Lausanne: FORS.

Feldman, J. M., & Lynch, J. G. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology*, 73(3), 421-435.

Freitag, M., & Bauer, P. C. (2013). Testing for Measurement Equivalence in Surveys: Dimensions of Social Trust across Cultural Contexts. *Public Opinion Quarterly*, 77(S1), 24-44.

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

Grice, H. (1975). Logic and conversation. In P. Cole & T. Morgan (Eds.), *Syntax and semantics: Vol. 3. Speech acts* (pp. 41-58). New York: Seminar Press.

Halman, L. (2009). Value change in Western European societies: Results from the European values study. *Bulletin of the Faculty of Social Studies, Kwansei Gakuin University*, 107, 35-48.

Harding, S., Phillips, D. R., Fogarty, M. P., & European Value Systems Study Group. (1986). *Contrasting values in Western Europe: Unity, diversity and change*. Basingstoke: Macmillan in association with the European Value Systems Study Group.

Harrison, D. A., & McLaughlin, M. E. (1993). Cognitive processes in self-report responses: Tests of item context effects in work attitude measures. *Journal of Applied Psychology,* 78(1), 129-140.

Hox, J. J., De Leeuw, E. D., & Zijlmans, E. A. O. (2015). Measurement equivalence in mixed mode surveys. *Frontiers in Psychology*, 6:87.

Hyman, H. H., & Sheatsley, P. B. (1950). The current status of American public opinion. In J. C. Payne (Ed.), *The teaching of contemporary affairs* (pp. 11-34). New York: National Education Association.

Klausch, T., Hox, J. J., & Schouten, B. (2013). Measurement Effects of Survey Mode on the Equivalence of Attitudinal Rating Scale Questions. *Sociological Methods & Research*, 42(3), 227-263.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology*, 55(2), 312-320.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5: 213-236.

Krosnick, J. A., Boninger, D. S., Chuang, Y. C., Berent, M. K., & Carnot, C. G. (1993). Attitude strength: One construct or many related constructs? *Journal of Personality and Social Psychology*, 65(6), 1132-1151.

Krosnick, J. A., & Petty, R. E. (1995). Attitude strength: An overview. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 1-24). Mahwah, NJ: Erlbaum.

Krosnick, J. A., & Schuman, H. (1988). Attitude intensity, importance, and certainty and susceptibility to response effects. *Journal of Personality and Social Psychology*, 54(6), 940-952.

Lavine, H., Huff, J. W., Wagner, S. H., & Sweeney, D. (1998). The moderating influence of attitude strength on the susceptibility to context effects in attitude surveys. *Journal of Personality and Social Psychology*, 75(2), 359-373.

Lavrakas, P. J. (2008). *Encyclopaedia of survey research methods*. Thousand Oaks, CA: SAGE Publications.

McGuire, W. J. (1969). The nature of attitudes and attitude change. In G. Lindzey, & E. Aronson (Eds.), *The handbook of social psychology* (2nd ed.), 3. Reading. Mass.: Addison-Wesley.

Moore, D. M. (2002). Measuring new types of question-order effects. Additive and subtractive. *Public Opinion Quarterly*, 66, 80-91.

Moors, G., & Wennekers, C. (2003). Comparing Moral Values in Western European Countries between 1981 and 1999. A Multiple Group Latent-Class Factor Approach. *International Journal of Comparative Sociology*, 44(2), 155-172.

Phillips D., & Harding, S. (1985). The Structure of Moral Values. In M. Abrams, D. Gerard, & N. Timms (Eds.), *Values and Social Change in Britain. Studies in the Contemporary Values of Modern Society* (pp. 93-108). London: Palgrave Macmillan.

Poznyak, D., Meuleman, B., Abts, K., & Bishop, G. F. (2014). Trust in American Government: Longitudinal Measurement Equivalence in the ANES, 1964–2008. *Social Indicators Research*, 118(2), 741-758.

Raghunathan, T. E., & Grizzle, J. E. (1995). A Split Questionnaire Survey Design. *Journal of the American Statistical Association*, 90, 54-63.

Rässler, S., Koller, F., & Mäenpää, C. (2002). A split questionnaire survey design applied to German media and consumer surveys. *Proceedings of the International Conference on Improving Surveys*, ICIS 2002, Copenhagen.

Roberts, C., Lipps, O., & Kissau, K. (2013). Using the Swiss population register for research into survey methodology. *FORS Working Paper Series*, paper 2013-1. Lausanne: FORS.

Rothschild, D., & Malhotra, N. (2014). Are public opinion polls self-fulfilling prophecies? *Research & Politics*, 1(2), 1-10.

Schuman, H. (1992). Context effects: State of the past/state of the art. In N. Schwarz, & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 5-20). New York, NY: Springer-Verlag.

Schuman, H. (2009). Context effects and social change. *Public Opinion Quarterly*, 73(1), 172-179.

Schuman, H., & Duncan, O. C. (1974). Questions about attitude survey questions. *Sociological Methodology*, 5, 232-251.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. New York: Academic Press.

Schuman, H., Presser, S., & Ludwig, J. (1981). Context effects on survey responses to questions about abortion. *Public Opinion Quarterly*, 45(2), 216-223.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54(2), 93-105.

Schwarz, N., & Bless, H. (1992). Constructing reality and its alternatives: assimilation and contrast effects in social judgment. In L.L. Martin, & A. Tesser (Eds.), *The construction of social judgment* (pp. 217-245). Hillsdale, NJ: Erlbaum.

Schwarz, N., & Hippler, H.-J. (1995). Subsequent questions may influence answers to preceding questions in mail surveys. *Public Opinion Quarterly*, 59(1), 93-97.

Schwarz, N., & Strack, F. (1991). Context effects in attitude surveys: Applying cognitive theory to social research. *European Review of Social Psycholoy*, 2, 31-50.

Schwarz, N., Strack, F., & Mai, H.P. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly*, 5, 3-23.

Schwarz, N., & Wyer, R. S., Jr. (1985). Effects of rank ordering stimuli on magnitude ratings of these and other stimuli. *Journal of Experimental Social Psychology*, 21, 30-46.

Shoemaker, D. M. (1973). *Principles and Procedures of Multiple Matrix Sampling*. Cambridge, MA: Ballinger.

Smith, T. W. (1979). Happiness: Time trends, seasonal variations, inter-survey differences, and other mysteries. *Social Psychology Quarterly*, 42, 18-30.

Stouffer, S. A., & DeVinney, L. C. (1949). How personal adjustment varied in the army – by background characteristics of the soldiers. In S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star, & R. M. Williams, Jr. (Eds.), *The American soldier: Adjustment during army life*. Princeton, NJ: Princeton University Press.

Strack, F. (1992). 'Order effects' in survey research: Activation and information functions of preceding questions. In N. Schwarz, & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 23-34). New York, NY: Springer-Verlag.

Strack F., & Martin L.L. (1987). Thinking, Judging, and Communicating: A Process Account of Context Effects in Attitude Surveys. In H.-J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social Information Processing and Survey Methodology. Recent Research in Psychology.* New York, NY: Springer.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

Thomas, N., Raghunathan, T. E., Schenker, N., Katzoff, M. J., & Johnson, C. L. (2006). An Evaluation of Matrix Sampling Methods Using Data From the National Health and Nutrition Examination Survey. *Survey Methodology*, 32, 217-232.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive Processes Underlying Context Effects in Attitude Measurement. *Psychological Bulletin*, 103(2), 299-314.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Vauclair, C., & Fischer, R. (2011). Do cultural values predict individuals' moral attitudes? A cross-cultural multilevel approach. *European Journal of Social Psychology*, 41, 645-657.

Willick, D. H., & Ashley, R. K. (1971). Survey question order and the political party preferences of college students and their parents. *Public Opinion Quarterly*, 35, 189-199.

Willits, F. K., & Saltiel, J. (1995). Question Order Effects on Subjective Measures of Quality of Life. *Rural Sociology*, 60(4), 654-665.

8. APPENDIX

Table 8.1: *Response frequencies by question battery version*

	(1) Whole question battery (N=1010) Response frequencies in %											(2) Splt question battery (N=1558) Response frequencies in %											(3) Difference between whole and split question battery (2) - (1) Δ Response frequencies[a]										
Items	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**NA**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**NA**	**Δ1**	**Δ2**	**Δ3**	**Δ4**	**Δ5**	**Δ6**	**Δ7**	**Δ8**	**Δ9**	**Δ10**	**ΔNA**
1- State benefits	58.3	14.6	9.0	3.5	5.9	1.9	1.9	1.3	0.5	2.0	1.2	57.6	17.1	9.8	3.5	5.6	1.0	1.4	0.8	0.6	1.2	1.3	-0.7	2.6	0.8	0.0	-0.4	-0.9	-0.5	-0.5	0.1	-0.8	0.2
2- Cheating on tax	60.0	16.4	8.8	2.7	4.1	2.1	0.8	1.4	0.6	1.8	1.4	60.3	17.6	8.4	2.4	4.7	1.2	1.2	1.2	0.6	1.4	1.0	0.3	1.2	-0.4	-0.2	0.6	-0.9	0.4	-0.2	0.0	-0.4	-0.4
3- Taking soft drugs	37.8	8.8	8.6	6.0	13.6	4.2	5.8	4.8	2.8	6.5	1.1	28.9	10.7	9.9	5.0	11.4	6.9	8.4	9.1	3.5	5.3	0.9	-8.9 ***	1.9	1.3	-1.0	-2.2	2.8 **	2.6 *	4.3 ***	0.8	-1.3	-0.2
4- Accepting a bribe	76.6	9.7	4.1	1.3	3.2	1.1	0.6	1.0	0.1	1.3	1.1	77.6	11.0	3.3	1.5	1.9	1.0	0.6	0.6	0.4	0.8	1.2	1.0	1.3	-0.7	0.3	-1.3 *	-0.1	0.0	-0.4	0.3 IN	-0.5	0.1
5- Homosexuality	5.2	2.1	2.9	2.2	14.4	3.4	4.4	7.4	6.6	49.3	2.2	5.2	1.5	3.7	1.9	11.0	4.4	4.9	9.1	8.4	48.6	1.3	0.0	-0.5	0.9	-0.3	-3.4 **	1.1	0.5	1.6	1.8	-0.7	-0.8
6- Abortion	6.7	5.0	4.2	3.4	23.0	6.8	8.7	14.4	5.6	20.9	1.4	6.0	3.6	6.5	4.6	17.7	7.8	10.4	15.5	8.3	18.4	1.2	-0.7	-1.4	2.4 *	1.3	-5.3 ***	1.0	1.7	1.2	2.6 *	-2.5	-0.2
7- Divorce	2.2	0.7	3.1	2.6	20.0	4.7	7.5	15.0	9.2	33.8	1.4	2.0	0.8	2.3	2.9	17.2	6.7	9.1	15.1	9.1	33.8	1.0	-0.2	0.1	-0.8	0.3	-2.8	2.0 *	1.6	0.1	-0.1	0.0	-0.4
8- Euthanasia	8.5	3.8	3.6	2.4	16.3	5.4	8.2	17.9	9.3	22.8	1.8	13.4	5.6	6.1	3.5	17.6	6.3	10.8	15.1	7.3	11.9	2.4	4.9 ***	1.8 *	2.5 **	1.2	1.3	0.8	2.6 *	-2.8	-2.0	-10.9 ***	0.6
9- Suicide	23.8	8.6	8.8	5.3	21.6	4.7	7.2	7.8	3.3	6.9	2.0	18.8	9.6	10.1	4.9	18.7	7.0	7.9	8.5	3.5	9.1	1.9	-5.0 **	0.9	1.3	-0.4	-2.9	2.3 *	0.7	0.7	0.3	2.2	-0.1
10- Having casual sex	19.5	7.9	6.9	3.7	18.7	5.2	7.1	9.6	5.0	14.3	2.1	17.5	6.8	8.2	4.4	18.0	5.6	7.8	10.3	5.1	15.0	1.3	-2.0	-1.1	1.3	0.8	-0.7	0.3	0.7	0.7	0.2	0.7	-0.8
11- Avoiding a fare	48.1	18.3	11.2	6.0	6.1	2.8	2.3	2.0	0.7	1.8	0.7	47.9	18.5	12.7	6.1	6.4	1.9	2.6	1.0	0.4	1.6	0.8	-0.2	0.2	1.5	0.1	0.3	-0.9	0.3	-1.0 *	-0.2	-0.2	0.1
12- Prostitution	19.9	10.9	7.9	5.1	21.7	7.0	7.8	7.0	4.3	6.5	1.8	19.3	7.8	11.0	6.7	19.8	7.1	8.9	8.9	2.0	7.1	1.5	-0.6	-3.1 **	3.1 *	1.6	-1.9	0.0	1.1	1.8	-2.3 ***	0.5	-0.3
13- Artificial insemination	6.8	4.6	5.5	3.6	17.1	7.5	9.9	13.2	7.9	22.0	1.9	5.7	3.2	5.9	5.1	15.1	8.0	12.2	16.3	9.1	18.0	1.4	-1.1	-1.3	0.4	1.5	-2.0	0.5	2.3	3.1 *	1.2	-4.0 *	-0.5
14- Political violence	60.6	13.5	8.8	3.0	8.1	1.4	0.8	0.8	0.1	1.2	1.8	69.0	12.1	7.1	3.5	3.5	1.1	0.9	0.3	0.4	0.6	1.5	8.4 ***	-1.3	-1.7	0.6	-4.7 ***	-0.3	0.1	-0.5	0.3 IN	-0.6	-0.3
15- Death penalty	46.4	13.2	5.5	2.9	13.5	3.2	3.6	4.9	1.5	4.3	1.2	49.2	10.7	8.5	3.3	10.2	3.7	4.2	3.7	1.7	3.3	1.4	2.7	-2.5	3.0 **	0.5	-3.3 *	0.5	0.7	-1.1	0.2	-0.9	0.2

Notes:
a: Chi-square tests of independence: *** p<.001, **p<.01, *p<.05.
IN: Invalid test because of not enough cases

Table 8.2: *Differences in morality items across groups of age, education level, political orientation and religiousness level (Full-length original questionnaire FL1)*

Religiousness level

		G1 Highest (N=66)			G2 Medium (N=237)			G3 Lowest (N=93)		K-W Test		Δ Mean[b]		
Item	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	X^2	**Sig.[a]**	**G2 - G1**	**G3 - G2**	**G3 - G1**
Taking soft drugs	65	2.57	2.69	234	3.64	2.69	93	5.30	3.21	34.21	***	1.07***	1.66***	2.73***
Euthanasia	63	5.02	3.13	233	6.95	2.63	93	7.58	2.58	28.29	***	1.94***	.63*	2.56***
Suicide	65	3.15	2.51	232	4.29	2.62	92	5.73	2.88	33.91	***	1.14***	1.44***	2.57***
Political violence	64	2.06	2.26	234	1.99	1.56	93	1.89	1.79	2.71		-.08	-.09	-.17

Education level

		G1 Lowest (N=66)			G2 Medium (N=237)			G3 Highest (N=93)		K-W Test		Δ Mean[b]		
Item	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	X^2	**Sig.[a]**	**G2 - G1**	**G3 - G2**	**G3 - G1**
Taking soft drugs	72	3.14	2.81	216	3.63	2.89	97	4.95	2.95	21.81	***	.50	1.31***	1.81***
Euthanasia	70	6.69	2.78	213	6.81	2.89	97	6.82	2.77	0.19		.12	.02	.14
Suicide	70	3.86	2.66	215	4.14	2.70	96	5.43	2.75	18.42	***	.29	1.28***	1.57***
Political violence	69	1.78	1.63	214	2.02	1.80	98	1.89	1.48	1.87		.24	-.13	.11

Age

		G1 Oldest (N=66)			G2 Middle (N=237)			G3 Youngest (N=93)		K-W Test		Δ Mean[b]		
Item	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	X^2	**Sig.[a]**	**G2 - G1**	**G3 - G2**	**G3 - G1**
Taking soft drugs	90	2.27	2.10	242	4.25	2.97	67	4.64	3.25	37.82	***	1.99***	.39	2.38***
Euthanasia	90	6.18	3.02	239	6.91	2.83	65	7.11	2.56	4.66		.73	.20	.93
Suicide	89	3.82	2.77	240	4.67	2.80	66	4.32	2.72	6.56	*	.85*	-.35	.50
Political violence	89	1.64	1.47	242	1.95	1.82	64	2.56	1.64	22.03	***	.31	.62***	.92***

Political orientation

		G1 Rightest (N=66)			G2 Medium (N=237)			G3 Leftest (N=93)		K-W Test		Δ Mean[b]		
Item	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	X^2	**Sig.[a]**	**G2 - G1**	**G3 - G2**	**G3 - G1**
Taking soft drugs	64	3.72	3.18	236	3.60	2.77	90	4.73	3.21	8.16	*	-.12	1.13**	1.01*
Euthanasia	64	7.52	2.24	232	6.71	2.95	89	6.53	2.84	3.65		-.80	-.18	-.99
Suicide	64	4.88	2.98	233	4.12	2.65	89	4.98	2.96	6.90	*	-.75	.86*	.10
Political violence	64	2.34	2.15	232	1.87	1.69	91	2.01	1.49	6.38	*	-.47*	.14	-.33

Notes:

a: Kruskal Wallis Equality of Populations Rank Test (non-parametric): *** p<.001, **p<.01, *p<.05.

b: Unpaired Two-Samples Wilcoxon Test (non-parametric): *** p<.001, **p<.01, *p<.05.