mda
methods, data, analyses
Journal for Quantitative Methods and Survey Methodology
Volume 9, 2015 | 1

methods, data, analyses is published by GESIS – Leibniz Institute for the Social Sciences.



Managing editor: Sabine Häder
GESIS – Leibniz Institute for the Social Sciences
PO Box 12 21 55
68072 Mannheim
Germany
Tel.: + 49.621.1246282
E-Mail: mda@gesis.org
Internet: www.gesis.org/mda

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects. We especially invite authors to submit articles extending the profession's knowledge on the science of surveys, be it on data collection, measurement, or data analysis and statistics. We also welcome applied papers that deal with the use of quantitative methods in practice, with teaching quantitative methods, or that present the use of a particular state-of-the-art method using an example for illustration.

All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. Mda appears in two regular issues per year (June, December).

Please register for a subscription via http://www.gesis.org/en/publications/journals/mda/subscribe

Print: Verlag Pfälzische Post GmbH, Neustadt, Germany
Printed on chlorine-free paper.

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)

# Content

methods, data, analyses | Vol. 9(1), 2015, pp. 3-56 DOI: 10.12758/mda.2015.001

# Mode System Effects in an Online Panel Study: Comparing a Probability-based Online Panel with two Face-to-Face Reference Surveys

*Bella Struminskaya*[1], *Edith de Leeuw*[2] & *Lars Kaczmirek*[1]

1 *GESIS – Leibniz Institute for the Social Sciences*
2 *Department of Methodology and Statistics, Utrecht University*

**Abstract**

One of the methods for evaluating online panels in terms of data quality is comparing the estimates that the panels provide with benchmark sources. For probability-based online panels, high-quality surveys or government statistics can be used as references. If differences among the benchmark and the online panel estimates are found, these can have several causes. First, the question wordings can differ between the sources, which can lead to differences in measurement. Second, the reference and the online panel may not be comparable in terms of sample composition. Finally, since the reference estimates are usually collected face-to-face or by telephone, mode effects might be expected. In this article, we investigate mode system effects, an alternative to mode effects that does not focus solely on measurement differences between the modes, but also incorporates survey design features into the comparison. The data from a probability-based offline-recruited online panel is compared to the data from two face-to-face surveys with almost identical recruitment protocols. In the analysis, the distinction is made between factual and attitudinal questions. We report both effect sizes of the differences and significances. The results show that the online panel differs from face-to-face surveys in both attitudinal and factual measures. However, the reference surveys only differ in attitudinal measures and show no significant differences for factual questions. We attribute this to the instability of attitudes and thus show the importance of triangulation and using two surveys of the same mode for comparison.

*Keywords*: mode system effect, online panel, benchmark survey comparison, data quality

# 1 Introduction

Several large-scale panel and repeated cross-sectional surveys incorporate or are planning to incorporate an online mode for data collection to reduce costs or to maximize contact and response rates. Some panels are designed as online panels from the beginning, with an interviewer-administered panel recruitment procedure and web-based data collection for the surveys in the panel (e.g., the LISS Panel[1] in the Netherlands and the KnowledgePanel[2] of GfK Custom Research, formerly Knowledge Networks in the United States). Other panels switch from interviewer-administered to online mode for their data collection (e.g., the Netherlands Kinship Panel Study[3]), employ an additional online component (e.g., ANES 2008-2009 Panel Study[4]), or experiment with including an online mode along with interview modes (ESS experiments on mixing modes[5], Understanding Society Innovation Panel in the UK[6], and Labor Force Survey and Crime Victimization Survey in the Netherlands[7]).

Data users need to know that irrespective of the mode in which data were collected, it is possible to make valid inferences about the processes that data users study, and that panel or trend data are comparable, that is, not influenced by the mode change.

In the literature on survey data collection modes, key dimensions are discerned on which modes vary that account for differences in responses across modes. In her theoretical model on mode effects, de Leeuw (1992, 2005) identified three sets of factors that explain differences between the modes: (1) media-related factors, (2) factors that are related to information transmission, and (3) interviewer effects. Media-related factors encompass social conventions and customs associated with the media utilized in survey methods. Media-related factors include socio-cultural

1 http://www.lissdata.nl/lissdata/

2 http://www.knowledgenetworks.com/knpanel/

3 http://www.nkps.nl/NKPSEN/nkps.htm; see also Hox, de Leeuw, & Zijlmans (2015).

4 http://www.electionstudies.org/studypages/2008_2009panel/anes2008_2009panel.htm

5 http://www.europeansocialsurvey.org/methodology/mixed_mode_data_collection.html

6 https://www.understandingsociety.ac.uk/about/innovation-panel; see also Auspurg et al. (2013)

7 See Schouten et al. (2013)

*Direct correspondence to*
Bella Struminskaya, GESIS – Leibniz Institute for the Social Sciences, B 2,1,
68159 Mannheim, Germany
E-mail: bella.struminskaya@gesis.org

effects such as familiarity with a medium, patterns of use of the medium, as well as norms of social interaction (e.g., interviewers having more control over the interviewing process and pace because they initiate the interaction). Factors related to information transmission involve more technical aspects of the communication process and include the manner in which information is presented (visual presentation, auditory presentation or both visual and auditory) as well as additional cues that are pertinent to the question-answering process (such as text and lay-out in the self-administered mode, gestures and tone of the interviewer in the face-to-face mode). The third set of factors includes the influence of an interviewer on the responses provided by respondents. Interviewer-administration can influence the respondents' feelings of privacy and lessen their willingness to disclose sensitive information. On the other hand, interviewers can provide clarification or motivate respondents to provide answers. In a meta-analysis of 52 early mode comparison studies, de Leeuw (1992, chapter 3) found that face-to-face and telephone interviews did not differ in response validity through record checks and on social desirability. When both interview modes were compared with self-administered mail surveys, the meta-analysis revealed an interesting picture. It is somewhat harder to have people answer questions in a self-administered mode (both response rate and item missing rates are higher), but the resulting answers show less social desirability and more openness on sensitive topics. These results emphasize the importance of the role of the interviewer.

Groves, Fowler, Couper, Lepkowski, Singer, and Tourangeau (2009) identify five dimensions along which data collection methods differ that partly overlap with the dimensions discussed above: (1) the degree of interviewer involvement, (2) the level of interaction with the respondent, (3) the degree of privacy for the respondent, (4) the channels of communication used, and (5) the degree of technology use. Compared to the interviewer-administered methods of data collection, online surveys eliminate interviewer involvement and offer a high level of privacy, and impose a low cognitive burden, because the questions can be easily reread on the computer screen (Tourangeau, Conrad, & Couper, 2013). In their meta-analysis of studies employing randomized experiments to compare different modes, Tourangeau et al. (2013) conclude that compared to interviewer-administered surveys, online surveys yield more reports of sensitive information and that compared to paper-and-pencil surveys, only small advantages are offered by online surveys. This finding is consistent with the finding of Klausch, Hox, & Schouten (2013), who find measurement differences between interviewer- and self-administration that are absent when comparing face-to-face to the telephone mode and mail to the online mode. It is also in line with the results of the early meta-analysis by de Leeuw (1992), and suggests a dichotomy in modes with and modes without an interviewer.

There are different ways to study mode effects (Groves, 1989). A common form is a field experiment, in which respondents are randomly assigned to a specific

mode (see also the meta-analyses by de Leeuw, 1992 and Tourangeau et al, 2013). These studies focus on a particular source or error (e.g., measurement error), investigate a particular mechanism that produces a mode difference (e.g., social desirability), and aim at estimating pure mode effects (Biemer & Lyberg, 2003, p. 207; Couper, 2011, p. 894-897).

However, in daily practice, associated with each mode is a set of decisions intended to take advantage of the benefits of each mode (Lyberg & Kasprzyk, 1991, p. 249) and in addition to mode other factors will vary, such as number of calls versus number of reminders in interview vs. web modes. Consequently, a good alternative is to measure a mode system effect, that is, to compare whole systems of data collection developed for different modes. A data collection system is defined as an "entire data collection process designed around a specific mode" (Biemer & Lyberg, 2003, p. 208; Biemer, 1988). A small number of studies focuses on the outcomes, examining total survey systems, where the final survey estimates are compared (Couper, 2011), and these studies are especially important for practitioners who want to switch from one mode of data collection to another (e.g, from interview to online survey) or employ a mixed-mode system for data collection.

In this study, we investigate mode system effects of three data collection systems, comparing the data from an online panel to two face-to-face reference surveys. Central is the question whether the different systems produce equivalent results given all the differences in data collection. Data from surveys that are carried out in different modes may differ for three reasons (de Leeuw & Hox, 2011, p. 53). Firstly, differences may be caused by the implementation of different question formats in different modes (question effects), secondly, different modes may lead to a different sample composition (selection effects), and, thirdly, the modes themselves may lead to the different response processes (mode effect).

In order to systematically investigate mode sysytem effects, we focus our comparison on questions with the same question wording and control step-by-step for differences in sample composition. As reference we use data from two face-to-face interview surveys, whose recruitment and administrative procedures are similar. Based on the mode comparison studies and meta-analyses cited above, we do not expect differences in the estimates based on the face-to-face interviews after adjusting for potential sample composition differences. However, we do expect differences in estimates based on data collected online compared to the estimates based on the reference interview surveys (Hypothesis 1).

We concentrate on two types of questions: factual and attitudinal questions. Couper (2011, p. 896) argues that factual questions and nonsensitive questions with only yes/no responses are not likely to be affected by the presence of an interviewer, or by visual or aural presentation. Questions about attitudes and opinions place different demands on respondents' cognitive processes than factual and behavioral questions do. Attitude is defined as a collection of feelings, beliefs, and knowledge

about an issue – considerations that have different levels of accessibility. While forming an answer, respondents process these considerations, which requires deliberation and effort (Tourangeau, Rips, & Rasinski, 2000, pp. 179-180). In a face-to-face survey, an interviewer initiates the interaction and thereby controls the pace of the interview (de Leeuw, 2005), so the time might not be sufficient for a respondent to process the considerations needed to generate responses to an attitudinal question. In a self-administered online survey, by contrast, the respondent controls the survey situation, including the pacing, which allows for taking more time if needed to answer an attitudinal question. Furthermore, according to Tourangeau, Conrad, and Couper (2013), online surveys impose a lower cognitive burden on the respondent due to the visual presentation of information, allowing respondents to consider the question and response alternatives better. In an experimental study specifically designed to evaluate mode effects in which respondents were randomly assigned to mail, web, telephone and face-to-face conditions in the Crime Victimization Survey of the general population in the Netherlands, Klausch, Hox, and Schouten (2014) indeed find measurement differences for attitudinal variables such as questions about the social quality and problems of the neighborhood, and no measurement differences for two factual variables on victimization.

We therefore expect that mode system effects between online and face-to-face interviews are more pronounced in attitudinal questions than in factual questions (Hypothesis 2).

## 2 Data

We use data from the GESIS Online Panel Pilot[8], a probability-based online panel of Internet users in Germany; we use two cross-sections from the German General Social Survey (ALLBUS 2010 and ALBUSS 2012) as reference surveys. Table 1 contains an overview of design characteristics for the GOPP and the reference surveys.

*GESIS Online Panel Pilot (GOPP)*

The GESIS Online Panel Pilot (GOPP) is a telephone-recruited online panel of Internet users aged 18 and older who live in private households in Germany. To recruit participants, the randomized last digit method was used, which is a variation of a random digit dialing (RDD) for Germany (Gabler & Häder, 2002). A dual-frame approach was used: the samples were drawn independently from landline and mobile phone frames, aiming at a final sample with 50% eligible landline numbers and 50% eligible mobile phone numbers. In order to handle the overlap between the

8 https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5582&db=e&doi=10.4232/1.11570

two frames, the inclusion probabilities for target persons were calculated according to the formulas of Siegfried Gabler, Sabine Häder and their colleagues under the assumption of independence of the two samples (Gabler, Häder, Lehnhoff, & Mardian, 2012). The inclusion probabilities account for the sample sizes and frame sizes of both landline and mobile phone components, as well as for the number of landline and mobile phone numbers at which a respondent can be reached. For the landline component, household size was also included in the calculation. The target person, that is, the person with the most recent birthday, provided this information. For the mobile phone component, no selection procedures were implemented, since mobile phone sharing in Germany is approximately 2% (Gabler et al. 2012).

The recruitment took place in three sequential study parts using almost identical recruitment protocols. Recruitment periods were in February 2011, in June-July 2011, and in July-August 2011. After a short telephone interview, respondents were asked to provide their email addresses in order to join an online panel. Respondents who agreed would then be sent email invitations to online surveys of 10-15 minutes every month for eight months in total. Prospective panel members were offered incentives of 0, 2, 5, or 10 Euros, which were varied experimentally. An additional bonus of 20 Euros for completing all eight online questionnaires was offered.

In the GESIS Online Panel Pilot questionnaires, a number of questions were replicated from other well-known social surveys. The first goal of such replication was to assess the feasibility of surveying rather complex constructs in an online setting. The second goal was to study data quality by comparing online estimates to external benchmarks. A substantial part of questions originated from the German General Social Survey ("ALLBUS") 2010. The original question wordings were retained, except for the cases when an adjustment was needed to make questions suitable for a self-administered mode. It is furthermore important to note that every questionnaire in the GOPP had a leading topic and most questions were asked once. Questions for analysis were selected from questionnaires (waves) 1, 2, 3, 4, 5, and 7.

*German General Social Survey "ALLBUS"*

The German General Social Survey is a general population survey on attitudes, behavior, and social structure in Germany. It has been conducted by GESIS biannually since 1980. The survey mode is face-to-face interviewing. For our analyses, we use ALLBUS 2010[9] and ALLBUS 2012[10], which both implemented a two-stage disproportionate random sample of individuals living in private households in Germany, aged 18 and older. The data collection for both ALLBUS 2010 and ALLBUS 2012 was conducted by the same fieldwork agency, using similar procedures for contacting and interviewing the respondents. A difference is that in

9 http://www.gesis.org/en/allbus/study-profiles/2010/

10 http://www.gesis.org/en/allbus/study-profiles/2012/

ALLBUS 2010, an incentive experiment was employed in which respondents could receive 10 Euros, 20 Euros, or no incentive (Wasmer, Scholz, Blohm, Walter, & Jutz, 2012, p. 51), whereas in ALLBUS 2012, all respondents were paid 10 Euros. It should be noted the target population of ALLBUS consists of both Internet and non-Internet users. A question on private Internet use is asked in both ALLBUS surveys. For our analysis, we are therefore able to compare GOPP with full samples of the ALLBUS surveys and with the subsamples of Internet users.

ALLBUS contains item batteries and single questions on opinions, which are repeated over the years in order to analyze social trends. The wording of the demographic questions generally does not change between ALLBUS surveys. Overall, the overlap in questions between ALLBUS 2010 and ALLBUS 2012 is 82 questions, 46 of which are not preceded by one or multiple filter questions[11], that is, 46 questions that were asked both in ALLBUS 2010 and in ALLBUS 2012 were posed to the total sample.

# 3 Measures

For the mode system comparison, we made a careful selection of factual and attitudinal questions from ALLBUS 2010 to avoid question format effects. About 73 questions from ALLBUS 2010 were also asked in the GESIS Online Panel Project. However, only a subset of these questions was asked in ALLBUS 2012. Only questions that were present in both ALLBUS 2010 and 2012 and were replicated in the GOPP are analyzed. These questions were not repeatedly measured in GOPP but distributed over the questionnaires, so that each question that we use for comparison was asked only once in the online panel. This prevents possible confounding of the answers due to learning effects (panel conditioning). We carefully inspected the questions and only included questions that had the same question wording (see Appendix for details) and were asked of the whole sample, that is, were not preceded by filter questions.

In total, 12 attitudinal and 7 factual questions fit these criteria. Attitudinal questions included respondents' assessment of the current economic situation in Germany and the economic situation in one year, the assessment of respondents' own financial situation and prospective financial situation in one year, general health, religiosity, self-assessed social class, four general attitude questions on societal functioning (anomie), and political orientation (right-left).

11 http://www.gesis.org/fileadmin/upload/dienstleistung/daten/umfragedaten/allbus/dokumente/VariablenlisteBis2012.pdf. This number does not include the variables with information provided by the interviewer as well as administrative variables.

*Table 1* Sample and survey design description for GOPP and ALLBUS surveys

| Survey characteristic | GOPP | ALLBUS 2010 | ALLBUS 2012 |
|---|---|---|---|
| Sample source | Adjusted RDD (landline+mobile) | Two-stage probability: municipalities and individuals from municipal registers | Two-stage probability: municipalities and individuals from municipal registers |
| Target population | Individuals living in Germany aged 18 and older, who use the Internet | Individuals living in households in Germany aged 18 and older | Individuals living in households in Germany aged 18 and older |
| Recruitment mode | CATI | CAPI | CAPI |
| Data collection mode | Online | CAPI | CAPI |
| Response/Participation rates | CATI AAPOR RR3: 17.1%; Panel participation rates per questionnaire: Q1: 54.92%; Q2: 46.89%; Q3 45.96%; Q4: 44.43%; Q5: 44.08%; Q7: 42.09% | 34.4% | 37.6% |
| Fieldwork | February 2011–May 2012 | May–November 2010 | April–September 2012 |
| Incentives | 0, 2, 5, 10, bonus 20 EUR (randomized experiment) | 0, 10, 20 EUR (experimental) | 10 EUR |
| N | 1114 (started Q1-Q5,Q7) | 2827 | 3480 |
| N Internet users | 1114 | 1869 | 2525 |
| Percent Internet users[a] | 71.5% | 67.1% | 73.6% |

[a] Unweighted proportion for GOPP, design-weighted for ALLBUS 2010 and 2012.
AAPOR RR3 is short for AAPOR Response Rate 3 (American Association for Public Opinion Research, 2011).

Although self-assessed health is not an attitudinal variable in the classical sense, it is a complex question, and the cognitive process required to answer a self-assessment question about health appears to be more similar to the cognitive processes of the attitudinal variables in the analysis, than to the factual questions. The important difference for the cognitive process is that the factual questions in our analysis only require simple processing with no extensive recall. The factual questions concern employment status, marital status, frequency of church attendance, religious confession, being born in Germany, citizenship, and type of dwelling. We recoded several variables to dichotomous variables. The variable "working for pay" generated from "employment status" makes a distinction between those who are in paid work (working full time, part time, or irregularly) and those who are not (not working). "Legal marital status" contrasts legally married persons (married and living together with their spouse, married and living apart) with persons who are not married (divorced, never married or widowed). Religious confession was recoded into an indicator variable, which takes a value of 1 if a confession was named vs. the value of 0 if no confession was named. The percentage of refusals on this variable is negligible. Citizenship of a specific country was recoded as either having German citizenship or not. A new variable, "owner of dwelling," was generated from the variable "type of dwelling" (for details on recoding the variables, see Table A5 in the Appendix).

# 4 Method

We study mode system effects by comparing the estimates from the GESIS Online Panel Pilot (GOPP) and two face-to-face reference surveys. To fully understand the processes when comparing the systems of data collection, we use a stepwise analysis procedure. First, we start with a direct comparison of the data from the GOPP and the two ALLBUS reference surveys. In this analysis, we compare the full samples of the two reference surveys ALLBUS 2010 and ALLBUS 2012, that is, Internet users plus those who do not use the Internet, with each other and with the full sample of the online panel that consists of Internet users only. This allows us to assess the differences between the online panel and the benchmark surveys that arise due to possible coverage bias, nonresponse bias, and mode effects. In this first step, we answer the practical question of what will happen if researchers switch from an interview mode to online surveys and more specifically how this will influence the unadjusted estimates.

In this first step, we do not control for sample composition and the possibility of coverage bias is substantive. Therefore, we expect to find differences between the face-to-face reference surveys and the online panel for both factual and attitudinal variables. Several studies on respondents in online surveys mention differences

in sample composition between those who use the Internet and those who do not (Bandilla, Kaczmirek, Blohm, & Neubarth, 2009; Bosnjak et al. 2013; Mohorko, de Leeuw, & Hox, 2013). Bandilla et al. (2009) found differences in age and education between the groups of Internet users and non-users. Bosnjak et al. (2013) found significant differences related to age, education, and sex between Internet users and non-users when respondents were recruited into an online panel. Mohorko, de Leeuw, and Hox (2013) give an overview of Internet coverage and coverage bias in Europe and point out that even in countries with high Internet coverage the digital divide can be observed, as Internet access is unevenly distributed across the population. In all countries, there are significant differences for age, sex, and education.

Our expectation of finding differences between the mode systems for attitudinal and factual variables is based on the relation of these variables to differences in the covered population, that is, Internet users, and not-covered population (non-Internet-users) in age, sex, and education for the online panel. For instance, due to differences in age, we might find differences in such variables as religiosity, confession, and frequency of church attendance since older people are more likely to be church members and attend religious services (Lois, 2011). In addition, we might find differences in employment status due to its relation with education.

In the second step, we match the reference population, used for the benchmark comparisons, to the target population of the online panel (Internet users). We compare the GESIS Online Panel Pilot to the subsamples of Internet users from the ALLBUS data in order to eliminate coverage as the possible cause of the differences between the GOPP online and the two face-to-face reference surveys.

However, potential differences due to selective nonresponse are not eliminated in this second step, as significant differences have been found between Internet users who are willing to participate in online surveys and Internet users who are not willing to participate. Couper et al. (2007) find that ethnicity, education, and age predict willingness to participate. Bandilla et al. (2009) show that younger and more educated Internet users are more likely to express their willingness to participate in an online survey. For face-to-face interviews, age, sex, and education have been repeatedly found to correlate with nonresponse; for an overview, see Croves & Couper (1998). Hence, differences in demographics between the face-to-face reference surveys and the online panel are expected to persist at this analysis stage. This allows us to assess the differences between the online panel and the benchmark surveys, which arise due to possible nonresponse bias and mode effects.

In the third step, we add weights to compensate for differences in nonresponse. To ensure that differences between the surveys are not caused by sample composition, we use post-stratification weighting based on age, sex, and education.

Post-stratification weighting helps to correct nonresponse errors. Previously we already corrected for coverage bias. In this final step, we are able to assess differences between the online panel and the benchmark surveys due to possible mode

effects as we correct for sample composition. According to hypothesis 1, we do not expect to find differences between the two face-to-face surveys, but we do expect to find differences between the online and the face-to-face surveys. According to hypothesis 2, we expect that differences between the online and the face-to-face surveys are more pronounced for attitudinal questions than for factual questions

To apply post-stratification weighting, classes of the sampled cases are built based on central characteristics (in our case, sex, age, and education), for which the population values are known. The weights are then assigned to the observations in each cell so that the sample data match at least the marginal totals of the population (Gabler & Ganninger, 2010). In standard social surveys, one would use the known population benchmarks to adjust for differences between the sample and the population. However, in our case we use post-stratification to correct for the sample composition bias between the surveys in order to achieve a better assessment of mode effects. Data from the ALLBUS 2010 are used as benchmark data. We treat the distributions of age (five age groups), sex, and education (recoded in three categories, see Table A5 in the Appendix for details) of Internet users in ALLBUS 2010 as reference values (Table A1 in the Appendix). We could have used population values, but unfortunately, neither the German Census nor the German Microcensus includes a question on Internet use. Post-stratification was performed using the iterative proportional fitting (IPF) algorithm also known as raking (Dehming & Stephan, 1940). Raking is an iterative process, the goal of which is to adjust the data furnished by a sample survey to the known marginal distributions obtained from other sources.[12] The weights are obtained stepwise so that the marginal distributions of the weighted data for specified variables match the benchmark marginal distributions.

The weights were calculated in Stata using the ipfweight procedure (Bergmann, 2011). Since the questions were spread over multiple waves in the GOPP, we calculated the post-stratification weights separately for each wave (Tables A2 and A3 in the Appendix). This allowed us to control for attrition and other confounding factors such as experiments with incentives.

Since we use data from single waves of the GOPP without making use of the longitudinal component, no additional panel weights were calculated. The demographic variables in the GOPP, which we use for post-stratification here, were all collected during the recruitment interview. In rare cases, when the previous multiple interview appointments with a respondent failed, or if a respondent was near a break-off, interviewers could ask only about Internet usage and proceed straight to the recruitment question. In such cases, demographic questions were asked later in the online questionnaires. For those cases with missing data on demograph-

12 For examples of IPF weighting compared with other weighting methods, see Gabler (1994) and Little & Wu (1991).

ics in the recruitment interview, we replaced the missing values with information obtained online if it was available. For the rare remaining cases with missing values on demographics, the missing values were then imputed using single hot deck imputation (Schonlau, 2012). In this procedure, the algorithm first identifies all observations that have no missing values for the specified variables (donor observations). In the second step, the algorithm replaces all missing values with values from a randomly chosen donor observation that is similar to the observation that has a missing value. The replacement is performed in such a way that correlations between variables are preserved. In our case, the only variable that had no missing values at all was sex.[13] The imputation was performed before forming the age and educational groups. For an educational group of respondents who were still at school, it could not be known what school-leaving qualification the respondents would obtain; for those who reported having an "other school-leaving degree," it could not be known how their school-leaving degrees related to the degrees of the German educational system. Those still in school and those with other school-leaving certificates were therefore marked as "missing" and imputed with the values of the variable "education" before the educational groups were formed.

In order to obtain the final weights, post-stratification weights were multiplied with design weights. Design weights correct for differences in selection probabilities. For example, in telephone surveys, persons who live in large households have a lower probability of selection than persons living in smaller households. Persons who have a very low chance of selection but have been selected into the sample "weigh" more than persons who have a high chance of being selected. Hence, a person with a low selection probability receives a high design weight, and a person with a high probability of selection receives a low weight (Gabler & Ganninger, 2010). Design weights for the GOPP, where recruitment was performed by telephone interview, were calculated using the Gabler-Häder method (Gabler, Häder, Lehnhoff, & Mardian, 2012). According to this method, the design weights equal the inverse probabilities of selection, which take into account the number of tele-

13 Unfortunately, we could not use other variables except sex since they contained missing values due to the specified recruitment condition in which respondents would only be asked about the Internet use and basic demographics (age and education – the variables for that we would like to impute missing values). We performed the analysis with post-stratification weighting using two data files: with and without imputed missing values on age and education. The results of these analyses differ for some of the variables we compare. However, the estimates that are based on the data if missing values are not imputed are far less precise due to the reduction of the sample size since cases with missing post-stratification weights but for which the information on the substantial analysis variables is present are discarded. Therefore, we report the analysis results with post-stratification weights for the calculation of which the variables age and education were imputed (Table 4). The analysis with post-stratification weights for the calculation of which variables age and education were not imputed are reported in Table A4 in the Appendix.

phone lines at which a respondent can be reached and household composition. If the design weight was missing after supplementing it with the data collected online, it was imputed with the modal category from the subsample of respondents willing to take part in the online panel. Design weights were normalized, that is, rescaled to have a mean of 1 and a sum that equals the unweighted number of cases. The weights were calculated separately for each questionnaire. Design weights in the ALLBUS account for the oversampling of persons from East Germany and were provided in both ALLBUS surveys.[14]

# 5 Results

We start with the results of the first step of the investigation of mode systems: the direct comparison without any adjustment for differences in coverage. Table 2 presents the estimates obtained from each survey and the results of the tests for statistical significance of the differences. The online GOPP differs from the ALLBUS face-to-face reference surveys in both attitudinal and factual questions. We find statistically significant differences between GOPP and ALLBUS 2010 on 11 out of 12 attitudinal items. The only variable for which no differences are found is the respondent's self-rated financial situation. GOPP differs from ALLBUS 2012 on all attitudinal variables, with the exception of 3 items from the anomie-battery, that is, the differences are found on 9 out of 12 attitudinal variables. For factual questions, GOPP differs from the two ALLBUS surveys on all 7 items at a statistically significant level. When we compare the two ALLBUS surveys, we find differences for 10 out of 12 attitudinal variables: current and prospective state of the German economy, current self-rated financial situation, religiosity, the items of the anomie-battery, left-right orientation, and self-assessed social class. For the factual questions, no significant differences between the two ALLBUS surveys were found.

In the next step, we investigate the effect of coverage bias. Table 3 presents the results of the comparisons between the online GOPP with subsamples of Internet users from the two ALLBUS surveys, excluding the non-Internet-users in ALLBUS from the analysis. For several attitudinal variables that showed statistically significant differences between GOPP and the two ALLBUS surveys, no differences are found when GOPP is compared to ALLBUS Internet users only. The difference with the results of the first step indicates potential coverage bias when measuring attitude questions online. For Internet users only, GOPP and ALLBUS 2010 only differ on 6 instead of 11 out of 12 attitudinal items: current and prospective state of German economy, health status, religiosity, one item of the anomie-battery,

14 For ALLBUS 2010, 69.33% of all interviews took place in West Germany and 30.67% in East Germany, for ALLBUS 2012, these proportions were 67.76% and 32.24% for the West and the East respectively.

and left-right orientation. The differences from the subsample of Internet users in ALLBUS 2012 are found for 4 instead of 9 attitudinal variables: prospective German economy, current self-rated financial situation, religiosity, and one category of the variable self-assessed social class. We find statistically significant differences between GOPP and the subsample of Internet users in ALLBUS 2010 for all factual variables with the exception of the variable "German citizenship." The differences between GOPP and ALLBUS 2012 Internet users are statistically significant for all 7 factual variables. With the exception of prospective financial situation and self-assessed social class, the ALLBUS surveys differ from one another on all attitudinal variables. In sum, when comparing the subsamples of Internet users, ALLBUS surveys differ from each other in 10 out of 12 attitudinal variables and none of the factual variables.

In the last step, we use weighting adjustment to compensate for potential differences in nonresponse between the surveys. Table 4 reports all comparisons using weighted data. The set of variables in which the online GOPP differs from the ALLBUS surveys is different from the set of variables that showed significant differences reported in Table 3, where sample composition bias due to coverage but not due to nonresponse was taken into account. The online GOPP now differs from both ALLBUS surveys in 6 out of 12 attitudinal items. In the following, we refer to the comparisons of GOPP with the subsamples of Internet users in ALLBUS with post-stratification weighting applied to GOPP and to ALLBUS 2012.

The variables in which GOPP differs from ALLBUS 2010 are not the same variables in which GOPP differs from ALLBUS 2012. Significant differences between ALLBUS 2010 and GOPP are now observed for the following attitudinal variables: state of German economy (both current and prospective), self-rated current financial situation, health status, religiosity, and left-right orientation. The set of attitudinal variables with statistically significant differences between GOPP and ALLBUS 2012 excludes the current state of German economy and left-right orientation, but includes two of the items of the anomie-battery. However, the two ALLBUS surveys still show significant mutual differences for the same 10 out of 12 items reported in Table 3.

For factual questions, GOPP differs from ALLBUS 2010 in 6 out of 7 items: in all factual questions with the exception of confession. Estimates from GOPP differ from ALLBUS 2012 in 5 out of 7 items, the exceptions being confession and frequency of going to church. As was the case with the unweighted data in Table 3, there are no significant differences between the two ALLBUS surveys for any of the factual questions. These two findings taken together indicate the presence of a mode system effect in our data. The two face-to-face reference surveys do not show any mutual differences regarding factual questions, but they do differ from the online results, even after adjusting for sample composition differences.

*Table 2* Comparison of the GESIS Online Panel Pilot to two ALLBUS surveys: attitudinal and factual questions

| | Q | GOPP | | | ALLBUS 2010 full sample | | | ALLBUS 2012 full sample | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | | Estimate | SE | 95% CI | Estimate | SE | 95% CI | Estimate | SE | 95% CI |
| *Attitudinal:* | | | | | | | | | | |
| German economy | 1 | 3.400 | 0.026 | [3.350, 3.450] | 3.037*** | 0.016 | [3.006, 3.068] | ***3.332**** | 0.015 | [3.303, 3.360] |
| German economy in 1 year | 1 | 2.868 | 0.029 | [2.811, 2.925] | 3.014*** | 0.017 | [2.980, 3.047] | ***2.662***** | 0.013 | [2.637, 2.688] |
| Self-rated financial situation | 1 | 3.454 | 0.027 | [3.401, 3.507] | 3.411 | 0.017 | [3.379, 3.444] | ***3.559**** | 0.014 | [3.531, 3.587] |
| Self-rated financial situation in 1 year | 1 | 3.187 | 0.027 | [3.134, 3.241] | 3.090** | 0.014 | [3.064, 3.117] | 3.094** | 0.011 | [3.071, 3.116] |
| Health status | 1 | 3.679 | 0.032 | [3.616, 3.743] | 3.561** | 0.021 | [3.521, 3.602] | 3.562** | 0.018 | [3.527, 3.597] |
| Religiosity | 4 | 4.462 | 0.125 | [4.217, 4.707] | 5.724*** | 0.059 | [5.608, 5.840] | ***6.111***** | 0.054 | [6.006, 6.216] |
| Self-assessed social class | 5 | | | | | | | | | |
| lower class | | 2.68% | 0.007 | [1.66, 4.32] | 3.16% | 0.003 | [2.56, 3.90] | ***1.94%*** | 0.002 | [1.53, 2.46] |
| working class | | 19.43% | 0.017 | [16.34, 22.96] | 25.75%** | 0.008 | [24.13, 27.45] | 26.21%*** | 0.008 | [24.73, 27.75] |
| middle class | | 61.12% | 0.020 | [57.18, 64.92] | 59.74% | 0.010 | [57.83, 61.61] | 61.12% | 0.009 | [59.41, 62.80] |
| upper middle class | | 16.14% | 0.014 | [13.56, 19.09] | 10.84%** | 0.006 | [9.67, 12.13] | 10.08%*** | 0.006 | [9.05, 11.22] |
| upper class | | 0.63% | 0.003 | [0.25, 1.56] | 0.51% | 0.001 | [0.29, 0.89] | 0.65% | 0.001 | [0.42, 1.00] |
| Life is getting worse | 7 | 78.66% | 0.017 | [75.20, 81.75] | 83.42%** | 0.007 | [81.93, 84.81] | ***77.44%*** | 0.008 | [75.91, 78.90] |
| Irresponsible to have children | 7 | 35.10% | 0.020 | [31.36, 39.04] | 42.07%** | 0.010 | [40.19, 43.98] | ***35.70%*** | 0.009 | [34.05, 37.39] |
| Politicians not interested | 7 | 76.86% | 0.018 | [73.23, 80.14] | 80.82%* | 0.008 | [79.24, 82.31] | ***76.18%*** | 0.008 | [74.62, 77.67] |
| People don't care about others | 7 | 65.30% | 0.020 | [61.33, 69.07] | 73.99%*** | 0.009 | [72.26, 75.64] | ***70.91%**** | 0.008 | [69.27, 72.49] |
| Left-right orientation | 7 | 4.848 | 0.070 | [4.712, 4.985] | 5.218*** | 0.034 | [5.152, 5.284] | ***5.045**** | 0.031 | [4.984, 5.106] |

*Table 2 continued*

| | Q | GOPP | | | ALLBUS 2010 full sample | | | ALLBUS 2012 full sample | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | | Estimate | SE | 95% CI | Estimate | SE | 95% CI | Estimate | SE | 95% CI |
| *Factual:* | | | | | | | | | | |
| Working for pay | 2 | 83.70% | 0.014 | [80.73, 86.29] | 57.80%*** | 0.010 | [55.91, 59.68] | 59.76%*** | 0.009 | [58.04, 61.45] |
| Legal marital status | 3 | 50.47% | 0.020 | [46.60, 54.34] | 57.55%** | 0.010 | [55.65, 59.42] | 57.49%** | 0.009 | [55.77, 59.20] |
| Confession | 4 | 64.56% | 0.019 | [60.74, 68.20] | 73.14%*** | 0.008 | [71.49, 74.73] | 73.13%*** | 0.007 | [71.65, 74.57] |
| Frequency of church attendance | 4 | 2.143 | 0.045 | [2.055, 2.231] | 2.385*** | 0.025 | [2.335, 2.434] | 2.338*** | 0.023 | [2.293, 2.384] |
| Born in Germany | 4 | 90.72% | 0.012 | [88.17, 92.76] | 84.01%*** | 0.007 | [82.53, 85.39] | 85.27%*** | 0.006 | [83.96, 86.49] |
| German citizenship | 4 | 96.25% | 0.007 | [94.51, 97.45] | 94.07%* | 0.005 | [93.07, 94.94] | 93.93%** | 0.004 | [93.00, 94.73] |
| Owner of dwelling | 5 | 47.78% | 0.020 | [43.86, 51.73] | 57.11%*** | 0.010 | [55.20, 58.99] | 59.35%*** | 0.009 | [57.64, 61.04] |

*Note*: *$p<.05$, **$p<.01$, ***$p<.001$. Differences between ALLBUS 2010 and 2012 in italic bold, Q short for GOPP questionnaire, SE short for standard error, CI short for confidence interval. N for GOPP (started): Q1=1010; Q2=838; Q3=800; Q4=775; Q5=761; Q7=729. Design weights are applied to all three surveys. Significance tests are pairwise t-tests for continuous/ordinal variables or tests of proportions.

*Table 3* Comparison of the GESIS Online Panel Pilot to the subsamples of Internet users from two ALLBUS surveys: attitudinal and factual questions

| | Q | GOPP | | | ALLBUS 2010 Internet users | | | ALLBUS 2012 Internet users | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | | Estimate | SE | 95% CI | Estimate | SE | 95% CI | Estimate | SE | 95% CI |
| *Attitudinal:* | | | | | | | | | | |
| German economy | 1 | 3.400 | 0.026 | [3.350, 3.450] | 3.086*** | 0.019 | [3.048, 3.124] | ***3.355*** | 0.017 | [3.321, 3.389] |
| German economy in 1 year | 1 | 2.868 | 0.029 | [2.811, 2.925] | 3.077*** | 0.021 | [3.036, 3.117] | ***2.663**** | 0.015 | [2.633, 2.694] |
| Self-rated financial situation | 1 | 3.454 | 0.027 | [3.401, 3.507] | 3.471 | 0.020 | [3.433, 3.510] | ***3.595***** | 0.016 | [3.564, 3.627] |
| Self-rated financial situation in 1 year | 1 | 3.187 | 0.027 | [3.134, 3.241] | 3.182 | 0.017 | [3.148, 3.216] | 3.145 | 0.014 | [3.117, 3.173] |
| Health status | 1 | 3.679 | 0.032 | [3.616, 3.743] | 3.823*** | 0.023 | [3.778, 3.868] | ***3.725*** | 0.020 | [3.686, 3.764] |
| Religiosity | 4 | 4.462 | 0.125 | [4.217, 4.707] | 5.974*** | 0.073 | [5.832, 6.117] | ***6.334***** | 0.062 | [6.213, 6.456] |
| Self-assessed social class | 5 | | | | | | | | | |
| lower class | | 2.68% | 0.007 | [1.65, 4.32] | 2.27% | 0.003 | [1.67, 3.06] | 1.58% | 0.002 | [1.16, 2.14] |
| working class | | 19.43% | 0.017 | [16.34, 22.96] | 20.23% | 0.010 | [18.42, 22.17] | 21.81% | 0.008 | [20.20, 23.52] |
| middle class | | 61.12% | 0.020 | [57.18, 64.92] | 63.38% | 0.012 | [61.08, 65.63] | 63.71% | 0.010 | [61.73, 65.64] |
| upper middle class | | 16.14% | 0.014 | [13.56, 19.09] | 13.80% | 0.008 | [12.22, 15.55] | 12.24%* | 0.007 | [10.93, 13.67] |
| upper class | | 0.63% | 0.003 | [0.25, 1.56] | 0.32% | 0.001 | [0.14, 0.73] | 0.66% | 0.002 | [0.40, 1.10] |
| Life is getting worse | 7 | 78.66% | 0.017 | [75.20, 81.75] | 82.00% | 0.009 | [80.10, 83.74] | ***75.71%*** | 0.009 | [73.88, 77.45] |
| Irresponsible to have children | 7 | 35.10% | 0.020 | [31.36, 39.04] | 37.35% | 0.012 | [35.10, 39.65] | ***31.07%*** | 0.010 | [29.22, 32.99] |
| Politicians not interested | 7 | 76.86% | 0.018 | [73.23, 80.14] | 77.42% | 0.010 | [75.37, 79.35] | ***73.63%*** | 0.009 | [71.75, 75.43] |
| People don't care about others | 7 | 65.30% | 0.020 | [61.33, 69.07] | 71.72%** | 0.011 | [69.55, 73.79] | ***68.69%*** | 0.010 | [66.74, 70.57] |
| Left-right orientation | 7 | 4.848 | 0.070 | [4.712, 4.985] | 5.143*** | 0.041 | [5.063, 5.222] | ***4.937*** | 0.035 | [4.868, 5.006] |

*Table 3 continued*

| | Q | GOPP | | | ALLBUS 2010 Internet users | | | ALLBUS 2012 Internet users | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | | Estimate | SE | 95% CI | Estimate | SE | 95% CI | Estimate | SE | 95% CI |
| *Factual:* | | | | | | | | | | |
| Working for pay | 2 | 83.70% | 0.014 | [80.73, 86.29] | 72.41%*** | 0.011 | [70.26, 74.46] | 72.11%*** | 0.009 | [70.25, 73.90] |
| Legal marital status | 3 | 50.47% | 0.020 | [46.60, 54.34] | 56.89%** | 0.012 | [54.56, 59.19] | 57.00%** | 0.010 | [54.98, 59.00] |
| Confession | 4 | 64.56% | 0.019 | [60.74, 68.20] | 70.57%** | 0.011 | [68.47, 72.59] | 71.00%** | 0.009 | [69.20, 72.74] |
| Frequency of church attendance | 4 | 2.143 | 0.045 | [2.055, 2.231] | 2.301** | 0.029 | [2.244, 2.359] | 2.287** | 0.026 | [2.237, 2.338] |
| Born in Germany | 4 | 90.72% | 0.012 | [88.17, 92.76] | 87.44%* | 0.008 | [85.76, 88.94] | 86.62%** | 0.007 | [85.13, 87.99] |
| German citizenship | 4 | 96.25% | 0.007 | [94.51, 97.45] | 94.53% | 0.006 | [93.32, 95.52] | 93.64%** | 0.005 | [92.52, 94.59] |
| Owner of dwelling | 5 | 47.78% | 0.020 | [43.86, 51.73] | 56.64%*** | 0.012 | [54.30, 58.95] | 59.19%*** | 0.010 | [57.18, 61.17] |

*Note*: *$p<.05$, **$p<.01$, ***$p<.001$. Differences between ALLBUS 2010 and 2012 in italic bold, Q short for GOPP questionnaire, SE short for standard error, CI short for confidence interval. N for GOPP (started): Q1=1010; Q2=838; Q3=800; Q4=775; Q5=761; Q7=729. Design weights are applied to all three surveys. Significance tests are pairwise t-tests for continuous/ordinal variables or tests of proportions.

*Table 4* Comparison of the GESIS Online Panel Pilot to the subsamples of Internet users from two ALLBUS surveys: attitudinal and factual questions with post-stratification

| | Q | GOPP | | | ALLBUS 2010 Internet users | | | ALLBUS 2012 Internet users | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | | Estimate | SE | 95% CI | Estimate | SE | 95% CI | Estimate | SE | 95% CI |
| *Attitudinal:* | | | | | | | | | | |
| German economy | 1 | 3.336 | 0.029 | [3.279, 3.393] | 3.086*** | 0.019 | [3.048, 3.124] | ***3.351*** | 0.018 | [3.317, 3.385] |
| German economy in 1 year | 1 | 2.844 | 0.033 | [2.779, 2.909] | 3.077*** | 0.021 | [3.036, 3.117] | ***2.667**** | 0.016 | [2.636, 2.698] |
| Self-rated financial situation | 1 | 3.394 | 0.030 | [3.334, 3.453] | 3.471* | 0.020 | [3.433, 3.510] | ***3.593**** | 0.017 | [3.560, 3.626] |
| Self-rated financial situation in 1 year | 1 | 3.168 | 0.031 | [3.107, 3.228] | 3.182 | 0.017 | [3.148, 3.216] | 3.157 | 0.015 | [3.128, 3.186] |
| Health status | 1 | 3.606 | 0.038 | [3.531, 3.681] | 3.823*** | 0.023 | [3.778, 3.868] | ***3.748**** | 0.020 | [3.709, 3.787] |
| Religiosity | 4 | 4.585 | 0.145 | [4.300, 4.870] | 5.974*** | 0.073 | [5.832, 6.117] | ***6.341**** | 0.063 | [6.218, 6.466] |
| Self-assessed social class | 5 | | | | | | | | | |
| lower class | | 3.36% | 0.010 | [1.85, 6.03] | 2.27% | 0.003 | [1.67, 3.06] | 1.55% | 0.003 | [1.13, 2.13] |
| working class | | 24.81% | 0.023 | [20.52, 29.65] | 20.23% | 0.010 | [18.42, 22.17] | 22.21% | 0.009 | [20.55, 23.97] |
| middle class | | 59.36% | 0.024 | [54.60, 63.94] | 63.38% | 0.012 | [61.08, 65.63] | 63.24% | 0.010 | [61.21, 65.22] |
| upper middle class | | 12.02% | 0.012 | [9.87, 14.56] | 13.80% | 0.008 | [12.22, 15.55] | 12.36% | 0.007 | [11.03, 13.83] |
| upper class | | 0.45% | 0.002 | [0.18, 1.16] | 0.32% | 0.001 | [0.14, 0.73] | 0.64% | 0.002 | [0.38, 1.09] |
| Life is getting worse | 7 | 79.51% | 0.019 | [75.47, 83.04] | 82.00% | 0.009 | [80.10, 83.74] | ***75.58%*** | 0.009 | [73.70, 77.36] |
| Irresponsible to have children | 7 | 40.41% | 0.024 | [35.87, 45.13] | 37.35% | 0.012 | [35.10, 39.65] | ***31.06%***** | 0.010 | [29.17, 33.01] |
| Politicians not interested | 7 | 79.28% | 0.019 | [75.38, 82.70] | 77.42% | 0.010 | [75.37, 79.35] | ***73.56%**** | 0.010 | [71.64, 75.40] |
| People don't care about others | 7 | 68.45% | 0.022 | [64.09, 72.50] | 71.72% | 0.011 | [69.55, 73.79] | ***68.63%*** | 0.010 | [66.64, 70.55] |
| Left-right orientation | 7 | 4.945 | 0.080 | [4.788, 5.102] | 5.143* | 0.041 | [5.063, 5.222] | ***4.946*** | 0.036 | [4.876, 5.016] |

*Table 4 continued*

| | Q | GOPP | | | ALLBUS 2010 Internet users | | | ALLBUS 2012 Internet users | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | | Estimate | SE | 95% CI | Estimate | SE | 95% CI | Estimate | SE | 95% CI |
| *Factual:* | | | | | | | | | | |
| Working for pay | 2 | 83.55% | 0.016 | [80.28, 86.37] | 72.41%*** | 0.011 | [70.26, 74.46] | 73.60%*** | 0.009 | [71.75, 75.37] |
| Legal marital status | 3 | 49.87% | 0.023 | [45.30, 54.44] | 56.89%** | 0.012 | [54.56, 59.19] | 56.23%* | 0.011 | [54.16, 58.28] |
| Confession | 4 | 67.29% | 0.021 | [63.10, 71.21] | 70.57% | 0.011 | [68.47, 72.59] | 71.29% | 0.009 | [69.46, 73.04] |
| Frequency of church attendance | 4 | 2.175 | 0.052 | [2.073, 2.277] | 2.301* | 0.029 | [2.244, 2.359] | 2.286 | 0.026 | [2.234, 2.337] |
| Born in Germany | 4 | 91.04% | 0.013 | [88.08, 93.32] | 87.44%* | 0.008 | [85.76, 88.94] | 86.25%** | 0.008 | [84.69, 87.67] |
| German citizenship | 4 | 96.50% | 0.008 | [94.63, 97.73] | 94.53%* | 0.006 | [93.32, 95.52] | 93.50%** | 0.006 | [92.33, 94.50] |
| Owner of dwelling | 5 | 45.38% | 0.023 | [40.86, 49.98] | 56.64%*** | 0.012 | [54.30, 58.95] | 58.06%*** | 0.010 | [56.99, 60.10] |

*Note*: *$p<.05$, **$p<.01$, ***$p<.001$. Differences between ALLBUS 2010 and 2012 in italic bold, Q short for GOPP questionnaire, SE short for standard error, CI short for confidence interval. N for GOPP (started): Q1=1010; Q2=838; Q3=800; Q4=775; Q5=761; Q7=729. Weighted data: ALLBUS 2010 design, GOPP & ALLBUS 2012 design and post-stratification. Significance tests are pairwise t-tests for continuous/ordinal variables or tests of proportions.

The fact that the two face-to-face ALLBUS surveys differ on attitudinal questions shows that attitudes, being more unstable constructs, would not have allowed us to single out a mode system effect. We see that two surveys with identical recruitment and design features do not differ in factual questions, whereas the difference between the online and the face-to-face surveys is clear.

However, finding statistically significant differences is not the only indicator of mode system effects. In addition to statistical significance, Biemer (1988) recommends examining the effect size, the direction of the difference, and the violations of the underlying assumptions for the mode comparison study, which could explain the magnitude of the difference.

Calculating effect sizes standardizes the comparisons between the means and proportions reported in Table 4 and allows for a better estimation of error, because effect sizes do not only take the difference of the estimates into account, but also the sample sizes and the precision of estimates.

We calculated the standardized mean difference effect sizes that are used to synthesize results from studies that contrast two groups on measures with a continuous underlying distribution for the continous and ordinal variables (Lipsey & Wilson, 2001, p. 172) and approximated standardized mean difference effect sizes that are calculated differencing the arcsine-transformed proportions (Lipsey & Wilson, 2001, p. 187) for binary variables.[15] The standardized mean difference effect sizes allow us to compare the size and direction of the difference between the GOPP and ALLBUS 2010 with the difference between GOPP and ALLBUS 2012 as well as with the difference between ALLBUS 2010 and ALLBUS 2012 for each variable. In order to compare the differences between the surveys for groups of variables (attitudinal vs. factual), we calculated the mean effect sizes. The mean effect size is computed by weighting each effect size by the inverse of its variance (Lipsey & Wilson, 2001, p. 114). Since the standardized mean effect sizes take the direction of the difference (indicated by the sign of the effect size) into account, the differences between the variables may be underestimated when the sum of the negative

15 The standardized mean difference effect size ($ES_{sm}$) is the difference between the group means ($\bar{X}$) divided by the pooled standard deviation ($s$), which is calculated based on the sample sizes for each group:

$$ES_{sm} = \frac{\bar{X}_1 - \bar{X}_2}{s_{pooled}}, \; s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

The approximations based on dichotomous data: $ES_{sm} = arcsine(p_1) - arcsine(p_2)$, where $p$ is the proportion for each group (Lipsey & Wilson, 2001, p. 198-200).

The mean effect size $\overline{ES} = \frac{\sum_{i=1}^{k}(w_i \times ES_i)}{\sum_{i=1}^{k} w_i}$; $w_i = \frac{1}{\overline{SE_i^2}}$,

where $w$ is the inverse variance weight, $ES$ is the unweighted effect size *(d)*, and $SE$ is the standard error of the difference between the effect sizes (Lipsey & Wilson, 2001, p. 114).

and positive effect sizes is calculated in order to compute the mean effect size. We therefore also calculate the absolute mean effect sizes, which are based on absolute values of the standardized mean difference effect sizes. In Table 5, we report the standardized mean difference effect sizes (*d*) and mean effect sizes as well as absolute mean effect sizes for attitudinal variables, factual variables and the overall mean effect size, which incorporates both attitudinal and factual variables.

From Table 5, we can conclude that the directions of all mean effect sizes are the same. For weighted mean effect sizes over all questions, the difference between GOPP and ALLBUS 2010 (-0.063) and the difference between GOPP and ALLBUS 2012 (-0.036) are larger than the difference between the two ALLBUS surveys (-0.030). However, the difference between GOPP and ALLBUS 2012 has almost the same magnitude as the difference between the two face-to-face reference surveys.

If we look at the effect sizes by type of questions (factual vs. attitudinal), the results mimic what we have seen when comparing the frequency counts of statistically significant results. For factual questions, the difference between the two face-to-face surveys (ALLBUS 2010 & 2012) is minimal (0.001) and for the online GOPP vs. face-to-face surveys the differences are almost the same (0.048 and 0.050). For attitudinal questions ALLBUS 2012 is closer to GOPP (-0.031) than to ALLBUS 2010 (-0.068). The difference between the two face-to-face surveys on attitudinal questions is even larger than the difference between the online GOPP and the ALLBUS 2012 (-0.042 vs. -0.031). Although effect sizes differ between surveys and between question types, most are very small (Ferguson, 2009, p. 533). The absolute weighted mean effect sizes reported in the end of Table 5 do not account for the direction of the effect, but only for the magnitude. Here, we see, once again, that for attitudinal items GOPP estimates are closer to ALLBUS 2012 estimates, but in the same magnitude as the difference between the face-to-face surveys. However, for factual questions the difference between the two face-to-face surveys is much smaller than between each of the two ALLBUS surveys and GOPP, resulting in an overall difference between GOPP and each individual ALLBUS survey that is larger than that between the two ALLBUS surveys. Although absolute mean effect sizes are larger than mean effect sizes, they are still small and do not exceed 0.2. From a practical point of view, the mode system effects between the surveys are therefore negligible when judged by their effect sizes and small when judged by the absolute effect sizes. Naturally, this is after differences in sample composition have been accounted for.

*Table 5* Effect sizes (*d*), inverse variance weights (*w*) and mean effect sizes across the surveys

| Variable | GOPP vs. ALLBUS 2010 Internet users | | GOPP vs. ALLBUS 2012 Internet users | | ALLBUS 2012 vs. 2010 Internet users | |
|---|---|---|---|---|---|---|
| | *d* | *w* | *d* | *w* | *d* | *w* |
| *Attitudinal*: | | | | | | |
| German economy | 0.295 | 625.000 | -0.017 | 688.705 | 0.307 | 1051.525 |
| German economy in 1 year | -0.247 | 623.053 | 0.206 | 682.128 | -0.488 | 1019.368 |
| Self-rated financial situation | -0.087 | 630.120 | -0.229 | 684.932 | 0.143 | 1062.699 |
| Self-rated financial situation in 1 year | -0.017 | 626.566 | 0.014 | 684.463 | -0.034 | 1052.632 |
| Health status | -0.206 | 621.504 | -0.136 | 679.348 | -0.075 | 1064.963 |
| Religiosity | -0.407 | 530.786 | -0.523 | 568.182 | 0.117 | 1059.322 |
| Self-assessed social class | | | | | | |
| lower class | 0.222 | 49.998 | 0.437 | 48.603 | -0.214 | 64.045 |
| working class | 0.145 | 311.915 | 0.080 | 346.380 | 0.065 | 571.755 |
| middle class | -0.094 | 416.840 | -0.090 | 452.284 | -0.003 | 800.000 |
| upper middle class | -0.087 | 191.022 | -0.018 | 201.086 | -0.070 | 393.236 |
| upper class | 0.189 | 7.004 | -0.195 | 9.100 | 0.384 | 13.942 |
| Life is getting worse | -0.088 | 269.469 | 0.125 | 305.344 | -0.213 | 552.181 |
| Irresponsible to have children | 0.071 | 406.669 | 0.225 | 428.266 | -0.154 | 776.398 |
| Politicians not interested | 0.060 | 284.738 | 0.175 | 310.945 | -0.115 | 626.174 |
| People don't care about others | -0.086 | 362.450 | -0.005 | 394.477 | -0.082 | 719.424 |
| Left-right orientation | -0.107 | 509.684 | -0.001 | 549.451 | -0.113 | 1022.495 |
| *Factual*: | | | | | | |
| Working for pay | 0.364 | 280.741 | 0.331 | 296.121 | 0.033 | 692.042 |
| Legal marital status | -0.156 | 453.309 | -0.141 | 490.196 | -0.015 | 861.326 |
| Confession | -0.085 | 383.877 | -0.104 | 413.736 | 0.019 | 720.461 |
| Frequency of church attendance | -0.097 | 539.374 | -0.084 | 582.072 | -0.012 | 1062.699 |
| Born in Germany | 0.209 | 157.109 | 0.266 | 169.033 | -0.058 | 397.931 |
| German citizenship | 0.258 | 66.934 | 0.359 | 72.417 | -0.101 | 193.836 |
| Owner of dwelling | -0.250 | 435.730 | -0.282 | 468.604 | 0.032 | 856.164 |

*Table 5 continued*

| Variable | GOPP vs. ALLBUS 2010 Internet users | | GOPP vs. ALLBUS 2012 Internet users | | ALLBUS 2012 vs. 2010 Internet users | |
|---|---|---|---|---|---|---|
| | *d* | *w* | *d* | *w* | *d* | *w* |
| Mean effect size: | | | | | | |
| attitudinal questions | | -0.068 | | -0.031 | | -0.042 |
| | | (0.012) | | (0.012) | | (0.009) |
| factual questions | | -0.048 | | -0.050 | | -0.001 |
| | | (0.021) | | (0.020) | | (0.014) |
| Overall | | -0.063 | | -0.036 | | -0.030 |
| | | (0.011) | | (0.010) | | (0.008) |
| Mean absolute effect size: | | | | | | |
| attitudinal questions | | 0.157 | | 0.141 | | 0.150 |
| | | (0.012) | | (0.012) | | (0.009) |
| factual questions | | 0.180 | | 0.185 | | 0.028 |
| | | (0.021) | | (0.020) | | (0.014) |
| Overall | | 0.163 | | 0.153 | | 0.115 |
| | | (0.011) | | (0.010) | | (0.008) |

*Note*: *d* is the unweighted effect size, *w* is the inverse variance weight, standard errors in parentheses.

# 6 Discussion

We investigated mode system effects by comparing the data from a probability-based telephone-recruited online panel with the data from two face-to-face surveys. All three sources implemented questions with identical wording and we controlled for differences in sample composition due to undercoverage and selective nonresponse in order to single out the mode system effect. We distinguished between factual and attitudinal questions. We hypothesized that the effects are more pronounced for attitudinal questions than for factual questions. This hypothesis finds no support. There are differences between the online collected data and interviewer collected data on both types of questions. However, the distinction between factual and attitudinal data remains important. We did find that for factual questions both face-to-face surveys differ from the online panel, but do not differ significantly from each other. This conclusion is supported by the analysis of effect sizes. For attitudinal questions, the difference between the two interviewer-administered face-to-face surveys is larger than the difference between one face-to-face survey (ALLBUS 2012) and the online panel. We attribute this result to the instability of attitudes. However, alternative explanations are possible. For example, the reference period

of the ALLBUS 2012 is closer to the online panel, which was administered in 2011-2012 than to the ALLBUS 2010. Surprisingly, the mean effect size for differences in attitudinal variables is larger for ALLBUS 2010 to ALLBUS 2012 comparison than for the GOPP to ALLBUS 2012 comparison, although all effect sizes are small. When we look at the content of the questions, we see that effect sizes are relatively large for questions that refer to the economic situation. The differences between the surveys on those variables might be attributed not to the mode system effect, but to the fact that Germany experienced real economic changes between 2010 and 2012. To examine the robustness of our results, we recalculated what the effect sizes are when the four variables that refer to the economic condition are excluded. This did not affect our overall conclusions, but it provided us with more accurate measures of the mode system effect.[16]

When comparing data from surveys to data from benchmark studies, Callegaro et al. (2014) advise that the following conditions be met: (1) question wording should be identical across compared surveys and (2) populations represented by each survey need to be comparable. Typically, studies comparing online panel data to benchmarks use demographic and behavioral measures (Callegaro et al., 2014). One of the reasons for this might be that the official statistics ideally used as benchmarks can only provide such data. If benchmarks come from high-quality surveys, using attitudinal measures can be considered. Our study meets both conditions (identical questions wording and comparable populations after adjustment) and shows the importance of the reference period depending on the nature of the measure. Ideally, the reference period should be the same for a benchmark source (survey) and the survey that is being compared to the benchmark – a situation given when a pure mode effect is estimated in a randomized experiment. However, in practical circumstances – when estimating a mode system effect – this might not be the case.

Our findings have important implications for estimating mode system effects. First, we have shown the importance of having more than one reference survey. In our case, both face-to-face surveys serve as control surveys for the online panel and allow us to draw conclusions about whether or not differences could be attributed to the mode. We use two surveys with equal recruitment procedures. Second, we show the importance of distinguishing between types of questions. In past studies, differences between online and interviewer-administered modes are well-documented for sensitive questions (e.g., meta-analysis by Tourangeau, Conrad, & Couper, 2013,

16 Mean effect sizes without the economic items. Overall: -0.082 (0.013), -0.048 (0.012), -0.036 (0.009); absolute: 0.164 (0.013), 0.168 (0.012), 0.073 (0.009); for attitudinal items: -0.102 (0.016), -0.047 (0.015), -0.058 (0.011); absolute for attitudinal items: 0.155 (0.016), 0.157 (0.015), 0.101 (0.011).

p. 142).[17] We draw on another dimension, in which a distinction is made based on the cognitive demands that questions place on a respondent. Our findings are in line with Klausch, Hox, and Schouten (2014), who also report that mode effects depend on the question type. For data analysts, the conclusion to be drawn from our analysis is that mode system effects differ across question types. Data users should therefore have different concerns about mode effects when analyzing only attitudinal, only factual, or both types of items. Third, we find a common denominator for the comparison of mode system effects by reporting effect sizes. The magnitude of the mode system effect when judged by the effect sizes is small. However, researchers who use data from large surveys might be misled if they rely solely on significance testing. We encourage other researchers to report effect sizes. This would allow for comparing our results to similar studies based on different data. Fourth, it is important to realize, that we corrected for sample composition bias, and found indications of coverage and nonresponse bias. In survey practice, one could use mixed-mode approaches to account for the digital divide; an example is the GESIS Panel[18], where a mix of postal mail and Internet is used for data collection.

Finally, we investigated the effect of mode systems on point estimates answering the practical question of survey practitioners, of what happens when we change our data collection procedures. To tease out the reasons for mode system effects, a series of carefully designed experiments is needed. These experiments should take both selection and measurement effects into account and not only use point estimates, but also rely on other indicators, such as indices for response tendencies (cf. Tourangeau, 2013).

17 We did not have sensitive items to use for our analysis. However, for religiosity, we find that the percentage online panel respondents who choose the answer "not religious" is much higher than for interviewer-administered surveys (29.44% with SE=0.021, CI=[25.56, 33.64] for GOPP and 7.72% with SE=0.007, CI=[6.54, 9.10] for ALLBUS 2010, 6.27% with SE=0.005, CI=[5.32, 7.38], $p<.001$ for GOPP vs. either ALLBUS). This could be indicative of social desirability.

18 www.gesis-panel.org

# References

American Association for Public Opinion Research. (2011). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 7th edition: AAPOR.

Auspurg, K., Burton, J., Cullinane, C., Delavande, A., Fumagelli, L., Iacovou, M., Zafar, B. (2013). Understanding Society Innovation Panel Wave 5: Results from Methodological Experiments. In J. Burton (Ed.), *Understanding Society Working Paper Series*. Essex: Institute for Social and Economic Research, University of Essex.

Bandilla, W., Kaczmirek, L., Blohm, M., & Neubarth, W. (2009). Coverage und Nonresponse-Effekte bei Online-Bevölkerungsumfragen [Coverage and nonresponse effects in online population-surveys]. In N. Jackob, H. Schoen & T. Zerback (Eds.), *Sozialforschung im Internet: Methodologie und Praxis der Online-Befragung* (pp. 129-143). Wiesbaden: VS Verlag.

Bergmann, M. (2011). Stata-ado "ipfweight": IPF-Algorithm to create adjustment survey weights. from http://fmwww.bc.edu/repec/bocode/i/ipfweight.ado

Biemer, P. P. (1988). Measuring Data Quality. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls & J. Waksberg (Eds.), *Telephone Survey Methodology* (pp. 273-283). New York: Wiley.

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. Hoboken, New Jersey: Wiley.

Bosnjak, M., Haas, I., Galesic, M., Kaczmirek, L., Bandilla, W., & Couper, M. P. (2013). Sample composition discrepancies in different stages of a probability-based online panel. *Field Methods*. doi: 10.1177/1525822X12472951

Callegaro, M., Villar, A., Yeager, D., & Krosnick, J. A. (2014). A critical review of studies investigating the quality of data obtained with online panels based on probability and nonprobability samples. In M. Callegaro, R. Baker, J. D. Bethlehem, A. S. Göritz, J. A. Krosnick & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective* (pp. 23-53). New York: Wiley.

Couper, M. P. (2011). The future of modes of data collection. *Public Opinion Quarterly, 75*, 5, 889-908.

Couper, M. P., Kapteyn, A., Schonlau, M., & Winter, J. (2007). Noncoverage and Nonresponse in an Internet Survey. *Social Science Research, 36*, 131-148.

Dehming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics, 11*(4), 427-444.

de Leeuw, E. D. (1992). *Data quality in mail, telephone, and face-to-face surveys.* Amsterdam: TT.Publikaties. http://edithl.home.xs4all.nl/pubs/disseddl.pdf

de Leeuw, E. D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics, 21*(2), 233-255.

de Leeuw, E. D., & Hox, J. J. (2011). Internet surveys as a part of a mixed-mode design. In M. Das, P. Ester & L. Kaczmirek (Eds.), *Social and behavioral research and the Internet* (Vol. 45-76). New York: Routledge.

Ferguson, C. J. (2009). An Effect Size Primer: A Guide for Clinicians and Researchers. *Professional Psychology: Research and Practice, 40*(5), 532–538.

Gabler, S. (1994). Eine allgemeine Formel zur Anpassug an Randtabellen [A general formula for adjustment for contingency tables with known margins]. In S. Gabler, J. H. P.

Hoffmeyer-Zlotnik & D. Krebs (Eds.), *Gewichtung in der Umfragepraxis [Weighting in the Survey Practice]* (pp. 88-105). Westdeutscher Verlag

Gabler, S., & Ganninger, M. (2010). Gewichtung [Weighting]. In C. Wolf & H. Best (Eds.), *Handbuch der sozialwissenschaftlichen Datenanalyse [Handbook of the Data Analyses in the Social Sciences]* (pp. 143-164). Wiesbaden: VS Verlag.

Gabler, S., & Häder, S. (2002). Telefonstichproben. Methodische Innovationen und Anwendungen in Deutschland [Telephone samples: Methodological innovations and applications in Germany]. Münster: Waxmann Verlag.

Gabler, S., Häder, S., Lehnhoff, I. & Mardian, E. (2012). Weighting for unequal inclusion probabilities and nonresponse in dual frame telephone surveys. In S. Häder, M. Häder & M. Kühne (Eds.), *Telephone surveys in Europe: Research and practice* (pp. 147-167). Heidelberg: Springer.

Groves, R. M. (1989). *Survey errors and survey costs.* New York: Wiley.

Groves, R. M. & Couper, M. P. (1998). *Nonresponse in household interview surveys.* New York: Wiley.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). New York: John Wiley.

Hox, J. J., de Leeuw, E. D., & Zijlmans, E. A. O. (2015). Measurement equivalence in mixed-mode surveys. Frontiers in Psychology, 6(87). doi: 10.3389/fpsyg.2015.00087

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.

Klausch, T., Hox, J. J., & Schouten, B. (2014). *Evaluating mixed-mode redesign strategies against benchmark surveys: The case of the Crime Victimization Survey.* Paper presented at the GOR General Online Conference, 05-07.03.2014 in Cologne.

Lee, S. (2006). An Evaluation of Nonresponse and Coverage Errors in a Prerecruited Probability Web Panel Survey. *Social Science Computer Review, 24*, 460-475.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, California: Sage.

Little, R. J. A., & Wu, M. (1991). Models for contingency tables with known margins when target and sampled populations differ. *Journal of the American Statistical Association, 86(413)*, 87-95.

Lois, D. (2011). Kirchenmitgliedschaft und Kirchgangshäufigkeit im Zeitverlauf – Eine Trendanalyse unter Berücksichtigung von Ost-West-Unterschieden [Church membership and frequency of church attendance over time – A trend analysis taking into consideration differences in East and West Germany]. *Comparative Population Studies – Zeitschrift für Bevölkerungswissenschaft, 36(1),* 127-160.

Lyberg, L., & Kasprzyk, D. (1991). Data collection methods and measurement error: An overview. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), Measurement errors in surveys (pp. 237-257). New York: Wiley.

Mohorko, A., de Leeuw, E., & Hox, J. (2013). Journal of Officila Statistics, 29, 4, 609-622. http://www.degruyter.com/view/j/jos.2013.29.issue-4/jos-2013-0042/jos-2013-0042.xml

Schonlau, M. (2012). Stata-ado package "hotdeckvar" – Single hotdeck imputation. from http://www.schonlau.net/stata/

Schouten, B., Van den Brakel, J., Buelens, B., & Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. Social Science Research, 42, 1555-1570.

Tourangeau, R. (2013). Confronting the challenges of household surveys by mixing modes. Keynote address at the 2013 Federal Committee on Statistical Methodology Research Conference, Nov 4, 2013, Washington DC. http://www.copafs.org/UserFiles/file/fcsm/Tourangeau2013FCSMResearchKeynote.pdf

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The science of web surveys*. Oxford: Oxford University Press.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Wasmer, M., Scholz, E., Blohm, M., Walter, J., & Jutz, R. (2012). Konzeption und Durchführung der "Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften" (ALLBUS) 2010. [Concept and implementation of the German General Social Survey (ALLBUS) 2010] *GESIS Technical Report* (Vol. 2012/12). Mannheim: GESIS – Leibniz-Institute for the Social Sciences.

# Appendix

*Table A1* Distributions of the demographic variables used for weighting

| | GOPP Questionnaires 1-5 & 7 | | | ALLBUS 2010 Internet users | | | ALLBUS 2012 Internet users | | |
|---|---|---|---|---|---|---|---|---|---|
| Variable | % | SE | CI | % | SE | CI | % | SE | CI |
| Male | 53.1 | .015 | [50.1, 56.0] | 51.5 | .012 | [49.3, 53.8] | 51.6 | .010 | [46.4, 50.3] |
| Age groups | | | | | | | | | |
| 18-24 | 13.0 | .010 | [11.2, 15.1] | 13.9 | .008 | [12.4, 15.6] | 13.9 | .007 | [12.6, 15.3] |
| 25-34 | 19.8 | .012 | [17.6, 22.3] | 18.6 | .009 | [16.9, 20.4] | 17.4 | .008 | [15.9, 18.9] |
| 35-49 | 34.9 | .014 | [32.2, 37.8] | 35.4 | .011 | [33.2, 37.6] | 31.4[a] | .009 | [29.6, 33.3] |
| 50-64 | 23.4 | .013 | [21.0, 26.0] | 23.2 | .010 | [21.4, 25.2] | 27.7[a] | .009 | [26.0, 29.5] |
| 65+ | 8.8 | .008 | [7.3, 10.6] | 8.9 | .007 | [7.6, 10.2] | 9.6 | .006 | [8.5, 10.9] |
| Education | | | | | | | | | |
| low | 12.5 | .010 | [10.7, 14.6] | 21.4 | .010 | [19.6, 23.4] | 20.9 | .008 | [19.4, 22.6] |
| medium | 30.0 | .014 | [27.4, 32.7] | 37.8 | .011 | [35.6, 40.0] | 39.4 | .010 | [37.5, 41.3] |
| high | 57.5 | .015 | [54.6, 60.4] | 40.8 | .011 | [38.5, 43.0] | 39.7 | .010 | [37.8, 41.6] |

[a] significant differences ($p<0.01$) between ALLBUS 2010 and 2012. N for GOPP is 1114 respondents, which includes the respondents who completed at least one of the questionnaires 1 to 5, or 7. SE is short for standard error, CI is short for confidence interval.

*Table A2* Distributions of demographic characteristics in the GOPP over questionnaires (percentages)

| Variable/Q | 1 | 2 | 3 | 4 | 5 | 7 |
|---|---|---|---|---|---|---|
| Male | 52.97 (.016)<br>[49.9, 56.1] | 52.74 (.017)<br>[49.4, 56.1] | 51.75 (.018)<br>[48.3, 55.2] | 51.48 (.018)<br>[48.0, 55.0] | 51.91 (.018)<br>[48.4, 55.5] | 51.99 (.019)<br>[48.4, 55.6] |
| Age groups | | | | | | |
| 18–24 | 13.37 (.011)<br>[11.4, 15.6] | 12.53 (.011)<br>[10.5, 15.0] | 12.12 (.012)<br>[10.0, 14.6] | 12.52 (.012)<br>[10.4, 15.0] | 13.27 (.012)<br>[11.0, 15.9] | 12.35 (.012)<br>[10.1, 14.9] |
| 25–34 | 19.31 (.012)<br>[17.0, 21.9] | 19.21 (.014)<br>[16.7, 22.0] | 19.25 (.014)<br>[16.7, 22.1] | 18.45 (.014)<br>[15.9, 21.3] | 18.53 (.014)<br>[15.9, 21.5] | 19.20 (.015)<br>[16.5, 22.2] |
| 35–49 | 34.85 (.015)<br>[32.0, 37.8] | 34.73 (.016)<br>[31.6, 38.0] | 35.00 (.017)<br>[31.8, 38.3] | 35.10 (.017)<br>[31.8, 38.5] | 34.03 (.017)<br>[30.7, 37.5] | 34.29 (.018)<br>[30.9, 37.8] |
| 50–64 | 23.47 (.013)<br>[20.9, 26.2] | 24.22 (.015)<br>[21.4, 27.2] | 24.50 (.015)<br>[21.6, 27.6] | 24.65 (.015)<br>[21.7, 27.8] | 24.97 (.017)<br>[22.0, 28.2] | 24.97 (.016)<br>[22.0, 28.2] |
| 65+ | 9.00 (.009)<br>[7.4, 10.9] | 9.31 (.010)<br>[7.5, 11.5] | 9.13 (.010)<br>[7.3, 11.3] | 9.28 (.010)<br>[7.4, 11.6] | 9.20 (.010)<br>[7.3, 11.5] | 9.19 (.011)<br>[7.3, 11.5] |
| Education | | | | | | |
| low | 12.18 (.010)<br>[10.3, 14.3] | 11.81 (.011)<br>[9.8, 14.2] | 10.25 (.011)<br>[8.3, 12.6] | 10.06 (.011)<br>[8.1, 12.4] | 10.25 (.011)<br>[8.3, 12.6] | 10.01 (.011)<br>[8.0, 12.4] |
| medium | 29.50 (.014)<br>[26.8, 32.4] | 29.00 (.016)<br>[26.0, 32.2] | 28.00 (.016)<br>[25.0, 31.2] | 28.52 (.016)<br>[25.4, 31.8] | 27.73 (.016)<br>[24.7, 31.0] | 27.57 (.017)<br>[24.4, 30.9] |
| high | 58.32 (.016)<br>[55.2, 61.3] | 59.19 (.017)<br>[55.8, 62.5] | 61.75 (.017)<br>[58.3, 65.1] | 61.42 (.017)<br>[57.9, 64.8] | 62.02 (.018)<br>[58.5, 65.4] | 62.41 (.018)<br>[58.8, 65.9] |
| missing on age | 2.08 | 1.55 | 1.88 | 1.68 | 1.45 | 1.37 |
| missing on education | 8.02 | 3.34 | 3.88 | 3.75 | 3.82 | 3.71 |
| missing design weight | 7.13 | 1.91 | 2.75 | 2.84 | 2.50 | 2.47 |
| N | 1010 | 838 | 800 | 775 | 761 | 729 |

*Note*: standard errors in parentheses, 95%-confidence intervals in brackets. The data are not weighted. Q is short for GOPP questionnaire, SE is short for standard error, CI is short for confidence interval.

*Table A3* Post-stratification weights obtained with IPF-Weighting

| Questionnaire | 1 | 2 | 3 | 4 | 5 | 7 |
|---|---|---|---|---|---|---|
| N | 1010 | 838 | 800 | 775 | 761 | 729 |
| Min | 0.585 | 0.539 | 0.520 | 0.510 | 0.507 | 0.498 |
| Max | 2.299 | 2.367 | 2.445 | 3.268 | 2.956 | 3.086 |
| 1% percentile | 0.585 | 0.539 | 0.520 | 0.510 | 0.507 | 0.498 |
| 99% percentile | 2.044 | 2.137 | 2.907 | 2.514 | 2.528 | 2.558 |

## Calculation of the effect sizes:

Formulas used for calculating the unweighted effect sizes (Lipsey & Wilson, 2001, p. 172 ff.):
The standardized mean difference effect size ($ES_{sm}$):

$$ES_{sm} = \frac{\overline{X}_1 - \overline{X}_2}{s_{pooled}}, \; s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

The approximations based on dichotomous data: $ES_{sm} = arcsine(p_1) - arcsine(p_2)$, where $p$ is the proportion for each group (Lipsey & Wilson, 2001, p. 198-200).
Formulas used for calculating the mean effect sizes:

$$\overline{ES} = \frac{\sum_{i=1}^{k}(w_i \times ES_i)}{\sum_{i=1}^{k} w_i}; \; w_i = \frac{1}{\overline{SE}_i^2},$$

where $w$ is the inverse variance weight, $ES$ is the unweighted effect size *(d)*, and $SE$ is the standard error of the difference between the effect sizes (Lipsey & Wilson, 2001, p. 114).

For the calculation of the effect sizes the effect size calculator at http://www.campbellcollaboration.org/resources/effect_size_input.php was used.
We used the unweighted sample sizes to calculate the mean effects.

## Questions used in the analysis

Asterisk marks the questions where a card with answer options was offered to the respondents by interviewers in ALLBUS interviews. Question wordings are provided as translated by ALLBUS team, original German question wordings were identical among surveys unless otherwise noted.

**German economy***

How would you generally rate the current economic situation in Germany?

*Very good*
*Good*
*Partly good/partly bad*
*Bad*
*Very bad*

**Own financial situation***

And your own current financial situation?

*Very good*
*Good*
*Partly good/partly bad*
*Bad*
*Very bad*

**German economy in 1 year***

What do you think the economic situation in Germany will be like in one year?

*Considerably better than today*
*Somewhat better than today*
*The same*
*Somewhat worse than today*
*Considerably worse than today*

**Own financial situation in 1 year***

And what will your own financial situation be like in one year?

*Considerably better than today*
*Somewhat better than today*
*The same*
*Somewhat worse than today*
*Considerably worse than today*

**General health***

A question about your health: How would you describe your health in general?

*Very good*
*Good*
*Satisfactory*
*Poor*
*Bad*

**Employment status***

And now let's continue with employment and your occupation. Which of the categories on the card applies to you?

*Full time employment*
*Part (''half'') time employment*
*Less than part (''half'') time employment*
*Not working*

If there are difficulties referring to the classification, here are some hints for you:
Trainees are considered employees in a regular occupation.
Family members assisting in a family business who are full- time or part- time (''halftime'') employees in the business of a household or a family member, without having a formal contract, are also considered employees in a regular occupation (either full- time or part- time).

''Employed less than part- time'' are persons who are gainfully employed while, at the same time, one of the following applies:
Attend a full- time school (pupils and students),
Are registered as unemployed or
Draw a retirement benefit / pension as a result of previous employment.
Persons on maternity / parental leave or on another type of leave of absence are not considered employees in a regular occupation.

**Marital status***

What is your marital status? Are you...

*Married and living with your spouse*
*Married and living apart*
*Widowed*
*Divorced*
*Never married*

*Civil partnership, living together*
*Civil partnership, living apart*
*Registered partner deceased*
*Civil partnership dissolved*

*Note that in GOPP this question involved a filter: "civil partnership" category was added to the first five answer options and followed by the four civil partnership answer choices in case a respondent selected the category "civil partnership".*

**Religiosity***

Would you describe yourself as tending to be religious or tending to be not religious?

*(1) "religious"*

*(10) "not religious"*

*Note: 10pt-scale, in ALLBUS end labels "religious", "not religious" and random letters for scale points, in GOPP end labels only.*

**Confession***

May I ask what religious confession you belong to?

*The Roman Catholic church*

*The German Protestant church (excluding free churches)*

*A Protestant free church*

*Another Christian denomination*

*Another non-Christian religion*

*No religious affiliation*

*Note: In GOPP "May we ask what religious confession you belong to?"*

**Frequency of church attendance**

As a rule, how often do you go to church?

*More than once a week*

*Once a week*

*Between one and three times a month*

*Several times a year*

*Less*

*Never*

*Note: in ALLBUS 2012 the question is split into two: "How often go to church?" for respondents who belong to a Christian religious denomination and "How often go you go to church or to mosque, synagogue or other chapel?"*

**Born in Germany**

Were you born within the current borders of Germany?

*Yes*

*No*

**Citizenship**
What citizenship do you have? If you have several citizenships, please name all of them.

*Germany*
*Greece*
*Italy*
*Former Yugoslavia:*
*Bosnia and Herzegovina,*
*Serbia,*
*Croatia,*
*Macedonia,*
*Slovenia*
*Poland*
*Turkey*
*Other country, please enter* _____
*None, stateless*

**Self-assessed social class**
There is a lot of talk about social class these days. What class would you describe yourself as belonging to?

*Lower class*
*Working class*
*Middle class*
*Upper middle class*
*Upper class*

**Ownership of dwelling***
The next question deals with the accommodation you/your family live in. Please tell me which of the categories on the card applies to you/your family.

*Sublet*
*In an official/company flat*
*In subsidized municipal housing*
*In a rented flat (not subsidized housing)*
*In a rented house (detached/semi-detached)*
*In a flat owned by you or your family*
*In a house owned by you or your family*
*Other type of accommodation, please enter* _____

**Attitudes**

I'm going to read you some statements now. Please tell me after each one whether you have the same or a different opinion.

No matter what some people say, life for ordinary people is getting worse rather than better.

With the future looking as it does, it's almost irresponsible to bring children into the world.

Most politicians are not really interested at all in the problems of ordinary people.

Most people don't really care in the slightest what happens to others.

*Have the same opinion*
*Have a different opinion*

*Note: question wording for GOPP: "Please indicate for each statement whether you have the same or a different opinion."*

**Left-right orientation***

Many people use the terms ''left'' and ''right'' when they want to describe different political views. Here we have a scale which runs from left to right. Thinking of your own political views, where would you place these on this scale?

*(1) "left"*
*(10) "right"*

*Note: 10pt-scale, in ALLBUS end labels "left", "right" and random letters for scale points, in GOPP end labels only. The text "Here we have a scale which runs from left to right." omitted in GOPP.*

*Table A4* Comparison of the GESIS Online Panel Pilot to the subsamples of Internet users from two ALLBUS surveys: attitudinal and factual questions with post-stratification, no imputation of variables education and age

| | Q | GOPP | | | ALLBUS 2010 Internet users | | | ALLBUS 2012 Internet users | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | | Estimate | SE | 95% CI | Estimate | SE | 95% CI | Estimate | SE | 95% CI |
| *Attitudinal:* | | | | | | | | | | |
| German economy | 1 | 3.788 | 0.121 | [3.551, 4.025] | 3.086*** | 0.019 | [3.048, 3.124] | ***3.351****** | 0.018 | [3.317, 3.385] |
| German economy in 1 year | 1 | 2.427 | 0.242 | [1.952, 2.902] | 3.077** | 0.021 | [3.036, 3.117] | ***2.667*** | 0.016 | [2.636, 2.698] |
| Self-rated financial situation | 1 | 3.210 | 0.120 | [2.974, 3.445] | 3.471* | 0.020 | [3.433, 3.510] | ***3.593**** | 0.017 | [3.560, 3.626] |
| Self-rated financial situation in 1 year | 1 | 3.112 | 0.066 | [2.982, 3.241] | 3.182 | 0.017 | [3.148, 3.216] | 3.157 | 0.015 | [3.128, 3.186] |
| Health status | 1 | 3.874 | 0.076 | [3.725, 4.024] | 3.823 | 0.023 | [3.778, 3.868] | ***3.748*** | 0.020 | [3.709, 3.787] |
| Religiosity | 4 | 5.734 | 0.145 | [4.334, 7.134] | 5.974 | 0.073 | [5.832, 6.117] | ***6.341*** | 0.063 | [6.218, 6.466] |
| Self-assessed social class | 5 | | | | | | | | | |
| lower class | | 3.60% | 0.011 | [1.96, 6.51] | 2.27% | 0.003 | [1.67, 3.06] | 1.55% | 0.003 | [1.13, 2.13] |
| working class | | 26.18% | 0.025 | [21.56, 31.39] | 20.23%* | 0.010 | [18.42, 22.17] | 22.21% | 0.009 | [20.55, 23.97] |
| middle class | | 58.11% | 0.025 | [53.09, 62.96] | 63.38% | 0.012 | [61.08, 65.63] | 63.24% | 0.010 | [61.21, 65.22] |
| upper middle class | | 11.65% | 0.012 | [9.50, 14.21] | 13.80% | 0.008 | [12.22, 15.55] | 12.36% | 0.007 | [11.03, 13.83] |
| upper class | | 0.46% | 0.002 | [0.18, 1.20] | 0.32% | 0.001 | [0.14, 0.73] | 0.64% | 0.002 | [0.38, 1.09] |
| Life is getting worse | 7 | 80.17% | 0.029 | [73.90, 85.24] | 82.00% | 0.009 | [80.10, 83.74] | ***75.58%*** | 0.009 | [73.70, 77.36] |
| Irresponsible to have children | 7 | 37.78% | 0.034 | [31.30, 44.73] | 37.35% | 0.012 | [35.10, 39.65] | ***31.06%*** | 0.010 | [29.17, 33.01] |
| Politicians not interested | 7 | 81.14% | 0.028 | [75.03, 86.03] | 77.42% | 0.010 | [75.37, 79.35] | ***73.56%**** | 0.010 | [71.64, 75.40] |

*Table A4 continued*

| | Q | GOPP | | | ALLBUS 2010 Internet users | | | ALLBUS 2012 Internet users | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | | Estimate | SE | 95% CI | Estimate | SE | 95% CI | Estimate | SE | 95% CI |
| People don't care about others | 7 | 62.74% | 0.034 | [55.82, 69.17] | 71.72%* | 0.011 | [69.55, 73.79] | ***68.63%*** | 0.010 | [66.64, 70.55] |
| Left-right orientation | 7 | 5.072 | 0.112 | [4.852, 5.291] | 5.143* | 0.041 | [5.063, 5.222] | ***4.946*** | 0.036 | [4.876, 5.016] |
| *Factual*: | | | | | | | | | | |
| Working for pay | 2 | 83.71% | 0.016 | [80.30, 86.62] | 72.41%*** | 0.011 | [70.26, 74.46] | 73.60%*** | 0.009 | [71.75, 75.37] |
| Legal marital status | 3 | 49.70% | 0.024 | [44.92, 54.49] | 56.89%** | 0.012 | [54.56, 59.19] | 56.23%* | 0.011 | [54.16, 58.28] |
| Confession | 4 | 85.53% | 0.082 | [61.73, 95.59] | 70.57% | 0.011 | [68.47, 72.59] | 71.29% | 0.009 | [69.46, 73.04] |
| Frequency of church attendance | 4 | 2.599 | 0.227 | [2.153, 3.044] | 2.301 | 0.029 | [2.244, 2.359] | 2.286 | 0.026 | [2.234, 2.337] |
| Born in Germany | 4 | 97.19% | 0.017 | [91.03, 99.16] | 87.44%*** | 0.008 | [85.76, 88.94] | 86.25%*** | 0.008 | [84.69, 87.67] |
| German citizenship | 4 | 99.07% | 0.006 | [96.53, 99.76] | 94.53%*** | 0.006 | [93.32, 95.52] | 93.50%*** | 0.006 | [92.33, 94.50] |
| Owner of dwelling | 5 | 45.31% | 0.024 | [40.59, 50.13] | 56.64%*** | 0.012 | [54.30, 58.95] | 58.06%*** | 0.010 | [56.99, 60.10] |

*Note*: *$p<.05$, **$p<.01$, ***$p<.001$. Differences between ALLBUS 2010 and 2012 in italic bold, Q short for GOPP questionnaire, SE short for standard error, CI short for confidence interval. N for GOPP (started): Q1=1010; Q2=838; Q3=800; Q4=775; Q5=761; Q7=729. Weighted data: ALLBUS 2010 design, GOPP & ALLBUS 2012 design and post-stratification. Significance tests are pairwise t-tests for continuous/ordinal variables or tests of proportions.

Datasets used in preparing the article:

**1 GESIS Online Panel Pilot (GOPP)**

Fieldwork: February 2011–May 2012
Data Collection agency: GESIS – Leibniz Institute for the Social Sciences
Archive Link to demographic variables dataset: http://info1.gesis.org/dbksearch19/SDesc2.asp?no=5590&tab=0&ll=10&notabs=&af=&nf=1&search=&search2=&db=E
DOI: 10.4232/1.11577
Archive Link to the first online questionnaire:
http://info1.gesis.org/dbksearch19/SDesc2.asp?no=5582&tab=0&ll=10&notabs=&af=&nf=1&search=&search2=&db=E
DOI: 10.4232/1.11570 (includes three datasets for three study parts).

Description: The GESIS Online Panel Pilot Study consists of an initial CATI-recruitment interview (probability based sample, including landline and mobile telephone numbers), followed by eight monthly online surveys. Respondents to the CATI-recruitment interview (ZA5581), who had agreed to participate in the online panel by providing their email addresses, received an email invitation with a link to the first online survey. In the following, the respondents received such invitations for participation every month for the total duration of eight months. The overall concept for the online part was to have a leading topic for each monthly questionnaire and to include some demographic questions in every wave in order not to burden respondents with unpleasant factual questions. The most important demographic information had been collected during the recruitment interview. (For participants of short interviews the demographic information, which was supposed to be collected on telephone during the recruitment interview, was collected online at a later time.) An integrated dataset of the demographic information of all waves (ZA5590) is provided. The average duration of each monthly questionnaire was conceptualized to be around fifteen minutes. Generally respondents have one month for filling out a questionnaire, but they also had the possibility to do so at a later stage. The last date for filling out the questionnaires for the first study was 31 December 2011, for the second and the third study it was 31 May 2012.
The design has both longitudinal and cross-sectional aspects: monthly waves are designed as cross-sections with various topics; the longitudinal component includes survey evaluation questions, which are asked at the end of every monthly questionnaire. Most of the questions used in the online questionnaires were originally conducted within other German and international surveys. The reasons for replicating the questions were twofold: first,

in the course of the project the feasibility of asking questions via the Internet which are of interest for the social sciences were assessed; second, for the purposes of data quality assessment, comparisons with external benchmarks were planned as part of the project. For these reasons, most questions stem from the German General Social Survey (ALLBUS) 2010 and the German version of the European Social Survey 2010, large cross-sectional face-to-face surveys conducted by GESIS. Some other surveys, from which questions were implemented in the course of the panel, include questions of the German Socio-economic Panel (GSOEP 2008), the Eurobarometer, and the International Social Survey Programme (ISSP). Borrowed questions were generally implemented using the original question wording. In some cases these were tailored to the self-administered mode.
The main topics of the monthly questionnaires were: Wave 1: Multitopic introductory wave (ZA5582), Wave 2: Education and employment (ZA5583), Wave 3: Family life (ZA5584), Wave 4: Religion and values (ZA5585), Wave 5: Ecology (ZA5586), Wave 6: Social networks (ZA5587), Wave 7: Politics (ZA5588), Wave 8: Multitopic (focus on personality) (ZA5589).

Link to the study description:
http://info1.gesis.org/dbksearch19/download.asp?db=E&id=48472

**2 The German General Social Survey (ALLBUS) 2010**

The following information about the study was taken from the webpages of the study.
Fieldwork: May 2010–November 2010
Data Collection agency: TNS Infratest Sozialforschung, Munich
Archive Link: http://www.gesis.org/en/allbus/data-access/ and http://info1.gesis.org/dbksearch13/SDesc.asp?nf=1&search=ALLBUS&field=TI&DB=E&sort=dbk_ext.SN+DESC&maxRec=100&ll=10&tab=0& choose study number "ZA4612"
DOI: 10.4232/1.10760
Description: The German General Social Survey (ALLBUS) is a biennial survey that has been conducted since 1980 on the attitudes, behaviour, and social structure of persons resident in Germany. A representative cross-section of the population is questioned using face-to-face interviews. As a service to social scientific research and teaching, ALLBUS data are disseminated to all interested persons and institutions as soon as the data and documentation have been prepared. ALLBUS is a substantively rich and methodologically sophisticated database which can be used for a variety of analytical purposes:

- to describe and analyse attitudes, behaviour and social structure of Germans with up-to-date cross-sectional data,
- for longitudinal analysis of German society (ALLBUS time series, replicated questions from other survey studies),
- for international comparative analysis (ALLBUS includes questions also asked in the American General Social Survey (GSS) and in the ISSP),
- to examine methodical issues (e.g., checking for context effects, question effects, nonresponse).

Link to the study description: http://www.gesis.org/en/allbus/general-information/; see also http://www.gesis.org/en/allbus/study-profiles/2010/

## 3 The German General Social Survey (ALLBUS) 2012

The following information about the study was taken from the webpages of the study.
Fieldwork: April 2012–September 2012
Data Collection agency: TNS Infratest Sozialforschung, Munich
Archive Link: http://www.gesis.org/en/allbus/data-access/ and http://info1.gesis.org/dbksearch13/SDesc.asp?nf=1&search=ALLBUS&field=TI&DB=E&sort=dbk_ext.SN+DESC&maxRec=100&ll=10&tab=0& choose study number "ZA4614"
DOI: 10.4232/1.11753
Link to the study description: http://www.gesis.org/en/allbus/study-profiles/2012/

*Table A5* Variable labels, value labels and recoding procedures for variables used in the analyses

| | GOPP | | ALLBUS 2010 | | ALLBUS 2012 | | |
|---|---|---|---|---|---|---|---|
| Variable | Variable label | Value label | Variable label | Value label | Variable label | Value label | How the variable was recoded for analysis |
| Current economic situation in Germany | h1u1gereco<br>h2u1gereco<br>h3u1gereco | 1 Very good<br>2 Good<br>3 Partly good/partly bad<br>4 Bad<br>5 Very bad<br>. Missing | v79 | 1 Very good<br>2 Good<br>3 Partly good/partly bad<br>4 Bad<br>5 Very bad<br>8 Don't know<br>9 Missing (No answer) | v118 | 1 Very good<br>2 Good<br>3 Partly good/partly bad<br>4 Bad<br>5 Very bad<br>8 Don't know<br>9 Missing (No answer) | 5 Very good<br>4 Good<br>3 Partly good/partly bad<br>2 Bad<br>1 Very bad<br>. Missing (ALLBUS 8, 9) |
| Own current financial situation | h1u1slfeco<br>h2u1slfeco<br>h3u1slfeco | 1 Very good<br>2 Good<br>3 Partly good/partly bad<br>4 Bad<br>5 Very bad<br>. Missing | v80 | 1 Very good<br>2 Good<br>3 Partly good/partly bad<br>4 Bad<br>5 Very bad<br>8 Don't know<br>9 Missing (No answer) | v119 | 1 Very good<br>2 Good<br>3 Partly good/partly bad<br>4 Bad<br>5 Very bad<br>8 Don't know<br>9 Missing (No answer) | 5 Very good<br>4 Good<br>3 Partly good/partly bad<br>2 Bad<br>1 Very bad<br>. Missing (ALLBUS 8, 9) |
| Economic situation in Germany in one year | h1u1gereco1<br>h2u1gereco1<br>h3u1gereco1 | 1 Considerably better than today<br>2 Somewhat better than today<br>3 The same<br>4 Somewhat worse than today<br>5 Considerably worse than today<br>. Missing | v81 | 1 Considerably better than today<br>2 Somewhat better than today<br>3 The same<br>4 Somewhat worse than today<br>5 Considerably worse than today<br>8 Don't know<br>9 Missing (No answer) | v120 | 1 Considerably better than today<br>2 Somewhat better than today<br>3 The same<br>4 Somewhat worse than today<br>5 Considerably worse than today<br>8 Don't know<br>9 Missing (No answer) | 5 Considerably better than today<br>4 Somewhat better than today<br>3 The same<br>2 Somewhat worse than today<br>1 Considerably worse than today<br>. Missing (ALLBUS 8, 9) |

| | GOPP | | ALLBUS 2010 | | ALLBUS 2012 | | |
|---|---|---|---|---|---|---|---|
| Variable | Variable label | Value label | Variable label | Value label | Variable label | Value label | How the variable was recoded for analysis |
| Own economic situation in one year | h1u1slfeco1<br>h2u1slfeco1<br>h3u1slfeco1 | 1 Considerably better than today<br>2 Somewhat better than today<br>3 The same<br>4 Somewhat worse than today<br>5 Considerably worse than today<br>. Missing | v82 | 1 Considerably better than today<br>2 Somewhat better than today<br>3 The same<br>4 Somewhat worse than today<br>5 Considerably worse than today<br>8 Don't know<br>9 Missing (No answer) | v121 | 1 Considerably better than today<br>2 Somewhat better than today<br>3 The same<br>4 Somewhat worse than today<br>5 Considerably worse than today<br>8 Don't know<br>9 Missing (No answer) | 5 Considerably better than today<br>4 Somewhat better than today<br>3 The same<br>2 Somewhat worse than today<br>1 Considerably worse than today<br>. Missing (ALLBUS 8, 9) |
| General health | h1u1health<br>h2u1health<br>h3u1health | 1 Very good<br>2 Good<br>3 Satisfactory<br>4 Poor<br>5 Bad<br>. Missing | v398 | 1 Very good<br>2 Good<br>3 Satisfactory<br>4 Poor<br>5 Bad<br>9 No answer | v268 | 1 Very good<br>2 Good<br>3 Satisfactory<br>4 Poor<br>5 Bad<br>9 No answer | 5 Very good<br>4 Good<br>3 Satisfactory<br>2 Poor<br>1 Bad<br>. Missing (ALLBUS 9) |
| Employment status | h1u2employed<br>h2u2employed<br>h3u2employed | 1 Full time employment<br>2 Part ("half") time employment<br>3 Less than part ("half") time employment<br>4 Not working<br>. Missing | v340 | 1 Full time employment<br>2 Part ("half") time employment<br>3 Less than part ("half") time employment<br>4 Not working<br>9 Missing (No answer) | v245 | 1 Full time employment<br>2 Part ("half") time employment<br>3 Less than part ("half") time employment<br>4 Not working<br>9 Missing (No answer) | 1 Working for pay (1, 2, 3)<br>0 Not working for pay (4)<br>. Missing (ALLBUS 9) |

| | GOPP | | ALLBUS 2010 | | ALLBUS 2012 | | |
|---|---|---|---|---|---|---|---|
| Variable | Variable label | Value label | Variable label | Value label | Variable label | Value label | How the variable was recoded for analysis |
| Marital status | h1u3famstdu3<br>h2u3famstdu3<br>h3u3famstdu3 | 1 Married and living with a spouse<br>2 Married and living apart<br>3 Widowed<br>4 Divorced<br>5 Never married<br>6 Civil partnership<br>. Missing<br>.r Missing (Refusal to provide answer) | v405 | 1 Married and living with your spouse<br>2 Married and living apart<br>3 Widowed<br>4 Divorced<br>5 Never married<br>6 Civil partnership, living together<br>99 Missing (No answer) | v274 | 1 Married and living with your spouse<br>2 Married and living apart<br>3 Widowed<br>4 Divorced<br>5 Never married<br>6 Civil partnership, living together<br>99 Missing (No answer) | 1 Married (ALLBUS 1, 2, 6; GOPP variable famstdu3 1, 2 & variable elpstd 1, 2)<br>0 Not married (ALLBUS 3, 4, 5; GOPP variable famstdu3 3, 4, 5, 6 & variable elpstd 3, 4)<br>. Missing (ALLBUS 99) |
| | h1u3elpstd<br>h2u3elpstd<br>h3u3elpstd | 1 Civil partnership, living together<br>2 Civil partnership, living apart<br>3 Civil partnership, dissolved<br>4 Civil partnership, partner deceased<br>. Missing | | | | | |
| Religiosity | h1u4relig<br>h2u4relig<br>h3u4relig | (1) “religious”<br>(10) “not religious”<br>. Missing | v729 | (1) “religious”<br>(10) “not religious”<br>99 Missing (No answer) | v134 | (1) “religious”<br>(10) “not religious”<br>99 Missing (No answer) | (10) religious<br>(1) not religious<br>. Missing (ALLBUS 99) |

| | GOPP | | ALLBUS 2010 | | ALLBUS 2012 | | |
|---|---|---|---|---|---|---|---|
| Variable | Variable label | Value label | Variable label | Value label | Variable label | Value label | How the variable was recoded for analysis |
| Confession | h1u4religcomm1<br>h2u4religcomm1<br>h3u4religcomm1 | 1 The Roman Catholic church<br>2 The German Protestant church (excluding free churches)<br>3 A Protestant free church<br>4 Another Christian denomination<br>5 Another non-Christian religion<br>6 No religious affiliation<br>. Missing | v730 | 1 The Roman Catholic church<br>2 The German Protestant church (excluding free churches)<br>3 A Protestant free church<br>4 Another Christian denomination<br>5 Another non-Christian religion<br>6 No religious affiliation<br>7 Missing (Refused to answer)<br>9 Missing (No answer) | v188 | 1 The Roman Catholic church<br>2 The German Protestant church (excluding free churches)<br>3 A Protestant free church<br>4 Another Christian denomination<br>5 Another non-Christian religion<br>6 No religious affiliation<br>7 Missing (Refused to answer)<br>9 Missing (No answer) | 1 Confession named (1, 2, 3, 4, 5)<br>0 Confession not named (6)<br>. Missing (ALLBUS 7, 9) |
| Frequency of church attendance | h1u4church<br>h2u4church<br>h3u4church | 1 More than once a week<br>2 Once a week<br>3 Between one and three times a month<br>4 Several times a year<br>5 Less<br>6 Never<br>. Missing | v732 | 1 More than once a week<br>2 Once a week<br>3 Between one and three times a month<br>4 Several times a year<br>5 Less<br>6 Never<br>9 Missing (No answer) | v269 | 0 Not applicable<br>1 More than once a week<br>2 Once a week<br>3 Between one and three times a month<br>4 Several times a year<br>5 Less<br>6 Never<br>9 Missing (No answer) | 6 More than once a week<br>5 Once a week<br>4 Between one and three times a month<br>3 Several times a year<br>2 Less<br>1 Never<br>. Missing (ALLBUS 9) |
| Frequency of attendance of a temple | n/a | n/a | n/a | n/a | v270 | 0 Not applicable<br>1 More than once a week<br>2 Once a week | 6 More than once a week<br>5 Once a week<br>4 Between one and three times a month |

| | GOPP | | ALLBUS 2010 | | ALLBUS 2012 | | |
|---|---|---|---|---|---|---|---|
| Variable | Variable label | Value label | Variable label | Value label | Variable label | Value label | How the variable was recoded for analysis |
| | | | | | | 3 Between one and three times a month<br>4 Several times a year<br>5 Less<br>6 Never<br>9 Missing (No answer) | 4 Between one and three times a month<br>3 Several times a year<br>2 Less<br>1 Never<br>. Missing (ALLBUS 9) |
| Born in Germany | h1u4birth_area<br>h2u4birth_area<br>h3u4birth_area | 1 Yes<br>2 No | v303 | 1 Yes<br>2 No | v222 | 1 Yes<br>2 No | 1 Yes<br>0 No |
| Citizenship | h1u4cititzenship1<br>h2u4cititzenship1<br>h3u4cititzenship1 | 0 No German citizenship<br>1 Has German citizenship<br>. Missing | v6 | 1 Yes<br>2 Yes, along with other citizenship<br>3 No<br>4 Stateless<br>9 Missing (No answer) | v6 | 1 Yes<br>2 No | 1 Yes (ALLBUS 1, 2)<br>0 No (ALLBUS 3, 4)<br>. Missing (ALLBUS 9) |
| Self-assessed social class | h1u5slfses<br>h2u5slfses<br>h3u5slfses | 1 Upper class<br>2 Upper middle class<br>3 Middle class<br>4 Working class<br>5 Lower class | v33 | 1 Lower class<br>2 Working class<br>3 Middle class<br>4 Upper middle class<br>5 Upper class<br>6 None of the above<br>7 Missing (Refused to answer)<br>8 Don't know<br>9 Missing (No answer) | v17 | 1 Lower class<br>2 Working class<br>3 Middle class<br>4 Upper middle class<br>5 Upper class<br>6 None of the above<br>7 Missing (Refused to answer)<br>8 Don't know<br>9 Missing (No answer) | 1 Lower class (GOPP 5)<br>2 Working class (GOPP 4)<br>3 Middle class<br>4 Upper middle class (GOPP 2)<br>5 Upper class (GOPP 1)<br>. Missing (ALLBUS 6, 7, 8, 9) |

| | GOPP | | ALLBUS 2010 | | ALLBUS 2012 | | |
|---|---|---|---|---|---|---|---|
| Variable | Variable label | Value label | Variable label | Value label | Variable label | Value label | How the variable was recoded for analysis |
| Ownership of dwelling | h1u5owner<br>h2u5owner<br>h3u5owner | 1 Sublet<br>2 In an official /company flat<br>3 In subsidized municipal housing<br>4 In a rented flat (not subsidized housing)<br>5 In a rented house (detached/semi-detached)<br>6 In a flat owned by you or your family<br>7 In a house owned by you or your family<br>8 Other type of accommodation, please enter<br>. Missing | v714 | 1 Sublet<br>2 In an official/company flat<br>3 In subsidized municipal housing<br>4 In a rented flat (not subsidized housing)<br>5 In a rented house (detached/semi-detached)<br>6 In a flat owned by you or your family<br>7 In a house owned by you or your family<br>8 Other type of accommodation, please enter<br>99 Missing (No answer) | v501 | 1 Sublet<br>2 In an official/company flat<br>3 In subsidized municipal housing<br>4 In a rented flat (not subsidized housing)<br>5 In a rented house (detached/semi-detached)<br>6 In a flat owned by you or your family<br>7 In a house owned by you or your family<br>8 Other type of accommodation, please enter<br>99 Missing (No answer) | 1 Yes (6, 7)<br>0 No (1, 2, 3, 4, 5, 8)<br>. Missing (ALLBUS 99) |
| No matter what some people say, life for ordinary people is getting worse rather than better. | h1u7attitude1<br>h2u7attitude1<br>h3u7attitude1 | 1 Have the same opinion<br>2 Have a different opinion<br>. Missing | v61 | 1 Have the same opinion<br>2 Have a different opinion<br>8 Don't know<br>9 Missing (No answer) | v32 | 1 Have the same opinion<br>2 Have a different opinion<br>8 Don't know<br>9 Missing (No answer) | 1 Yes (ALLBUS 1)<br>0 No (ALLBUS 2)<br>. Missing (ALLBUS 8, 9) |

| | GOPP | | ALLBUS 2010 | | ALLBUS 2012 | | |
|---|---|---|---|---|---|---|---|
| Variable | Variable label | Value label | Variable label | Value label | Variable label | Value label | How the variable was recoded for analysis |
| With the future looking as it does, it's almost irresponsible to bring children into the world. | h1u7attitude2<br>h2u7attitude2<br>h3u7attitude2 | 1 Have the same opinion<br>2 Have a different opinion<br>. Missing | V62 | 1 Have the same opinion<br>2 Have a different opinion<br>8 Don't know<br>9 Missing (No answer) | v33 | 1 Have the same opinion<br>2 Have a different opinion<br>8 Don't know<br>9 Missing (No answer) | 1 Yes (ALLBUS 1)<br>0 No (ALLBUS 2)<br>. Missing (ALLBUS 8, 9) |
| Most politicians are not really interested at all in the problems of ordinary people. | h1u7attitude3<br>h2u7attitude3<br>h3u7attitude3 | 1 Have the same opinion<br>2 Have a different opinion<br>. Missing | V63 | 1 Have the same opinion<br>2 Have a different opinion<br>8 Don't know<br>9 Missing (No answer) | v34 | 1 Have the same opinion<br>2 Have a different opinion<br>8 Don't know<br>9 Missing (No answer) | 1 Yes (ALLBUS 1)<br>0 No (ALLBUS 2)<br>. Missing (ALLBUS 8, 9) |
| Most people don't really care in the slightest what happens to others. | h1u7attitude4<br>h2u7attitude4<br>h3u7attitude4 | 1 Have the same opinion<br>2 Have a different opinion<br>. Missing | V64 | 1 Have the same opinion<br>2 Have a different opinion<br>8 Don't know<br>9 Missing (No answer) | v35 | 1 Have the same opinion<br>2 Have a different opinion<br>8 Don't know<br>9 Missing (No answer) | 1 Yes (ALLBUS 1)<br>0 No (ALLBUS 2)<br>. Missing (ALLBUS 8, 9) |
| Left-right orientation | h1u7leftright<br>h2u7leftright<br>h3u7leftright | (1) "left"<br>(10) "right"<br>. Missing | v78 | (1) "left"<br>(10) "right"<br>99 Missing (No answer) | v101 | (1) "left"<br>(10) "right"<br>99 Missing (No answer) | . Missing (ALLBUS 99) |

| | GOPP | | ALLBUS 2010 | | ALLBUS 2012 | | |
|---|---|---|---|---|---|---|---|
| Variable | Variable label | Value label | Variable label | Value label | Variable label | Value label | How the variable was recoded for analysis |
| Sex of respondent | sex_cati<br>sex_onl | 1 Male<br>2 Female<br>. Missing | v298 | 1 Male<br>2 Female | v217 | 1 Male<br>2 Female | 1 Male<br>0 Female |
| Age of respondent/ Year of birth | ybirth_cati<br>age_onl | . Missing<br>. r Missing (Refusal to provide answer) | v301 | 999 Missing | v220 | 999 Missing | GOPP: 2011-ybirth_cati |
| Education: highest school leaving certificate | graduation_cati_h1 | 1 Still at school (general education)<br>2 Still at school (vocational education)<br>3 Finished school without school leaving certificate<br>4 Lowest formal qualification of 5 Germany's tripartite secondary school system, after 8 or 9 years of schooling<br>5 Intermediary secondary qualification, after 10 years of schooling<br>6 Secondary qualification from Eastern Germany (Polytechnic secondary school) | v327 | 1 Finished school without school leaving certificate<br>2 Lowest formal qualification of Germany's tripartite secondary school system, after 8 or 9 years of schooling<br>3 Intermediary secondary qualification, after 10 years of schooling<br>4 Certificate fulfilling entrance requirements to study at a polytechnical college<br>5 Higher qualification, entitling holders to study at a university<br>6 Other school leaving certificate<br>7 Still at school<br>99 Missing (No answer) | v230 | 1 Finished school without school leaving certificate<br>2 Lowest formal qualification of Germany's tripartite secondary school system, after 8 or 9 years of schooling<br>3 Intermediary secondary qualification, after 10 years of schooling<br>4 Certificate fulfilling entrance requirements to study at a polytechnical college<br>5 Higher qualification, entitling holders to study at a university<br>6 Other school leaving certificate<br>7 Still at school<br>99 Missing (No answer) | 1 Low (ALLBUS 1,2)<br>2 Middle (ALLBUS 3)<br>3 High (ALLBUS 4, 5)<br>. Missing (ALLBUS 6, 7, 99) |

| | GOPP | | ALLBUS 2010 | | ALLBUS 2012 | | |
|---|---|---|---|---|---|---|---|
| Variable | Variable label | Value label | Variable label | Value label | Variable label | Value label | How the variable was recoded for analysis |
| | | 7 Certificate fulfilling entrance requirements to study at a polytechnical college<br>8 Higher qualification, entitling holders to study at a university<br>9 Other school leaving certificate<br>. Missing<br>.r Missing (Refusal to provide answer) | | | | | |
| | graduation_cati_h2 | 1 Still at school<br>2 Finished school without school leaving certificate<br>3 Lowest formal qualification of Germany's tripartite secondary school system, after 8 or 9 years of schooling<br>4 Intermediary secondary qualification, after 10 years of schooling<br>5 Secondary qualification from Eastern Germany (Polytechnic secondary school, 8 or 9 years) | | | | | |

| | GOPP | | ALLBUS 2010 | | ALLBUS 2012 | | |
|---|---|---|---|---|---|---|---|
| Variable | Variable label | Value label | Variable label | Value label | Variable label | Value label | How the variable was recoded for analysis |
| | | 6 Secondary qualification from Eastern Germany (Polytechnic secondary school, 10 years)<br>7 Certificate fulfilling entrance requirements to study at a polytechnical college<br>8 Higher qualification, entitling holders to study at a university<br>9 Higher qualification, entitling holders to study at a university obtained later<br>10 Other school leaving certificate<br>. Missing<br>.r Missing (Refusal to provide answer) | | | | | |
| Person-related East-West-weight | n/a | n/a | v977 | | v743 | | |
| Internet use for private purposes | n/a | n/a | v772 | 1 Yes, I use the Internet for private purposes<br>2 No, I don't use the Internet | v530 | 1 Yes, I use the Internet for private purposes<br>2 No, I don't use the Internet | |

| | GOPP | | ALLBUS 2010 | | ALLBUS 2012 | | |
|---|---|---|---|---|---|---|---|
| Variable | Variable label | Value label | Variable label | Value label | Variable label | Value label | How the variable was recoded for analysis |
| Design weight | pweightcomb | | n/a | n/a | n/a | n/a | |
| Study part (CATI) | cati | 1 h1 (HS1)<br>2 h2 (HS2)<br>3 h3 (HS3) | n/a | n/a | n/a | n/a | |
| Study part (online) | study | 1 h1 (HS1)<br>2 h2 (HS2)<br>3 h3 (HS3)<br>. Missing | n/a | n/a | n/a | n/a | |
| Recruitment question at the end of the CATI interview | panelquestion | 1 Agreed to participate, provided Email<br>2 Respondent in doubt (homepage, call later)<br>3 No Email address<br>4 Not interested, don't want to participate<br>. Missing | n/a | n/a | n/a | n/a | |
| dispcode | Disposition code for the first online questionnaire | 12 Active (not started)<br>22 Break-off<br>31 Completed without pause<br>32 Completed with pause<br>. Missing | n/a | n/a | n/a | n/a | |
| pseudonym | Respondent ID (Online) | | n/a | n/a | n/a | n/a | |

methods, data, analyses | Vol. 9(1), 2015, pp. 57-86 DOI: 10.12758/mda.2015.002

# How Much Inequality of Earnings Do People Perceive as Just? The Effect of Interviewer Presence and Monetary Incentives on Inequality Preferences

*Stefan Liebig*[1], *Meike May*[1], *Carsten Sauer*[1], *Simone Schneider*[2] & *Peter Valet*[1]

1 *Bielefeld University*
2 *Trinity College Dublin*

**Abstract**

This paper describes two studies designed to test how two structural conditions of an interview situation – the presence of an interviewer and use of incentives – influence respondents' preferences regarding inequality. According to goal-framing theory and findings from empirical justice research, different goal frames are activated in different types of relationships, producing different distributional preferences: Cooperative situations induce a normative goal frame resulting in a stronger preference for equality whereas competitive situations induce a gain frame in which individuals have stronger preferences for inequality. Assuming the former type of relationship is established by the presence of an interviewer and the latter type by incentivizing, we conducted two studies to test our hypotheses. The results suggest that building a cooperative relationship through interviewer presence and cooperation priming leads to a preference for equality, while use of incentives leads to a clear preference for inequality.

*Keywords*: justice attitudes, inequality preferences, interviewer presence, incentives, priming, survey methodology

# 1 Introduction

Research on survey methodology shows that the mode of data collection affects respondents' willingness to participate and to provide information. Although survey errors such as coverage error, non-response error, and measurement error are not constant across survey modes, there are "typical" survey errors that are related to individual survey modes (Groves, 1989; Lindhjem & Navrud, 2011; Vannieuwenhuyze, Loosveldt, & Molenberghs, 2010). Some general findings indicate that certain topics, issues, and items are more likely to produce higher rates of survey errors; however, theories remain to be formulated that can predict response behaviors under certain structural interview conditions accurately. This inability to anticipate responses is especially relevant when it comes to understanding how different interview settings affect inequality preferences.

In order to identify measurement errors and to choose appropriate adjustment methods, it is essential to understand how certain structural conditions of an interview situation evoke different response sets. This is particularly important in public opinion and attitude research where aggregate measures are used to draw substantial conclusions about a society's sentiments and opinions regarding political and societal issues.

Drawing on the "logic of justice" and goal-framing theory (GFT) (Lindenberg, 2006), this paper attempts to explain how certain structural features of interview situations affect respondents' inequality preferences. We focus on two crucial interview conditions that are found in most large-scale population surveys: interviewer presence and conditional incentives. The logic of justice suggests that different principles of justice are associated with specific types of social relations (Tyler, Boeckmann, Smith, & Huo, 1997); accordingly, goods should be distributed equally in cooperative relationships and equitably in competitive relationships, and should correspond to individual needs in family and kinship relationships. If we apply these findings to large-scale survey research, inequality preferences should be affected by the kind of relationship respondents share in the situation in which they are completing the questionnaire.

---

*Direct correspondence to*
Stefan Liebig, Universität Bielefeld, Postfach 100131, 33501 Bielefeld
E-mail: stefan.liebig@uni-bielefeld.de

*Acknowledgment*: This article is an outcome of the research project "The Legitimation of Inequalities – Structural Conditions of Justice Attitudes over the Lifespan" which is funded by the German Research Foundation (DFG) within the Collaborative Research Center (Sonderforschungsbereich 882) "From Heterogeneities to Inequalities" at Bielefeld University, Germany. We thank two anonymous reviewers for helpful comments, critiques, and suggestions, and William White and René Reinholz for language editing.

By coupling the logic of justice with GFT, we argue that interviewer presence and use of conditional incentives establish different types of social relationships in an interview setting. These types of relationships activate certain cognitive frames which in turn lead to individual preferences for more or less inequality. The first assumption is that interviewers try to establish a cooperative situation with the respondent in order to increase response rates. The second assumption is that the use of conditional incentives as a way of increasing response rates establishes an (economic) exchange situation that has the same character as a competitive social relationship. Therefore, we hypothesize that respondents will prefer a lower degree of inequality (i.e., more equality) in the presence of an interviewer and a higher degree of inequality (i.e., less equality) if they are given a conditional incentive for completing the questionnaire.

To test these predictions, we conducted two experiments involving undergraduate students at a German university ($N_{Study1} = 145$; $N_{Study2} = 210$). We varied the interview situations experimentally by having an experimenter in the room, by displaying stylized eyes on the computer screen (see Figure A1 in Appendix), by introducing priming techniques, and by offering monetary incentives. In order to lower the risk of social-desirability bias, individual preferences for inequality were measured indirectly through the use of a vignette module that focused on fair earnings.

This paper is organized as follows: First, the concept of the logic of justice is introduced with reference to empirical justice research. We then report findings concerning the situational vulnerability of justice behavior and attitudes. The next section presents the theoretical model we used, which was based on GFT and our hypotheses concerning the situational dependence of (in)equality preferences. In the methods section, we describe the experimental design, measurement of variables, and sample issues. The paper concludes with a presentation and discussion of the findings.

# 2 Theoretical Background

## 2.1 The Logic of Justice

Because most of the resources we value in our lives are scarce, conflicts arise about the allocation and distribution of such goods. To solve this kind of conflict, several norms and principles have emerged during human cultural evolution (Krebs, 2008; Miller, 1999). As psychological justice research has shown, four principles are fundamental (Deutsch, 1975; Konow, 2001; Miller, 1999): *equality* (everybody should receive the same), *equity* (benefits and burdens should be distributed proportionally according to individual investments), *need* (everybody should get

enough to cover basic needs), and *entitlement* (people should receive either benefits or burdens based on their ascribed or achieved status characteristics, such as gender, education, occupation, or origin). One key empirical finding is that there is a homology between types of social relationships and the application of these justice principles. In very close long-term relationships, the principles of equality and need are preferred, whereas in competitive short-term relationships, the equity principle dominates, and in hierarchical relationships, entitlement is important (Törnblom, 1992). The stronger the subjects' social involvement in decisions concerning allocation (e.g., through direct personal contact with other subjects or the experimenter), the sooner subjects will base their decisions on the principles of equality or need (Konow, 2001). This applies even when respondents are asked to consider what would be a fair distribution of income in a society (Traub, Seidl, & Schmidt, 2009). In the early 1980s, researchers found that the type of social relationship determines which principle of justice is selected, but that there is also a reverse effect: the validity and application of the relevant principle of justice also has an impact on the nature of the social relationship. Equality and need-based rules strengthen close and long-term relationships, whereas rules based on the equity principle lead to the development of competitive short-term relationships (Schwinger, 1981).

The logic of justice evident in the homology between types of social relationship and the respective justice principles can be further differentiated by using a typology introduced by Alan P. Fiske (1991; this paragraph is based on Liebig & Sauer 2013, 2015). Fiske assumes that there are four ideal types of social relationships (Figure 1). The first type is driven by strong ties and long-term relationships between individuals. Here the individual is integrated into a community (*Gemeinschaft*; see Tönnies, 1887) in which all members share the same origin and identity. These communities define themselves as solidary communities in which helping one another is a given and will assure the survival of one's group or family and – from an evolutionary perspective – the survival of one's descendants and one's genes. Classic examples are families and clans based on kinship relations. In such relationships, the distribution rule that would be considered "just" is the need principle: all individuals get as much as they need to survive. The second type is characterized by hierarchical relationships. Examples include pre-modern hierarchical societies (*Ständegesellschaften*) and bureaucratic organizations with different hierarchical levels in which each level has authority over the subordinate levels. In these contexts, the responsibility of higher-ranking individuals is to give instructions to lower-ranking individuals. Higher-ranking individuals have the power to sanction subordinates should they not follow the instructions, and they take responsibility for the lower-ranking individuals, who recognize their superiors' authority in return. The corresponding principle of justice in these situations is entitlement. All individuals get what they are entitled to according to their position in the hierarchical structure. The third type is marked by an absence of rank differ-

| | |
|---|---|
| Type 1<br>Solidary communities:<br>*Need* | Type 2<br>Hierarchical relationships:<br>*Entitlement* |
| Type 3<br>Social-exchange relationships:<br>*Equality* | Type 4<br>Economic-exchange relationships:<br>*Equity* |

*Figure 1* The logic of justice: homology between types of social relationship and justice principles

ences. Examples include non-hierarchical networks, peer groups, and cooperatives. Members treat each other as equal despite individual differences, and everyone has the same rights and duties. Relationships are characterized primarily by mutual exchange in which – following strict reciprocity – the rewards and inputs between individuals are balanced. The dominant justice principle in this type of social relationship is equality. The fourth type is defined by short-term relationships between strangers and is typical of market relations. These are economic-exchange relationships in which individual actors offer goods and services for maximum personal benefit. The related justice principle here is equity.

The conclusion from this model is, first, that the meaning of justice is not based on any single principle that is superior to any other competing principle. Individuals can regard different distributive principles as just, applying each of them to different situations. Second, for each "ideal type" of social relationship there is a corresponding justice principle that constitutes what may be called the logic of justice. Therefore, the normative expectation is that goods will be distributed equally in social-exchange relations but equitably in economic-exchange situations. For our purposes, the crucial point is that within an interview situation, either type of social relationship may be established. If a respondent is offered a (conditional) incentive for filling out the questionnaire, an *economic-exchange situation* between the interviewer and the interviewee is established. Once there is no monetary or non-monetary (conditional) incentive, a *social-exchange relation* is established – respondents may invest their time in filling out a questionnaire because they think, for example, that it is their duty as a citizen, they see a chance to express their views, they want to get social approval from the interviewer, or they simply want to attract the interviewer. According to the logic of justice, different justice principles will be appropriate in both situations. In Section 2.2, we discuss research providing empirical evidence that the situational conditions of an experiment or an interview influence respondents' preferences regarding how goods should be distributed.

## 2.2 Research on the Situational Conditions of Justice Behavior and Attitudes

Studies in behavioral economics and social psychology have shown that the structural conditions of an interview situation affect the inequality preferences of respondents. These structural conditions are (1) framing with regard to the type of social relationship, (2) anonymity, and (3) incentivizing.

(1) *Framing of a situation:* In determining moves in an Iterated Prisoner's Dilemma game, Liberman, Samuels, and Ross (2004) framed the dilemma as either a competitive or a cooperative situation by labeling it "Wall Street Game" or "Community Game." Participants in the condition with the "Community Game" label were significantly more willing to cooperate with the other participants. Hertel and Fiedler (1994) primed their subjects by using evaluative and emotional priming with positive and negative connotations of cooperation and competition. As expected, the positive connotations of cooperation led the participants to cooperate more. Hole (2011) tested how priming would influence the decisions of dictators during the distribution phase by using a communication phase on fairness before a one-shot Dictator Game. Participants in the treatment group were asked during this communication phase what they thought might be a fair distribution. This resulted in significantly higher offers of the framed subjects (for an overview with precise descriptions of the games applied, see Fehr & Schmidt, 2006). This finding was supported by Liebig (2001) with regard to justice attitudes. Subjects who were given a detailed description of how philosophers understand justice and how justice attitudes must rely on moral judgment to overcome self-interest showed an attitude pattern on the justice of taxes that was less affected by self-interest. In sum, research shows the relational framing of a situation has an effect on the distributional decision-making process.

(2) *Anonymity:* Hoffman et al. (1994, 1996) varied the anonymity between participants in a Dictator Game. The more anonymous the situation was for the dictators, the less they were willing to give. Gächter and Fehr (1999) show similar results for investments in a public-good game. Communication between participants also led them to invest more in a trust game (Ellingsen & Johannesson, 2004) or to give the recipient more money in a Dictator Game (Xiao & Houser, 2009). Findings were similar when the family names of the players were revealed to the others (Charness & Gneezy, 2008) or the anonymity between subject and experimenter was varied (Bolton & Zwick, 1995).

There is some evidence that simply placing visual cues of eyes on a computer screen alters behavior and inequality preferences (Bateson, Nettle, & Roberts, 2006; Burnham & Hare, 2007; Haley & Fessler, 2005; Rigdon, Ishii, Watabe, & Kitayama, 2009). In dictator games, images of eyes presented to the dictator led to higher-than-average monetary gifts for the recipient (Haley & Fessler, 2005; Rig-

don et al., 2009). Similar results were found by Burnham and Hare (2007) in a public-goods game displaying the image of a robot with human eyes on the computer screen. Subjects in the "eyes" condition contributed significantly more (29% more) to the public good than did subjects in the control condition (Burnham & Hare, 2007). Bateson et al. (2006) displayed an image of a pair of eyes above an honesty box in a university coffee area and observed that people paid more for their coffee when they were "watched" by these eyes than in a control condition with neutral images. Their explanation of this finding is that cues of eyes subconsciously activate those brain regions responsible for detecting human faces, including gaze and facial expression. People are subconsciously aware of "being watched" or "being observed." Aside from these findings, a more recent study did not find any effects of an image or cues of eyes on altruistic or equality-oriented behavior (Vogt, Efferson, Berger, & Fehr, 2014).

(3) *Incentivizing:* In laboratory experiments, incentivizing is used in two ways: first as a "show-up" fee for participants and second as performance-related incentives incorporated into the experimental design. Average performance is not changed substantively by increased incentives, although the variance of responses often decreases (for a review, see Camerer & Hogarth, 1999, p. 9). Although increases in stakes in the dictator game apparently do not have a strong effect on the amount offered (Fehr & Schmidt, 2006, p. 625), people are less generous in allocation decisions when the stakes are real as opposed to hypothetical (Hertwig & Ortmann, 2001).

In survey research, the positive effect of incentives on willingness to participate and data quality is well documented (Becker & Mehlkop, 2011; Singer, Van Hoewyk, & Maher, 2000; Toepoel, 2012). However, the results for response bias are mixed. Most studies were not able to confirm incentive effects on responses (Becker & Mehlkop, 2011; James & Bolstein, 1990), but some found significant differences, at least for some attitudinal measures (James & Bolstein, 1990; Singer et al., 2000).[1] However, we know little about the effects of incentives on inequality preferences in surveys.

To summarize, research suggests that (1) inequality preferences and distributive behavior are influenced by the framing of a situation, (2) inequality preferences depend on the social relationship established between actors, and (3) incentives shift behavior and inequality preferences to more selfish allocation strategies and higher-inequality preferences. Although each of the studies we referred to provides some theoretical reasons for the empirical results reported, there is no theoretical model that allows one to deduce testable assumptions as to which situational conditions we may expect to result in which inequality preferences. One theory

1 James and Bolstein (1990) found that with larger monetary incentives, respondents expended greater effort in completing the questionnaire and made more favorable comments about the survey sponsor.

that could potentially serve as such a general theoretical model is Lindenberg's GFT because it clarifies the theoretical link between the situational conditions and inequality preferences and helps to formulate specific hypotheses regarding the structural conditions in which people prefer more or less inequality in society.

## 2.3 Using GFT to Explain Inequality Preferences

Following the seminal work of Kahneman and Tversky (1984, 2000) the literature on framing processes shows that when people make decisions they pay selective attention to situational cues. They retain cues that give them the information they need to pursue their current motivational goal while suppressing other information. At the same time, certain situational cues will activate specific information and knowledge in an individual's memory. Both selective attention and the activation process are governed by a cognitive-motivational – or "framing" – process. This process includes the mechanism by which motivational goals influence the cognitive processes of actors and their mental models of a given situation. A mental model is particularly relevant in social relationships because it contains information about the prototype of a relationship and the behavioral rules, expectations, and social norms it involves. Therefore, it helps to know what kind of behavior is appropriate or expected in a specific situation. Actors use situational cues in order to define the kind of situation or relationship they are confronting and to act appropriately and efficiently according to their goals.

As part of a general theory of human action (Lindenberg, 1990, 2001, 2006), GFT distinguishes three types of "master goal frames" and forms the core motivation for behavior: a *normative frame*, the goal being "to act appropriately"; a *gain frame*, the goal being "to increase one's resources"; and a *hedonic frame*, the goal being "to feel better" (Lindenberg, 2006). Each of the three frames has a different a priori strength to govern behavior. The hedonic frame is considered the strongest because it is closest to the psychological self and to emotions. The gain frame is considered second in strength, and the normative frame is the weakest because it depends largely on the support of other motivational goals that require the same behavior.

In a given situation, the master frame and mental model that are active depend on the structural conditions of that situation, an individual's cultural knowledge (e.g., what rules exist for solving distributional conflicts), and, according to the more general theory of social production functions (Ormel, Lindenberg, Steverink, & Verbrugge, 1999), the actual level of physical and social well-being. We assume that the two basic structural conditions within an interview or experiment – i.e., the presence of an interviewer and use of incentives – will determine which of the master goal frames will be activated. Once a goal frame is activated, it is part of an individual's "cultural" knowledge to know the "right" mental model for behav-

ing appropriately. Here, the findings from justice research come into play: the corresponding mental model is the logic of justice in which an ideal type of social relationship connects with one of the four general justice principles.

In order to derive empirical assumptions about how the structural conditions of an interview situation result in "biases" in inequality preferences, we must (1) consider what type of social relationship is established within a survey interview and (2) ask what kind of expected behavior or attitude is related to it (Lindenberg, 2006). In a non-incentivized survey or experiment, respondents receive no money for participating yet they bear costs in terms of time and answering cognitively demanding or personal questions. Assuming that other motivations to participate – curiosity, entertainment, or sympathy of the interviewer – cannot fully explain the response rates, non-incentivized surveys or experiments basically rely on the willingness of the participant to cooperate in order to produce a public good. Such cooperation is supported by interviewers in that they often try to motivate respondents to participate by asking for "help" or "support." Under these conditions, GFT assumes that the normative goal frame will be activated and a respondent's goal is to act appropriately. The behavior associated with a normative goal frame is prosocial behavior, in which a person is prepared to bear costs to benefit others because it is appropriate to do so.[2] According to the logic of justice, the dominant norm of allocating or distributing resources within a cooperative relationship is the equality principle. Under the condition of an activated normative goal frame, we expect respondents, when asked to evaluate inequalities, to formulate their judgments in light of the equality principle. Hence, respondents should reveal greater preferences for equal distribution in non-incentivized interview situations. The presence of an interviewer will confirm this equality orientation because the interviewer tries actively to establish a cooperative relationship; the interviewer will also be perceived as an agency controlling whether or not respondents are acting appropriately.

Incentivizing participation in a survey establishes a social relationship in which respondents would be oriented more toward their self-interest. Giving money *conditionally* for completing the questionnaire establishes an economic-exchange relationship, that is, respondents are rewarded with a certain amount of money for investing their time and effort. Under these conditions, GFT assumes that respondents frame the situation according to their personal gain. Following the logic of justice, the principle required to solve distributional conflicts within economic-exchange situations is the equity principle; individuals receive a distribution according to their individual contributions. Hence, we would expect respondents with an activated gain frame to assess inequalities on the basis of the equity principle, which leads to greater inequality preferences.

2 Prosocial behavior can also be observed in the gain or hedonic frame, but the motivation is a different one. In the gain frame, people act prosocially when it is an efficient means to increase gain, and they do so in the hedonic frame when it feels good.

Our baseline assumption is that the type of social relationship established in an interview setting will determine the activation of a specific cognitive frame that in turn will influence the response behavior (in this case, inequality preferences). In particular, we assume a cooperative relationship to activate a normative frame that induces preferences for equal distributions, whereas an economic-exchange situation or a competitive relationship (in contrast to a cooperative relationship) is likely to activate the gain frame, which induces greater inequality preferences. These structural conditions may cause biases in response behavior in survey research, especially regarding attitudes toward just distributions. We therefore expect two situational cues to be of specific importance: (a) the presence of others (e.g., an interviewer) and (b) use of incentives. Four hypotheses are derived from the theoretical reasoning outlined above. The first two are more general assumptions related to the logic of justice and the stated correlation between type of social relationship and inequality preference:

- ($H_1$) In *cooperative* situations, the equality principle is the predominant justice norm, and therefore a more *equal* distribution of resources will be preferred.
- ($H_2$) In *competitive* situations, the equity principle is the predominant justice norm, and therefore a more *unequal* distribution of resources will be preferred.

From these two hypotheses follow two more specific hypotheses about the influence of the structural conditions within an interview situation and the normative preferences:

- ($H_3$) In *interviewer-assisted* situations, the *equality* orientation is stronger compared to situations in which the interviewer is not present.
- ($H_4$) In *incentivized* interview situations, the *inequality* orientation is stronger compared to situations that are not incentivized.

# 3 Empirical Design

In order to test our hypotheses, we needed an appropriate research design and an accurate measure of inequality preferences. We opted for a laboratory experiment (Study 1) and an experimental survey study (Study 2). The laboratory experiment had several advantages, one of which was that it allowed us to test the direct effect of a factor on the dependent variable in an artificial situation in which we were able to control for other factors. Furthermore, the random assignment of subjects to either the control group or the treatment group ensured that no external traits of the subjects would influence the measured effect (see also Webster & Sell, 2007, p. 12). In addressing our research question, the experiment provided the opportunity to control the interview situation in terms of presence of other individuals and interviewer behavior. The second study was an experimental survey study. Respondents

were randomly assigned to treatments. In both studies, we chose from among three possible experimental conditions: (a) the activation of a specific mindset (cooperation vs. neutrality, and cooperation vs. competition) through *priming* techniques ($H_1$, $H_2$), (b) the *presence of another person* while a self-administered questionnaire was completed ($H_3$), and (c) use of conditional *incentives* for participation ($H_4$).

***Priming:*** In order to activate the normative or gain goal frame subliminally, we used the scrambled-sentence test as a priming technique. Originally introduced by Srull and Wyer (1979), this test appeared to be the most appropriate technique for our study. Before being presented with the actual questionnaire, respondents were asked to participate in a cognitive language test (which we called a *Sprachfertigkeitsübung*). They had to build logical sentences out of a given number of word sets. We expected to activate either a "cooperative" or a "competitive" mindset through the use of specific words associated with the two mindsets evoked. We varied the priming conditions in the two experiments and tested cooperation priming versus a neutral control group in Study 1 and cooperation priming versus competition priming in Study 2.

***Presence of others:*** We used two different experimental setups to measure the influence of interviewer presence on inequality preferences: (a) the presence or absence of another person in the room (i.e. the laboratory) where the respondent filled out the questionnaire on a computer (simulating interviewer presence or absence) in Study 1, and (b) the presence or absence of an image of eyes on the computer screen while the participants filled out the online questionnaire in Study 2 (see Figure A1 in the Appendix). In Study 1, the control group filled out the questionnaire without another person in the room; in Study 2, the control group completed the online questionnaire on a computer screen that did not show images of eyes.

***Incentives:*** In Study 2, we tested the effect of incentives on inequality preferences. We randomly selected two groups of study subjects: the first group was asked to participate in our online survey without any incentive, whereas the second group was offered a payment of €5, to be paid *after* they had completed the questionnaire (conditional incentive), for participating in the study. Table 1 presents an overview of the research design for both studies.

***Inequality preference:*** One-item measures of attitudes toward earnings inequality are problematic because they produce virtually no variation in the responses regarding inequality preferences. Social desirability among other aspects may be a reason. Therefore, use of a factorial survey design (see Jasso, 2006; Rossi & Anderson, 1982; Wallander, 2009) appeared to be advisable because it allows the indirect measurement of specific preferences regarding earnings inequality by asking respondents to evaluate the justice of earnings on the basis of several descriptions of fictitious employees (vignettes). These multiple evaluations of just earnings

*Table 1* Research design of Study 1 and Study 2, according to the four study hypotheses

| Study | Priming (H1, H2) | Presence of others (H3) | Incentives (H4) |
|---|---|---|---|
| 1 | Cooperation vs. neutral | Experimenter present: yes/no | No treatment: Every participant received €10 as a conditional "show-up fee" (paid after completing the questionnaire) |
| 2 | Cooperation vs. competition | Eyes on screen: yes/no | Participants received either no incentive or €5 (paid after completing the questionnaire) |

can be used to reconstruct the individual inequality preference measured by an individual Gini coefficient.[3]

For our study, we used vignettes that described full-time employees (working 40 hours per week) who differ in ascribed and labor market–related characteristics and who earn a specific monthly gross income (see Table A1 in the Appendix).[4] The selection of these dimensions was based on theoretical considerations grounded in previous studies (Alves, 1982; Jasso & Rossi, 1977; Jasso & Webster, 1997; Sauer, Auspurg, Hinz, & Liebig, 2011; Sauer et al., 2009; Struck et al., 2006). The vignettes were presented on a computer screen using the same layout in all experimental settings (see Figure 1A in the Appendix). The following is an example of the wording used:

> A 55-year-old woman with no vocational training has three children and works as a clerk.
>
> She works in a company with a stable economic situation. Her performance is above-average.
>
> Her monthly gross earnings (before tax and other deductions) are €1,500.

A sample of 20 vignettes was drawn randomly from the vignette universe and presented to the subjects, meaning that each participant was asked to rate exactly

3 The individual Gini coefficient measures the inequality preference of respondents. The coefficient reflects the inequality among values of a frequency distribution. Perfect equality is expressed by Gini=0, maximal inequality is expressed by Gini=1.

4 For an overview for the use of vignette studies in justice research, see Liebig et al. 2015.

the same vignettes.[5] The purpose of this evaluation task was to decide whether or not the specific amount of gross earnings was just for the person described in the vignette and, if not, what a just amount of gross earnings would be in local currency (euros).[6] The just earnings provided by the participants were then used to calculate the Gini coefficient. This inequality measure is the outcome variable for all the following analyses.

# 4 Description and Results of Study 1

Study 1 was designed to test the effects of two experimental conditions on individual inequality preferences: (1) the effect of a cooperation frame (induced by a priming instrument) ($H_1$) and (2) the presence of others (presence or absence of the experimenter) ($H_3$). The experiment was conducted at a German university during the winter term of 2011/12. All of the participants were undergraduate students who had responded to handouts containing basic information about the study (time, place, duration, and compensation) that were distributed in the main building of the university.

The sample consisted of 145 participants. These were randomly assigned to the different experimental treatments, as shown in Table 2: 65 participants (45%) completed the survey in the presence of another person, while 80 participants (55%) filled out the questionnaire with no other person present in the room; 79 (54%) of all participants were primed on cooperation, while 66 (46%) were given neutral primes.

5 In other research designs, it is useful to draw several decks to collect ratings of as many vignettes as possible. Furthermore, sophisticated sampling techniques are recommended in order to arrive at efficient estimations of the coefficients (Atzmüller & Steiner, 2010; Dülmer, 2007). However, this was not necessary in the setup described because we were investigating differences using different experimental settings.

6 To avoid response heuristics that would make it easier to state a preference for the earnings described in the vignette by simply checking a box instead of typing in a value, participants had to insert a specific amount of money even if they thought the earnings stated were just. In the latter case, subjects had to type into the blank field the same value as was given in the vignette description. In the factorial survey literature, this is known as the "direct approach," that is, an approach in which respondents use an open scale to insert a value, as opposed to the "indirect approach," in which respondents do not provide a specific monetary value but instead use a justice scale to evaluate whether the value is just (sometimes by means of a rating scale but also with the use of open scales) and, subsequently, the researcher estimates the just earnings by means of individual regression techniques (for details, see Jasso, 2006; Jasso & Wegener, 1997). As discussed in the literature, the direct answers may lead to anchor effects – i.e., respondents adjusting their ratings to the income provided in the vignette – but this effect was not a problem in our study because only the differences between experimental groups were analyzed.

*Table 2* Number of respondents per experimental condition in absolute values

| | Cooperation-primed group | Control group | N |
|---|---|---|---|
| No other person present | 44 | 36 | 80 |
| Another person present | 35 | 30 | 65 |
| N | 79 | 66 | 145 |

*Data*: "Experiment on the influence of cooperative relationships on justice evaluations." (doi: 10.4119/unibi/sfb882.2012.1).

The experiments were conducted in two laboratories equipped with a computer screen on a table, along with a chair for the participant, as well as a table and chair for the experimenter when called for. Participants were asked to take the scrambled-sentence test and later to fill out a questionnaire regarding issues of social justice (including the vignette study on inequality preferences), personal background, and other questions used to control for side effects (e.g., using a social-desirability scale).[7] On average, the questionnaire took 35 minutes to complete.

## 4.1 Measurement

*Inequality preferences* were measured by the factorial survey design described in Section 3.[8]

*Priming:* Participants were given 20 sets of four or five words in a scrambled order and were asked to construct grammatically correct sentences. The cooperative mindset should be activated by words closely related to "cooperation" such as "together," "help," "cooperation," "fair," "trust," and "sharing." In total, 10 out of 20 sentences (50%) included primed wordings. The priming instrument was developed on the basis of previous studies on cooperation priming (Bargh, Gollwitzer, Lee-Chai, Barndollar, & Trötschel, 2001; Bry, Meyer, Oberlé, & Gherson, 2009; Drouvelis, Metcalfe, & Powdthavee, 2010; Kay & Ross, 2003). The control group received neutral sentences.[9]

7 The questionnaire was programed using the web survey software Unipark.

8 For descriptive statistics of the variables, see Table A2 in the Appendix.

9 Four participants reported that they were aware of being primed. To avoid contrast effects, we excluded these participants from the analyses. Another respondent was excluded because that participant failed to provide information on age and gender. The final analysis was thus based on 140 participants.

*Presence of others:* All participants were welcomed by an experimenter[10] who introduced them to the setup of the study. If the participant was randomly assigned to the experimental group, the experimenter stayed in the laboratory while the participant filled out the questionnaire at the computer. The experimenter was asked to remain quiet, without watching the computer screen during the task. The experimenter paid the show-up fee to the participants after they completed the questionnaire. Participants in the control group were shown to the room by a secretary and filled out the questionnaire with no other person present in the room. These participants were paid the show-up fee in cash by the secretary after they completed the questionnaire.

*Incentives:* All participants received a show-up fee of €10 in cash after completing the questionnaire. Because we did not vary incentivizing in this study, we are not able to test the effect of incentives.

*Social desirability:* To ensure that our results would not be biased by social desirability, we used three items from the impression management scale, which is based on the work of Paulhus (1984, 1991) and which was empirically tested by Winkler, Kroh and Spiess (2006).

## 4.2 Results

A comparison of the means of the individual Gini coefficients of participants with and without the presence of an experimenter revealed no significant difference in inequality preferences ($Gini_{w/o} = 0.29$; $Gini_{w/} = 0.31$; $t = -1.16$; $p(T>t) = 0.88$). The same applied to the priming condition: participants who received cooperation priming did not deviate significantly in their reports on inequality preferences from those in the neutral priming condition ($Gini_{coop.} = 0.29$; $Gini_{neutral} = 0.30$; $t = 0.37$; $p(T>t) = 0.36$). It would appear that neither presence of others nor cognitive priming on cooperation is a sufficient condition for activating a normative goal frame that would influence inequality preferences.

To filter the true effects of the two experimental conditions from effects induced by characteristics of the respondents, we performed an ordinary least squares (OLS) regression, controlling for social desirability by including the impression management scale and the respondents' gender and age. In a second model, we extended this basic model by adding interactional effects between the two experimental conditions to determine whether priming would have an effect on inequality preference depending on the absence or presence of an experimenter.

10 We recruited graduate students from a class on social stratification (master's degree level) to work as experimenters (N = 14; 50% female).

*Table 3* Preference for earnings inequality (Gini) regressed on two experimental conditions (cooperative priming and experimenter presence)

| | Preference for earnings inequality (Gini) | | | |
|---|---|---|---|---|
| | Model 1 | | Model 2 | |
| *Treatments variables* | | | | |
| Priming (cooperative = 1) | -0.001 | (0.013) | 0.018 | (0.017) |
| Experimenter (present = 1) | 0.009 | (0.012) | 0.032* | (0.018) |
| Priming * experimenter | | | -0.043* | (0.024) |
| *Control variables* | | | | |
| Social desirability | -0.004 | (0.008) | -0.003 | (0.008) |
| Gender (female = 1) | -0.022+ | (0.013) | -0.023* | (0.013) |
| Age (in years) | -0.005*** | (0.001) | -0.005*** | (0.001) |
| Constant | 0.470*** | (0.034) | 0.464*** | (0.034) |
| N | 140 | | 140 | |
| $R^2$ | 0.154 | | 0.172 | |

*Notes*: OLS regression coefficients, robust standard errors in parentheses, one-sided t-tests, +p<.10, * p<.05, ** p<.01, *** p<.001, doi:10.4119/unibi/sfb882.2012.1.

Table 3 shows the findings of both OLS regression models. The first model reports the main effects of the two experimental conditions while controlling for personal characteristics (social desirability, gender, and age); the second includes the interactional effects. Even when controlling for personal characteristics, we found that our two experimental conditions had no main effects on the preference for earnings inequality (Model 1). Testing for interaction effects (Model 2) revealed that the cooperation priming condition was effective only if the experimenter was present in the room. Participants primed on cooperation preferred less inequality only in the presence of another person. If no one else was in the room, respondents in the two priming conditions showed the same response pattern regarding earnings inequality. Contrary to our prediction, presence of another person in the neutral priming condition was related to a preference for greater inequality.

Of the control variables, only age and gender showed significant effects: older students and female students preferred a lower earnings inequality than did the younger students and male students. The scale for detecting a response bias toward social desirability had no effect.

## 4.3 Summary

As all respondents were paid for completing the questionnaire we induced an economic exchange situation for all participants. Following our theory this results in an activation of the gain frame. Accordingly we expected a higher level of inequality preference for all respondents. Evidence is provided by comparing the constant of the regression models in Table 3 and Table 5 (Table 5 reports results from an experiment were payment is experimentally varied). While under the general payment condition of Study 1 the mean Gini the respondents considered as fair is .470 (Model 1) resp. .464 (Model 2), the general level of preferred inequality in gross income in Study 2 is substantial lower (Model 1: .315, Model 2: .334). Seemingly, in Study 1 all respondents regardless the experimental treatment start with a higher level of income inequality as "economic exchange" is the default definition of the interview situation. The results for the experimental treatments indicate that only the combination of presence of the experimenter in the room and cooperation priming (activating a cooperative mindset) reduced individual inequality preferences significantly. Therefore, $H_1$ and $H_3$ are only partly confirmed: only when a cooperative relationship was established by the presence of an experimenter *and* an induced cooperative mindset, respondents showed a stronger equality orientation. What is quite unclear, however, is why priming on cooperation had no significant main effect. In accordance with GFT, single activation of a normative frame should be enough to alter the behavior of individuals. There are at least three possible explanations for this finding: (1) the priming was too weak and the difference between the two priming conditions was not distinct enough; (2) the priming instrument was filled out more seriously by the respondents when an experimenter was present and thus worked well only under this condition; and/or (3) participants knew that they would receive money for completing the interview, the gain frame was activated and the presence of another person was perceived as a "control mechanism" for checking how respondents behaved and to make sure they completed the questionnaire properly. In the last-mentioned case, the structural situational cues (incentive, person present) were stronger than the "psychological" cues of priming. The results on the condition "neutral priming, experimenter present" show that the presence of the experimenter may be interpreted as a control mechanism enforcing the structural induced framing. To better understand how incentivizing affects the response behavior on inequality preferences, we conducted a second study.

# 5 Description and Results of Study 2

Study 2 was designed as a follow-up study to test the effects of three experimental conditions on individual inequality preferences: (1) the differing effects of priming (cooperation vs. competition) ($H_1$, $H_2$), (2) the presence of others as simulated by eyes on the computer screen (eyes vs. no eyes) ($H_3$), and (3) the influence of incentives (incentivized vs. voluntary participation) ($H_4$). The study was conducted during the summer term of 2012 at the same German university. Respondents were recruited from an undergraduate course. Students received an email inviting them to participate in an online survey on inequality of earnings. Out of the 724 students invited to participate, 210 completed the questionnaire (response rates: 41% with incentive and 20% without incentive) (see Table 4). All participants were randomly selected; first, they had a 27.6 percent chance of being selected to the incentive sample; second, all participants had a 50 percent chance of being selected for the experimental conditions (priming and presence of eyes on the computer screen).

## 5.1 Measurement

*Inequality preferences* were measured using the factorial survey design described in Section 3.[11]

*Priming for cooperation versus competition:* We induced two mindsets: cooperation and competition. To improve the scrambled-sentence test we used in Study 1, we reduced the number of word sets to 12 and increased the number of primed words sets to 9. For the most part, the "cooperation" primes resembled the words chosen for the first priming condition in Study 1. The "competition" primes were developed based on examples from the recent literature (Bargh et al., 2001; Bry et al., 2009; Kay & Ross, 2003). Words such as "competition," "comparison," "arguing," "power," "assertion," "provocation," "winning," and "inconsiderate behavior" were used to induce a competitive mindset. To ensure comparability of the two experimental conditions, the sentences differed only in the specific priming but not in their structure.[12]

*Presence of others:* We chose to simulate the presence of others by displaying eyes on the computer screen while the respondent answered the questions. We used natural-looking eyes in the top right-hand corner of the screen (see Figure A1 in the Appendix). The eyes were not meant to be too prominent because we wanted to induce a feeling of someone being present at a subconscious level. Questions

11 For descriptive statistics of the variables, see Table A2 in the Appendix.

12 To avoid any contrast effects, we excluded 3 of the 210 participants from our analyses because they showed signs that they were aware of being primed, and 11 participants because they failed to provide complete information on just earnings in the vignettes, gender, or age. The analysis is thus based on 191 respondents.

*Table 4* Number of respondents per experimental condition in absolute values

| | No incentive | | Incentive | | |
|---|---|---|---|---|---|
| | No eyes | Eyes | No eyes | Eyes | N |
| Priming on cooperation | 32 | 31 | 23 | 19 | 105 |
| Priming on competition | 35 | 31 | 19 | 20 | 105 |
| N | 67 | 62 | 42 | 39 | 210 |

*Source*: "Experiment on the influence of interviewer presence and incentivizing on justice evaluations." (doi: 10.4119/unibi/sfb882.2012.2).

concerning awareness and interpretation of the eyes were asked at the end of the questionnaire. In the control group, participants answered the same questionnaire, but the eyes were not present on the screen.

*Social desirability:* Again, we used the three items on social desirability, that is, impression management (see Winkler et al., 2006).

## 5.2 Results

When we compared the means of the individual Gini coefficients for participants with and for those without the presence of eyes on the computer screen, we found no difference in inequality preferences ($Gini_{w/o} = 0.28$; $Gini_{w/} = 0.28$; $t = -0.01$; $p(T<t) = 0.49$). The priming condition also showed no significant differences in inequality preferences; although there is a tendency that participants who received cooperation priming preferred less inequality than did participants who received competition priming ($Gini_{coop.} = 0.27$; $Gini_{comp} = 0.29$; $t = 1.55$; $p(T>t) = 0.06$). The difference between the incentive and non-incentive condition was significant ($Gini_{payment} = 0.29$; $Gini_{nonpay.} = 0.27$; $t = -2.01$; $p(T<t) = 0.02$). To test whether these results were sensitive to the contextual setting and whether they were biased due to social desirability, we performed the following analysis using OLS regression.

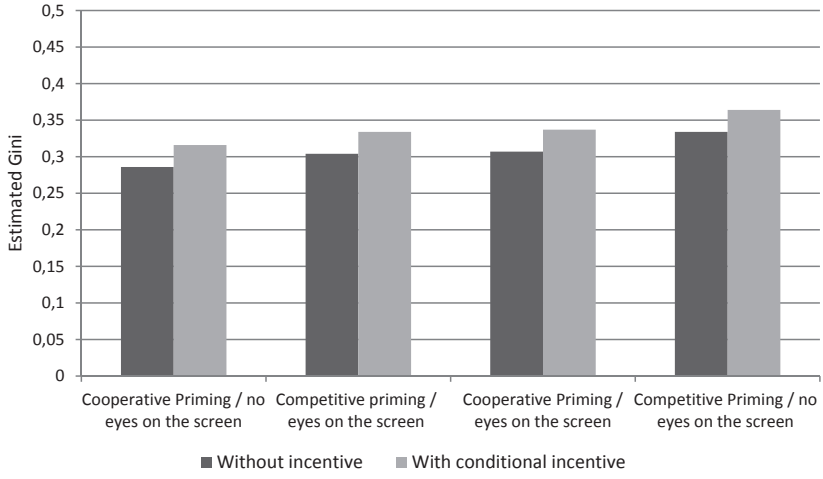
Again, we estimated two models, one testing only for the main effects of the three experimental conditions (Model 1) and the other including an interaction term for the priming condition and the "eyes on the screen" condition (Model 2) (Table 5). In both models we controlled for gender, age, and response bias (social desirability). We also added two variables to control for the interview situation: (1) if respondents were aware of the eyes on the screen and (2) if they completed the questionnaire while someone was in the room. The second question was asked to control for the presence of other people while the respondents filled out the online questionnaire.

*Table 5* Preference for earnings inequality (Gini coefficient) regressed on three experimental conditions – competition, control (eyes), and payment)

| | Preference for earnings inequality (Gini) | | | |
|---|---|---|---|---|
| | Model 1 | | Model 2 | |
| *Treatments* | | | | |
| Incentive (yes = 1) | 0.030** | (0.012) | 0.030** | (0.012) |
| Priming (Cooperation = 1, competition = 0) | -0.021* | (0.012) | -0.048** | (0.016) |
| Eyes on screen (eyes = 1, no eyes = 0) | -0.005 | (0.012) | -0.030* | (0.017) |
| Cooperation priming with eyes on the screen | | | 0.051** | (0.024) |
| *Control variables* | | | | |
| Social desirability | -0.005 | (0.007) | -0.007 | (0.008) |
| Awareness of eyes on screen (yes = 1) | -0.025 | (0.026) | -0.027 | (0.026) |
| Any person present during completion of questionnaire (yes = 1) | 0.009 | (0.012) | 0.009 | (0.012) |
| Gender (female = 1) | 0.003 | (0.013) | 0.004 | (0.013) |
| Age (in years) | -0.001* | (0.001) | -0.001* | (0.001) |
| Constant | 0.315*** | (0.037) | 0.334*** | (0.026) |
| N | 191 | | 191 | |
| $R^2$ | 0.021 | | 0.088 | |

*Notes*: N = 191, OLS regression coefficients; robust standard errors in parentheses, one-sided t-tests. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. doi:10.4119/unibi/sfb882.2012.1.

Two of our experimental conditions influenced the preferred earnings inequality. Respondents who were paid after they completed the questionnaire preferred greater earnings inequality than did those who participated in the study without payment. Priming also had a significant effect: those with cooperation priming (normative goal frame) preferred lower earnings inequality as compared with those with competition priming (gain frame). This is in line with our Hypotheses 3 and 4. Contrary to our prediction, there was no significant main effect of "eyes on the screen," although the regression coefficient showed the expected direction: those with eyes on their computer screen had slightly smaller Gini coefficients, which indicates that they preferred lower earnings inequality. Of our control variables, only the age of a respondent had a significant effect: older students – mostly from the social sciences – preferred less inequality. The results showed no bias for social




*Note*: Based on Model 2 (Table 5) only considering the constant and the coefficients of incentive, priming, eyes on the screen and the interaction of the two latter main effects.

*Figure 2* Estimated levels of preferred inequality (Gini) for the combination of two experimental treatments

desirability whether respondents became aware of the eyes on the screen or they were not alone while completing the questionnaire.

In Model 2, we tested our previous finding that in the presence of an experimenter or an interviewer, respondents who underwent cooperation priming endorsed less earnings inequality. Although the main effects of the two conditions "incentive" and "priming" remained more or less the same, the main effect of eyes on the screen was now significant. But, contrary to our predictions and to the findings from Study 1, the interaction effect with priming was positive, meaning that those with a cooperative mindset and eyes on the screen preferred greater earnings inequality when compared with those in the experimental condition "cooperation priming with no eyes on the screen" ($p_F = .106$, one sided F-test). These results showed no biases for the control variables social desirability, awareness of eyes, and people present while the questionnaire was being filled out. As for Study 1 we provide the estimated Gini level for the eight experimental combinations in Figure 2. Again, it can be seen that cooperative primed respondents without incentive (and without eyes on the screen) showed the lowest level of estimated Gini (0.286), while competitive primed respondents with incentive the highest level (0.364).

## 5.3 Summary

Study 2 provided evidence that respondents who received a payment for completing the questionnaire showed a significantly greater preference for inequality than did those who completed the questionnaire voluntarily. Together with the findings from Study 1 – where all participants received an incentive and we observed a general higher level of inequality preference (constant in Study 1 regression: .470/.464 vs. constant in Study 2 regression: .315/.334) – these results support Hypothesis 4. The finding holds independently of any priming effect. Our explanation is that by using conditional incentives, an economic-exchange relation is established and the respondents will act according to their gain frame. However, the results on type of relationship (relational mindsets) and presence of an interviewer only partly confirmed our other three hypotheses ($H_1$, $H_2$, and $H_3$): only in the no-eyes condition did respondents who had a cooperative mindset prefer more equality and who received competition priming more inequality. The presence of eyes on the screen revealed no clear effects: respondents with a competitive mindset showed slightly lower inequality preferences, whereas those with a cooperative mindset preferred slightly greater inequality (differences are on the third decimal place). These findings leave room for speculation as to whether the condition "eyes on the screen" can truly function as a substitute for the presence of another person, as we had assumed at the outset. Eyes on the screen may not necessarily induce cooperation but may be perceived as situational cues for social control and anticipated sanctions for norm-violating behavior as we already assumed in Study 1. Because recent studies have reported heterogeneous and contradictory results when investigating whether the presence of eyes on the computer screen would affect altruistic or egoistic behavior (see Vogt et al., 2014), we suspect that this instrument might not be valid as a substitute for the presence of an interviewer.

# 6 Discussion

Large-scale population surveys are predominantly based on interviewer-assisted data collection, and incentivizing is becoming a more common practice in survey research. In this study, we investigated whether these structural conditions of interview situations influence individual inequality preferences. In keeping with GFT and empirical justice research, the main argument was that in both structural conditions – presence of an interviewer and use of incentives – different types of social relationships are established and either a normative frame or a gain frame is activated in the respondent. The consequence of different framing is that respondents will apply different distributive principles when evaluating earnings inequality according to the "logic of justice." In normative framing, the equality principle

is dominant, whereas in gain framing, the equity principle is the appropriate distributive principle. Therefore, in a survey setting, having another person present in the interview room in addition to establishing a cooperative relationship by means of priming was expected to influence respondents' preferences in favor of a more equal distribution of earnings. In contrast, when respondents are paid for participating and a competitive relationship is established through priming, their preference will be for less equal distributions.

The results of two experiments we conducted involving students at a German university showed that establishing a competitive relationship by incentivizing respondents led to a response bias toward greater earnings inequality. The results concerning the presence of an interviewer were not as straightforward: the combination of interviewer presence and the inducing of a cooperative relationship led to an equality bias. Hence, the mere presence of an interviewer was not sufficient to trigger a normative framing of the interview situation. The use of styled eyes on the computer screen while respondents filled out the online questionnaire was intended to serve as a substitute for interviewer presence, but the results of this test were heterogeneous and in fact contradictory. Instead, the respondents appeared to feel watched or controlled by the eyes on the screen, which undermined the perception of a cooperative relationship and failed to activate a normative goal frame.

In general, our results raise the concern that measures of inequality preferences are affected by situational conditions and cues. The effects of paying respondents for completing a questionnaire were the most robust, incentives induced preferences for higher inequality. But, since we used student samples and interviewer settings that are not strictly comparable with those used in large population surveys, our results cannot be generalized and simply transferred to the "survey reality." Our assumptions must therefore be tested under more appropriate conditions. Nevertheless, we were able to show that certain effects of the study design must be controlled for because they can influence the substantive findings of surveys on inequality preferences.

Overall, our results suggest that population surveys on attitudes toward social inequality or social justice should not rely on one mode of data collection alone. Each mode is characterized by different structural conditions and produces certain situational cues that affect respondents' behavior systematically. If our experimental data on the effects of the presence of experimenter in combination with a cooperative mindset reflect a general phenomenon of framing and adapting of inequality preferences to situational cues, survey data using interviewer-assisted modes may overestimate the equality orientation within a population. This might be the case when interviewers establish a cooperative relationship with their respondents within the interview situation and by doing so strengthen a normative, equality oriented framing. Interviewers from a recent German employee-survey on the perceptions and evaluations of social inequality (LINOS1, DOI: 10.4119/unibi/sfb882.2014.9,

CAPI-split, N = 1007) report for 59 percent of the completed interviews that the interview situation was characterized by a cooperative and trustful mindset of the respondent. If respondents of this survey show the same response pattern under the conditions of a collaborative mindset and present of interviewer, we expect a substantial equality oriented response bias.

As effects of conditional incentivizing on inequality preferences were observed in both studies, our results contribute to the ongoing discussion about whether large-scale population surveys should use incentives to increase respondents' willingness to participate. Aside from the question of whether payment really contributes to higher-quality data by increasing the response rates in population surveys, our study showed that paying for participation is relevant not only for methodological purposes but also for substantive issues. If other respondents behave towards incentives the same way as our student sample did, we may observe a very different picture of the inequality preferences within a society. Therefore, the question to be addressed in future research is which attitudes are relevant to political or other types of behavior – those resulting from a normative frame or those resulting from a gain frame.

# 7 References

Alves, W. M. (1982). Modeling Distributive Justice Judgments. In P. H. Rossi & S. L. Nock (Eds.), *Measuring Social Judgments: The Factorial Survey Approach* (pp. 205-234). Beverly Hills: Sage.

Atzmüller, C., & Steiner, P. M. (2010). Experimental Vignette Studies in Survey Research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 6*(3), 128-138. doi:10.1027/1614-2241/a000014

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Trötschel, R. (2001). The Automated Will: Nonconscious Activation and Pursuit of Behavioral Goals. *Journal of Personality and Social Psychology, 81*(6), 1014-1027. doi:10.1037/0022-3514.81.6.1014

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of Being Watched Enhance Cooperation in a Real-World Setting. *Biology Letters, 2*(3), 412-414. doi:10.1098/rsbl.2006.0509

Becker, R., & Mehlkop, G. (2011). Effects of Prepaid Monetary Incentives on Mail Survey Response Rates and on Self-reporting about Delinquency – Empirical findings. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 111*(1), 5-25. doi:10.1177/0759106311408870

Bolton, G. E., & Zwick, R. (1995). Anonymity versus Punishment in Ultimatum Bargaining. *Games and Economic Behavior, 10*(1), 95-121. doi:10.1006/game.1995.1026

Bry, C., Meyer, T., Oberlé, D., & Gherson, T. (2009). Effect of Priming Cooperation or Individualism on a Collective and Interdependent Task: Changeover Speed in the 4 × 100-meter Relay Race. *Journal of Sport and Exercise Psychology, 31*(1), 380-389.

Burnham, T., & Hare, B. (2007). Engineering Human Cooperation. *Human Nature, 18*(2), 88-108. doi:10.1007/s12110-007-9012-2

Charness, G., & Gneezy, U. (2008). What's in a Name? Anonymity and Social Distance in Dictator and Ultimatum Games. *Journal of Economic Behavior & Organization, 68*(1), 29-35. doi:10.1016/j.jebo.2008.03.001

Deutsch, M. (1975). Equity, Equality, and Need: What Determines Which Value Will Be Used as the Basis of Distributive Justice? *Journal of Social Issues, 31*(3), 137-149. doi:10.1111/j.1540-4560.1975.tb01000.x

Drouvelis, M., Metcalfe, R., & Powdthavee, N. (2010). *Priming Cooperation in Social Dilemma Games*. IZA Discussion Paper No. 4963. Bonn: Institute for the Study of Labor.

Dülmer, H. (2007). Experimental Plans in Factorial Surveys: Random or Quota Design? *Sociological Methods Research, 35*(3), 382-409. doi:10.1177/0049124106292367

Ellingsen, T., & Johannesson, M. (2004). Promises, Threats and Fairness. *The Economic Journal, 114*(495), 397-420. doi:10.1111/j.1468-0297.2004.00214.x

Fehr, E., & Schmidt, K. M. (2006). The Economics of Fairness, Reciprocity and Altruism – Experimental Evidence and New Theories. In S.-C. Kolm & J. M. Ythier (Eds.), *Handbook of the Economics of Giving, Altruism and Reciprocity, Vol. 1: Foundations* (pp. 615-691). Amsterdam: North-Holland.

Fiske, A. P. (1991). *Structures of Social Life: The Four Elementary Forms of Human Relations*. New York: Free Press.

Gächter, S., & Fehr, E. (1999). Collective Action as a Social Exchange. *Journal of Economic Behavior & Organization, 39*(4), 341-369. doi:10.1016/s0167-2681(99)00045-1

Groves, R. M. (1989). *Survey Errors and Survey Costs*. New York: Wiley.

Haley, K. J., & Fessler, D. M. T. (2005). Nobody's Watching? Subtle Cues Affect Generosity in an Anonymous Economic Game. *Evolution & Human Behavior, 26*(3), 245-256. doi:10.1016/j.evolhumbehav.2005.01.002

Hertel, G., & Fiedler, K. (1994). Affective and Cognitive Influences in Social Dilemma Game. *European Journal of Social Psychology, 24*(1), doi:131-145. 10.1002/ejsp.2420240110

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences, 24*(03), 383-403.

Hoffman, E., McCabe, K., Shachat, K., & Smith, V. L. (1994). Preferences, Property Rights, and Anonymity in Bargaining Games. *Games and Economic Behavior, 7*(3), 346-380. doi:10.1006/game.1994.1056

Hoffman, E., McCabe, K., & Smith, V. L. (1996). Social Distance and Other-Regarding Behavior in Dictator Games. *The American Economic Review, 86*(3), 653-660. doi:10.1257/aer.89.1.340

Hole, A. D. (2011). Communication and Fair Distribution. *Rationality and Society, 23*(2), 234-264. doi:10.1177/1043463110387266

James, J. M., & Bolstein, R. (1990). The Effect of Monetary Incentives and Follow-Up Mailings on the Response Rate and Response Quality in Mail Surveys. *Public Opinion Quarterly, 54*(3), 346-361. doi:10.1086/269211

Jasso, G. (2006). Factorial Survey Methods for Studying Beliefs and Judgments. *Sociological Methods & Research, 34*(3), 334-423. doi:10.1177/0049124105283121

Jasso, G., & Rossi, P. H. (1977). Distributive Justice and Earned Income. *American Sociological Review, 42*(4), 639-651. doi:10.2307/2094561

Jasso, G., & Webster, Jr., M. (1997). Double Standards in Just Earnings for Male and Female Workers. *Social Psychology Quarterly, 60*(1), 66-78. doi:10.2307/2787012

Jasso, G., & Wegener, B. (1997). Methods for Empirical Justice Analysis: Part 1: Framework, Models, and Quantities. *Social Justice Research, 10*(4), 393-430. doi:10.1007/bf02683292

Kahneman, D., & Tversky, A. (1984). Choices, Values, and Frames. *American Psychologist, 39*(4), 341-350. doi:10.1037/0003-066x.39.4.341

Kahneman, D., & Tversky, A. (Eds.). (2000). *Choices, Values and Frames*. New York: Russell Sage Foundation; Cambridge: Cambridge University Press.

Kay, A. C., & Ross, L. (2003). The Perceptual Push: The Interplay of Implicit Cues and Explicit Situational Construals on Behavioral Intentions in the Prisoner's Dilemma. *Journal of Experimental Social Psychology, 39*(6), 634-643. doi:10.1016/S0022-1031(03)00057-X

Konow, J. (2001). Fair and Square: The Four Sides of Distributive Justice. *Journal of Economic Behavior & Organization, 46*(2), 137-164. doi:10.1016/S0167-2681(01)00194-9

Krebs, D. L. (2008). The Evolution of a Sense of Justice. In J. Duntley & T. K. Shackelford (Eds.), *Evolutionary Forensic Psychology: Darwinian Foundations of Crime and Law* (pp. 230-248). Oxford: Oxford University Press.

Liberman, V., Samuels, S. M., & Ross, L. (2004). The Name of the Game: Predictive Power of Reputations versus Situational Labels in Determining Prisoner's Dilemma Game Moves. *Personality and Social Psychology Bulletin, 30*(9), 1175-1185. doi:10.1177/0146167204264004

Liebig, S. (2001). Lessons from Philosophy? Interdisciplinary Justice Research and Two Classes of Justice Judgments. *Social Justice Research, 14*(3), 265-287. doi:10.1023/a:1014367907348

Liebig, S., & Sauer, C. (2013). Soziologische Gerechtigkeitsanalyse. Überlegungen zur theoretischen Fundierung eines Forschungsfeldes. *Analyse & Kritik. Zeitschrift für Sozialtheorie*, *35*(2), 371–394

Liebig, S., & Sauer, C. (2015). Sociology of Justice. In C. Sabbagh & M. Schmitt (Eds.), *Handbook of Social Justice Theory and Research* (in press). New York: Springer.

Liebig, S., Sauer, C. & Friedhoff, S. (2015). Empirische Gerechtigkeitsforschung mit dem faktoriellen Survey. *Soziale Welt*, Sonderband 22: 315–333.

Lindenberg, S. (1990). Homo Socio-Oeconomicus: The Emergence of a General Model of Man in the Social Sciences. *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft, 146*(4), 727-748.

Lindenberg, S. (2001). Social Rationality versus Rational Egoism. In J. H. Turner (Ed.), *Handbook of Sociological Theory* (pp. 635-668). New York: Kluwer Academic/Plenum.

Lindenberg, S. (2006). Prosocial Behavior, Solidarity, and Framing Processes. In D. Fetchenhauer, A. Flache, A. P. Buunk, & S. Lindenberg (Eds.), *Solidarity and Prosocial Behavior: An Integration of Sociological and Psychological Perspectives* (pp. 23-44). New York: Springer.

Lindhjem, H., & Navrud, S. (2011). Using Internet in Stated Preference Surveys: A Review and Comparison of Survey Modes. *International Review of Environmental and Resource Economics, 5*(4), 309-351. doi:10.1561/101.00000045

Miller, D. (1999). *Principles of Social Justice*. Cambridge: Harvard University Press.

Ormel, J., Lindenberg, S., Steverink, N., & Verbrugge, L. M. (1999). Subjective Well-Being and Social Production Functions. *Social Indicators Research, 46*(1), 61-90. doi:10.1023/A:1006907811502

Paulhus, D. L. (1984). Two-Component Models of Socially Desirable Responding. *Journal of Personality and Social Psychology, 46*(3), 598-609. doi:10.1037/0022-3514.46.3.598

Paulhus, D. L. (1991). Measurement and Control of Response Bias. In J. P. Robinson, P. R. Shaver & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 17-59). San Diego: Academic Press.

Rigdon, M., Ishii, K., Watabe, M., & Kitayama, S. (2009). Minimal Social Cues in the Dictator Game. *Journal of Economic Psychology, 30*(3), 358-367. doi:10.1016/j.joep.2009.02.002

Rossi, P. H., & Anderson, A. B. (1982). The Factorial Survey Approach: An Introduction. In P. H. Rossi & S. L. Nock (Eds.), *Measuring Social Judgments: The Factorial Survey Approach* (pp. 15-67). Beverly Hills: Sage.

Sauer, C., Auspurg, K., Hinz, T., & Liebig, S. (2011). The Application of Factorial Surveys in General Population Samples: The Effects of Respondent Age and Education on Response Times and Response Consistency. *Survey Research Methods, 5*(3), 89-102.

Sauer, C., Liebig, S., Auspurg, K., Hinz, T., Donaubauer, A., & Schupp, J. (2009). *A Factorial Survey on the Justice of Earnings within the SOEP-Pretest 2008*. IZA Discussion Paper No. 4663. Bonn: Institute for the Study of Labor.

Schwinger, T. (1981). Steuerung und Rechtfertigung sozialer Prozesse durch Gerechtigkeitsnormen. In W. Grunwald & H. G. Lilge (Eds.), *Kooperation und Konkurrenz in Organisationen* (pp. 97-107). Bern: Haupt.

Singer, E., Van Hoewyk, J., & Maher, M. P. (2000). Experiments with Incentives in Telephone Surveys. *Public Opinion Quarterly, 64*(2), 171-188. doi:10.1086/317761

Srull, T. K., & Wyer, R. S. (1979). The Role of Category Accessibility in the Interpretation of Information about Persons: Some Determinants and Implications. *Journal of Personality and Social Psychology, 37*(10), 1660-1672. doi:10.1037/0022-3514.37.10.1660

Struck, O., Stephan, G., Köhler, C., Klause, A., Pfeifer, C., & Sohr, T. (Eds.). (2006). *Arbeit und Gerechtigkeit: Entlassungen und Lohnkürzungen im Urteil der Bevölkerung.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Toepoel, V. (2012). Effects of Incentives in Surveys. In L. Gideon (Ed.), *Handbook of Survey Methodology for the Social Sciences* (pp. 209-223). New York: Springer.

Tönnies, F. (1887). *Gemeinschaft und Gesellschaft*. Leipzig: Fues.

Törnblom, K. (1992). The Social Psychology of Distributive Justice. In K. R. Scherer (Ed.), *Justice: Interdisciplinary Perspectives* (pp. 177-236). Cambridge: Cambridge University Press.

Traub, S., Seidl, C., & Schmidt, U. (2009). An Experimental Study on Individual Choice, Social Welfare, and Social Preferences. *European Economic Review, 53*(4), 385-400. doi:10.1016/j.euroecorev.2008.04.002

Tyler, T. R., Boeckmann, R. J., Smith, H. J., & Huo, Y. J. (1997). *Social Justice in a Diverse Society.* Boulder: Westview.

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2010). A Method for Evaluating Mode Effects in Mixed-Mode Surveys. *Public Opinion Quarterly, 74*(5), 1027-1045. doi:10.1093/poq/nfq059

Vogt, S., Efferson, C., Berger, J., & Fehr, E. (2014). Eye Spots Do Not Increase Altruism in Children. *Evolution & Human Behavior.* doi:10.1016/j.evolhumbehav.2014.11.007

Wallander, L. (2009). 25 Years of Factorial Surveys in Sociology: A Review. *Social Science Research, 38*(3), 505-520. doi:10.1016/j.ssresearch.2009.03.004

Webster, M., & Sell, J. (Eds.). (2007). *Laboratory Experiments in the Social Sciences*. London: Academic Press.

Winkler, N., Kroh, M., & Spiess, M. (2006). *Entwicklung einer deutschen Kurzskala zur zweidimensionalen Messung von sozialer Erwünschtheit*. Discussion Paper No. 579. Berlin: German Institute for Economic Research.

Xiao, E., & Houser, D. (2009). Avoiding the Sharp Tongue: Anticipated Written Messages Promote Fair Economic Exchange. *Journal of Economic Psychology, 30*(3), 393-404. doi:10.1016/j.joep.2008.12.002

# Appendix

*Table A1* Vignette dimensions and their levels as used in both studies

| # | Dimensions | Levels |
|---|---|---|
| 1 | Age | 30/40/50/60 years |
| 2 | Sex | Male/female |
| 3 | Vocational degree | Without degree/vocational degree/university degree |
| 4 | Occupation | Unskilled worker/door(wo)man/engine driver/clerk/ hairdresser/social worker/software engineer/electrical engineer/manager/medical doctor |
| 5 | Gross earnings/month | 10 values, ranging from €500 to €15,000 |
| 6 | Children | 1 to 5 (1 = No child, 2 = one child, 3 = two children, 4 = three children, 5= four children) |
| 7 | Performance | Below-average/above-average |
| 8 | Economic situation of the firm | High profits/threatened by bankruptcy/solid |

*Table A2* Descriptive statistics for variables used in the models (studies 1 and 2)

| | Mean | SD | Min. | Max. |
|---|---|---|---|---|
| ***Study 1 (N = 140)*** | | | | |
| Preferred earnings inequality (Gini) | 0.298 | 0.079 | 0.065 | 0.619 |
| Priming (cooperation = 1, no cooperation = 0) | 0.529 | — | 0 | 1 |
| Person present (yes = 1) | 0.450 | — | 0 | 1 |
| Social desirability | 0.000 | 0.803 | -2.288 | 1.509 |
| Gender (female = 1) | 0.671 | — | 0 | 1 |
| Age (in years) | 24.379 | 5.136 | 18 | 59 |
| ***Study 2 (N = 191)*** | | | | |
| Preferred earnings inequality (Gini) | 0.278 | 0.086 | 0.014 | 0.595 |
| Incentive (yes = 1) | 0.403 | — | 0 | 1 |
| Priming (cooperation = 1, no cooperation = 0) | 0.476 | — | 0 | 1 |
| Eyes on screen (yes = 1) | 0.518 | — | 0 | 1 |
| Social desirability | 0.000 | 0.838 | -2.246 | 1.547 |
| Awareness of eyes on screen (yes = 1) | 0.094 | — | 0 | 1 |
| Person present (yes = 1) | 0.267 | — | 0 | 1 |
| Gender (female = 1) | 0.408 | — | 0 | 1 |
| Age (in years) | 22.869 | 4.132 | 19 | 60 |

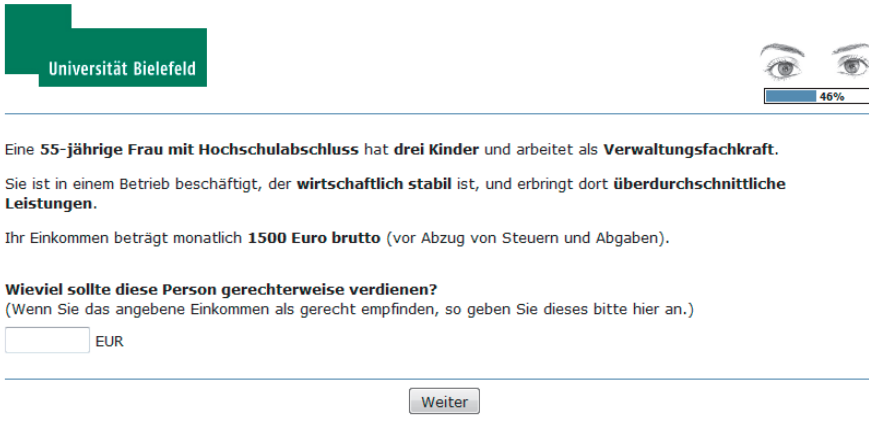



*Figure A1* Image of "eyes on the screen" condition in Study 2 (cf. p. 68)

methods, data, analyses | Vol. 9(1), 2015, pp. 87-110 DOI: 10.12758/mda.2015.003

# The Development and Test of a Measure of Youth's Ethnic and National Identity

*Lars Leszczensky & Aitana Gräbs Santiago*
*Mannheim Centre for European Social Research (MZES)*

**Abstract**

Comparatively few studies quantitatively examine the mechanisms underlying the formation of and change in young immigrants' ethnic and host country national identifications. A key reason for this research gap is the lack of an accurate measure of ethnic and national identity that meets the demands of integration research, i.e., includes a native reference group and is applicable to various age groups. In this article, we propose and test such a measure.

As ethnic identity and national identity both are types of social identity, our measure distinguishes three crucial dimensions of social identity. The cognitive dimension not only captures whether immigrants and their descendants actually conceive of themselves as belonging to the country of origin of their families but also captures the presence of potential dual identities. The evaluative dimension assesses how non-native and native youths evaluate their group memberships, respectively. Finally, the emotional dimension measures their respective strength of commitment towards their family's country of origin as well as towards the host country.

After presenting our measure of ethnic and of national identity, we test it quantitatively on native and non-native children and youths aged between 9 and 17 years. Our analyses confirm the suspected multi-dimensionality of both ethnic and national identity. We also ascertain the invariance of our measure across immigrants and natives as well as across different immigrant generations and age groups. The results further indicate strong reliability and construct validity. We therefore conclude that our proposed measure not only adequately captures different dimensions of ethnic and of national identity but that it is also applicable to different ethnic and age groups, thereby providing a valuable tool for studying immigrants' identification.

*Keywords*: Ethnic identity, national identity, social identity, identification, integration, migration, immigrants

# 1 Introduction

Ethnic diversity caused by immigration is nowadays a key feature of many European nations. In these societies, immigrants and their descendants face the challenge of combining ethnic and host country national identities (Phinney, Berry, Vedder, & Liebkind, 2006; Verkuyten & Martinovic, 2012).[1] This struggle is particularly pronounced among immigrant youths, identity development having long been recognized as a key task of adolescence (Meeus, 2011; Phinney, 1990). Yet, even though classical assimilation theories regard immigrants' identification with the host country as the last step in a successful integration process (e.g., Gordon, 1964; Nauck, 2001; Steinbach, 2004), comparatively few quantitative studies focus on explaining immigrants' emotional integration (Kalter 2008, p. 26).

Studying immigrants' identification, however, is important for two major reasons. First, weak national identification, or even dis-identification, is often considered to be a problem in and of itself since it threatens social cohesion and intensifies interethnic conflict (see Verkuyten & Martinovic, 2012). In most countries, even the descendants of immigrants show lower levels of identification with the host country than their native peers do (Phinney et al., 2006). Especially in Western Europe, an (alleged) lack of immigrants' identification with their host countries stands at the center of political as well as of scientific debates (e.g., Diehl & Schnell, 2006; Ersanilli & Saharso, 2011). Second, it is crucial to learn more about the determinants of ethnic and national identifications, because these identities are potentially consequential for other dimensions of integration, such as ethnic inequalities in the labor market and in the educational system (e.g., Altschul, Oyserman, & Bybee, 2006; Casey & Dustmann, 2010; Nekby & Rödin, 2010) or the formation of interethnic friendships (e.g., Leszczensky, 2013; Rutland et al., 2012).

A major reason for lack of research on immigrants' identification is the lack of appropriate data and, especially, the lack of an adequate measure of ethnic and national identities of young immigrants (see Leszczensky & Gräbs Santiago, 2014a; Nandi & Platt, 2012). In this article, we propose and test such a measure[2]. Given that ethnic identity is a notoriously vague term, we proceed by providing a conceptual understanding of ethnic and of national identity (1.1). Then we briefly discuss shortcomings of established measures of ethnic identity as well as the need for an

1 For the sake of brevity, we use the term "immigrant" to denote actual immigrants as well as their children and grandchildren.

2 Our measurement is ready for use and available via ZIS/GESIS (Leszczensky & Gräbs Santiago 2014b).

*Direct correspondence to*
Lars Leszczensky, Mannheim Centre for European Social Research (MZES), A5, 6, 68159 Mannheim, Germany
E-mail: Lars.Leszczensky@mzes.uni-mannheim.de

adequate measure of ethnic identity that must be structurally similar across different age groups and applicable to both immigrants and natives (1.2).

## 1.1 Defining Ethnic and National Identity

Ethnic identity is a special case of *social identity* (see Ashmore, Deaux, & McLaughlin-Volpe, 2004; Phinney, 1990). Social identity can generally be defined as

> "that part of an individual's self-concept which derives from his *knowledge* of his membership of a social group (or groups) together with the *value* and *emotional significance* attached to that membership" (Tajfel, 1978, p. 63, our italics).

According to this definition, a cognitive, an evaluative, and an emotional dimension of social identity can be distinguished (see Ashmore et al., 2004; Ellemers, Kortekaas, & Ouwerkerk, 1999; Jackson, 2002).

The *cognitive* dimension refers to the subjective knowledge of being a member of a social group. This self-categorization is a necessary condition for developing a sense of belonging to and attachment towards this group (Ashmore et al., 2004, pp. 84f.). Accordingly, a measure of ethnic and of national identity first has to capture whether or not immigrants actually identify themselves as members of the host country and/or as members of their own ethnic group. This includes as well capturing the presence of a dual identity, such as German-Turkish, which many immigrants may prefer above an exclusive ethnic or national identity (Verkuyten & Martinovic, 2012). However, dual identity does not necessarily mean that national and ethnic identities are equally strong (Simon & Ruhs, 2008; Simon & Grabow, 2010).

The *evaluative* dimension captures the value attached to a group membership, as well as related attitudes (Ashmore et al., 2004; Sellers, Smith, Shelton, Rowley, & Chavous, 1998). The subjective evaluation of a group is referred to as private regard. Immigrants may evaluate their membership of their own ethnic group and of the host country group as either positive or negative.

The *emotional* dimension refers to the affective commitment to a group, which is often considered to be the most important component of ethnic identity (Phinney & Ong, 2007, p. 272). The particular importance of the emotional dimension arises from the fact that it is the dimension most consequential for individual actions and in-group bias (see Ashmore et al., 2004; Ellemers et al., 1999; Jackson, 2002).

To sum up, e*thnic identity* is a multidimensional construct that encompasses not only the cognitive awareness of being a member of a particular ethnic group but also the subjective evaluation of this group membership and the emotional attachment to this group. *National identity*, by contrast, refers to the host country instead of to the immigrants' own ethnic group (see Phinney, 1990; Schwartz et al., 2012;

Verkuyten & Martinovic, 2012). In many European nations, ethnic and national identities are indistinguishable for the native majority population. Immigrants and their descendants, by contrast, can generally identify with their own ethnic group, with the host country, or with a combination of these two in form of a dual identity (see Berry, 2001; Phinney et al., 2006; Verkuyten & Martinovic, 2012).

## 1.2 Shortcomings of Existent Measures of Ethnic Identity

Compared to studies on ethnic identity, relatively little research has focused on immigrants' identification with the host country (see Verkuyten & Martinovic, 2012, p. 85). In addition, large-scale studies on emotional integration often rely on rather rough measures of ethnic identity (see Nandi & Platt, 2012). In contrast, social-psychological literature on how to measure ethnic identity is vast. Especially American social-psychologists have proposed various multidimensional measures of ethnic identity (for a review see Cokley, 2007). Most prominent are multigroup measures that can be applied to ethnically heterogeneous samples. The most frequently used of these measures is the *Multigroup Ethnic Identity Measure* (MEIM) developed by Phinney (1992). In the last two decades, the MEIM has been steadily revised and tested (e.g., Roberts, Phinney, Masse, Chen, Roberts, & Romero, 1999; Yoon, 2011). The most recent version is the MEIM-R (Phinney & Ong, 2007). As an alternative to the MEIM, Umaña-Taylor, Yazedjian, and Bámaca-Gómez (2004) proposed the *Ethnic Identity Scale* (EIS). Due to a similar substantive approach, however, the difference between EIS and MEIM-R is rather marginal (see Cokley, 2007). Based on the MEIM, Schwartz and colleagues (2012) recently proposed an analogous measure for American national identity.

Even though these social-psychological measures are well-established, for two reasons we believe that they are of limited use for the purpose of integration research. The first reason is that these measures are based on specific developmental approaches to social identity by Erikson (1968) and Marcia (1980). For the study of immigrants' emotional integration, however, a theoretically more open and flexible measure seems preferable (see Nandi & Platt, 2012). For instance, while the MEIM-R and EIS assess the process of identity exploration, they do not explicitly capture the evaluative and emotional dimensions of social identity. This is why the developers of these measures themselves stress that, depending on the research question, their measures have to be complemented by additional measures (Phinney & Ong, 2007, p. 278).

The second reason is that most established measures assess only ethnic identity but neglect national identity. Forcing immigrants to choose an ethnic group, however, dismisses the cognitive dimension of social identity. As a consequence, neither the MEIM-R nor EIS tells researchers whether respondents identify with the host country as well. As these measures typically do not include a native refer-

ence group, it is not possible to infer whether immigrants adjust towards the native population over time.

Making comparisons between different groups requires an adequate measure of ethnic and of national identity to be structurally similar across groups (Cokley 2007, p. 231; Schwartz et al., 2014). Importantly, for national identity this also includes the *native population*, which is needed as a baseline comparison group. Given that integration is an intergenerational process (e.g., Diehl & Schnell, 2006), the measure also has to be invariant for immigrants of the first, second, and third *immigrant generations*. Finally, children become aware of the societal significance and evaluation of ethnic groups at the age of 10 and start to develop an ethnic identity during adolescence (Phinney, 1990; Quintana, 1999, 2007). Evidence of measurement invariance across different *age groups* is a necessary prerequisite to understanding these individual developments (Phinney & Ong, 2007, p. 279).

## 2 Proposed Measure

In the development of our measure of ethnic and of national identity we draw on previous research, in particular on American measurements, e.g. *MEIM-R* (Phinney & Ong, 2007) and *EIS* (Umaña-Taylor et al., 2004). We created an item-pool and chose the items most understandable to native and non-native children and youths. To assess the comprehensibility we conducted two cognitive pretests (Leszczensky, 2012). The approved items were chosen for our measurement and further tested in a primary study. The findings supported the theoretically expected dimensionality of ethnic and of national identity and the applicability of our measure (Leszczensky & Pink, 2013). Due to the small number of cases, however, we could not conduct more extensive analyses (but see Leszczensky & Gräbs Santiago, 2014a).

The questions were used in paper-and-pencil questionnaires administered during lessons in school. First, students answered questions concerning their national identity, as these applied to both native and non-native students.[3] Next, the ethnic group of immigrant children and youths had to be defined. We used the phrasing "my family's country of origin", because students understood it much better than an alternative formulation like "ethnic group" (Leszczensky, 2012). In addition, by referring to the country of origin of the family rather than to the country of origin of the respondent himself, this formulation explicitly includes children whose parents or grandparents were born abroad, but who themselves were born in Germany. If the family members have different countries of origin, we asked the students to choose the most important one. Subsequently, immigrant students answered ques-

3 In future applications, a randomization of the sequence of national and ethnic identity measures may be considered to investigate the possibility of effects of the ordering of questions.

*Table 1* Proposed Measure of National and Ethnic Identity by Dimensions

| Dimension | Item | Response Categories |
|---|---|---|
| *Cognitive* | | |
| Self-Categorization | What do you consider yourself to be? | "Only as German"<br>"More German"<br>"Both equally"<br>"More like a person from my family's country of origin"<br>"Only as a person from my family's country of origin" |
| Dual Identity | Some people consider themselves German, others, for example, Turkish, and others again Turkish-German. How about you? What do you consider yourself to be? | Half-open: German; Turkish; German-Turkish; Kurdish; German-Kurdish; Italian; German-Italian; Polish; German-Polish; Something else, namely: |
| *Evaluative* | | |
| Private Regard | I am satisfied to belong to Germany/my family's country of origin. | Five-Point Scale (applies) |
| | I am glad to belong to Germany/my family's country of origin. | Five-Point Scale (applies) |
| *Emotional* | | |
| Attachment | It bothers me if somebody speaks ill about Germany/my family's country of origin. | Five-Point Scale (applies) |
| | Germany/My family's country of origin is dear to me. | Five-Point Scale (applies) |
| | I feel strongly attached to Germans/people from my family's country of origin. | Five-Point Scale (applies) |
| | I feel like I am part of Germany/my family's country of origin. | Five-Point Scale (applies) |

tions regarding the cognitive dimension and ethnic identity. Table 1 shows our proposed measure of national and of ethnic identity (see table A1 in the appendix for the original German wording of the items).

We use two questions to assess the cognitive dimension. On one hand, immigrant children and youths indicated on a five-point scale their *self-categorization* as German or as a person of their family's country of origin. Through the half-

open question on *dual identity* we further distinguish between students with a single identity regarding Germany or their family's country of origin and those with a dual identity. We measure the evaluative and emotional dimensions for both national and ethnic identity, respectively. Two items capture *private regard*, which is a key element of the evaluative dimension *Attachment,* a key element of the emotional dimension, is assessed by four items. Students rated their agreement with the items on five-point Likert scales. The responses were coded such that higher scores indicate stronger approval.

# 3 Data and Methods

## 3.1 Sample

We use data from the first wave of the project "Friendship and Identity in School" to test our proposed measure of ethnic and of national identity (Leszczensky, Pink, & Kalter, 2014). The data were collected in the fifth, sixth, and seventh grades of nine schools in North Rhine-Westphalia. The school sample consists of lower secondary, intermediate secondary, and comprehensive schools with a higher share of immigrants. Schools were randomly chosen within predefined strata regarding different numbers of non-native students. The overall participation rate was 76.5%. Therefore, our analyses are based on the data of 1,668 students.

At the time of the survey in April and May 2013 students were between 9 and 17 years old ($M = 12.77$; $SD = 1.14$). 18% of the students attended a lower secondary, 36% an intermediate secondary, and 46% a comprehensive school. 63% of the respondents had a migration background. The majority of them stem from Turkey (38%), followed by Poland (10%) and Russia (7%).[4] Due to the method of collection the data are not representative. The sample is negatively selected in regard to the school type, and thus, to the social background of the students. However, there is no reason to doubt that the measure would operate at least equally well in a representative sample.

## 3.2 Covariates

In the analyses we differentiate between native and immigrant children and youths. Persons with at least one grandparent born abroad were defined as *immigrants*. We also consider the *immigrant generation*. First-generation immigrants are students who were born abroad and migrated themselves to Germany within the first six

4 The ethnic origin of the students is based on the information regarding their country of birth, and those of their parents and grandparents, e.g. if at least one parent or grandparent was born in Turkey the student is defined as Turkish.

years of their lives. Students who were born abroad and migrated themselves to Germany before they were six years old are defined as second-generation immigrants, as are of whom at least one parent was born abroad. Students who were born in Germany and whose parents were also born in Germany are third-generation immigrants if at least one grandparent was born abroad. Furthermore, students were categorized into three different *age groups*: 9 to 11, 12 to 13, and 14 to 17 years.

To test the construct validity of our measure, we examine the correlation between ethnic and national identities as well as that between indicators of social and of cultural integration, which have been found to be associated with immigrants' identification (Phinney et al., 2006; Zander & Hannover, 2013; Schulz & Leszczensky, 2015). We assess social integration by the *share of friends* from Germany as well as by the family's country of origin. Regarding their friends who do not visit the same school the students were asked to rate on a five-point scale ranging from "all" to "none" the share of friends from Germany or from the family's country of origin, respectively. The scales were coded such that higher values indicate a greater number of friends. Cultural integration is measured by self-evaluation of *the knowledge of the German language* and *the language of the family's country of origin*. The students judged their skills with respect to speaking, comprehension, writing and reading on a four- and on a three-point scale, respectiveley. We constructed two indices by taking the average of the four items for the German language (Cronbach's $\alpha = 0.82$) and the four items for the language of the family's country of origin (Cronbach's $\alpha = 0.84$). Higher scores on the scales express better skills in the respective language.

## 3.3 Analytical Strategy

Since the subjective knowledge of being a member of the host country and/or a member of the family's country of origin is a necessary condition for the evaluation of and emotional attachment to a respective identity, we first analyze the cognitive dimension. In particular, we describe the self-categorization and the presence of dual identity distinguished by immigrant generation (4.1).

Next, we analyze the evaluative and emotional dimensions of national and of ethnic identity, respectively (4.2). For an initial evaluation of the items we show descriptive statistics and intercorrelations. Then we conduct confirmatory factor analyses to check whether the evaluative and emotional dimensions of ethnic and of national identity can actually be identified empirically (Brown, 2006, p. 49).[5]

5 To empirically assess the underlying factor structure we previously ran explanatory factor analyses for each national and ethnic identity. In accordance with our theoretical expectations, in both cases the items loaded on two factors that can be labeled private regard and attachment. Because the cognitive dimension forms the basis for the evalua-

Subsequently, we test for measurement invariance across native and non-native students, across immigrant generations, and across age groups (Vandenberg & Lance, 2000).[6] For this purpose, we conduct separate confirmatory factor analyses for the single groups, as well as multiple-group confirmatory factor analysis (Brown, 2006, pp. 268ff.). Thereby, we test for measurement invariance on three different levels (Jöreskog, 1971). *Configural* invariance assumes the same factor structure across all groups; *metric* invariance supposes additionally the same factor loadings; s*calar* invariance requires, besides the previous conditions, the same item intercepts across groups. Only if scalar invariance is ascertained, are comparisons between groups legitimated (Brown, 2006, pp. 268ff.; Kline, 2011, pp. 251ff.).

In the last step, we construct subscales for national and for ethnic identity and examine their reliability (4.3). We also test for construct validity by looking at the correlation between the dimensions of national and of ethnic identity as well as their respective relation to indicators of social and of cultural integration.

# 4 Results

## 4.1 Cognitive Dimension

As presented in table 2, we examine the distribution of both self-categorization and of dual identity dependent on immigrant generation. In line with the hypothesis of intergenerational assimilation, the vast majority of first-generation immigrants in our sample categorize themselves as members of their family's country of origin rather than as German. In contrast, half of the youths in the second generation see themselves as members of both countries, and nearly half of the third-generation immigrants identify themselves as German only. It bears mentioning that about one third of both the first-generation and the third-generation immigrants categorize themselves as members of both countries. The importance of dual identity is confirmed by the fact that half of the immigrant youths state having a dual identity. Differences between immigrant generations are surprisingly small.

To further compare both measures, table 3 displays the relation of self-categorization and dual identity. Almost two thirds of the children and youths with a dual identity see themselves both as German and as persons of their family's country of origin. Consistent with previous research, however, a dual identity does not seem to imply that national and ethnic identities are equally strong (see Simon & Ruhs, 2008; Simon & Grabow, 2010). This finding is further supported by the fact that a

tive and emotional dimensions, we did not include this dimension in the factor analyses (see, e.g., Phinney, 1992).

6 We checked whether immigrant generation and age are correlated. As they are not, we treat them separately in our analyses of measurement invariance.

*Table 2* Self-Categorization and Dual Identity by Generation

| | Immigrant Generation | | | | |
|---|---|---|---|---|---|
| Self-Categorization | 1st gen. | 2nd gen. | 3rd gen. | Total | N |
| Only German | 3.39% | 7.28% | 31.58% | 10.15% | 90 |
| More German | 10.17% | 5.88% | 14.91% | 7.33% | 65 |
| Both equally | 27.12% | 49.44% | 35.96% | 46.22% | 410 |
| More FCO | 25.42% | 20.87% | 11.40% | 19.95% | 177 |
| Only FCO | 33.90% | 16.53% | 6.14% | 16.35% | 145 |
| Total | 100.00% | 100.00% | 100.00% | 100.00% | |
| N | 59 | 714 | 114 | | 887 |
| Dual Identity | 30.00% | 52.59% | 41.59% | 49.66% | 440 |
| Only German Identity | 0.00% | 10.52% | 42.48% | 13.88% | 123 |
| Only FCO Identity | 70.00% | 36.89% | 15.93% | 36.46% | 323 |
| Total | 100.00% | 100.00% | 100.00% | 100.00% | |
| N | 60 | 713 | 113 | | 886 |

*Note:* FCO = (person from) family's country of origin

*Table 3* Self-Categorization and Dual Identity

| | Presence of Dual Identity | | | | |
|---|---|---|---|---|---|
| Self-Categorization | Dual Identity | Only German Identity | Only FCO Identity | Total | N |
| Only German | 4.23% | 50.00% | 2.49% | 9.74% | 94 |
| More German | 8.46% | 14.62% | 1.38% | 6.63% | 64 |
| Both equally | 64.27% | 28.46% | 25.41% | 44.87% | 433 |
| More FCO | 16.70% | 3.08% | 32.60% | 20.83% | 201 |
| Only FCO | 6.34% | 3.85% | 38.12% | 17.93% | 173 |
| Total | 100.00% | 100.00% | 100.00% | 100.00% | |
| N | 473 | 130 | 362 | | 965 |

*Note:* FCO = (person from) family's country of origin

considerable number of youths with a single ethnic or national identity nevertheless indicate seeing themselves both as German and as a person of their family's country of origin.

## 4.2 Evaluative and Emotional Dimension

Descriptive statistics for the evaluative and emotional dimensions of both national and ethnic identity are presented in table 4, showing left-skewed distributions for all items for national identity. The mean values lie between around 3.0 and 4.0 on the five-point scale. The items for ethnic identity show even stronger left-skewed distributions and higher mean values. As in other studies, immigrants' ethnic identity thus is stronger than their identification with the host country (e.g., Phinney et al., 2006; Zander & Hannover, 2013). The items of each identity are strongly correlated (see table A2 in the appendix). We conduct confirmatory factor analyses to assess the two-dimensionality of national and of ethnic identity. Due to the non-normality of the items, we use maximum-likelihood estimators with robust standard errors and Satorra-Bentler scaled values (Brown, 2006, p. 76; Satorra & Bentler, 2001).[7]

In the case of national identity the results indicate a good fit of the model with the two factors private regard and attachment ( $\chi^2 = 62.36$, df = 8, TLI = 0.97, CFI = 0.99, RMSEA = 0.07, SRMR = 0.02, AIC = 26261.55). Concerning ethnic identity the analysis yield similar results, showing an adequate fit of the two-dimensional model ( $\chi^2 = 22.47$, df = 8, TLI = 0.97, CFI = 0.98, RMSEA = 0.04, SRMR = 0.02, AIC = 12985.94).[8]

Based on these findings we test the two-dimensional model of each identity for measurement invariance. We begin with the analyses of national identity (see table A3 in the appendix). First, we check the equivalence of the respective measurements for immigrants and for natives. The single-group analyses show a good model fit for both groups. Also the results of the multiple-group confirmatory factor analysis indicate scalar invariance between immigrants and natives.[9] Second, we consider different immigrant generations. The results of the test suggest the equivalence of the respective measurements for immigrants of the first, second, and third generations. Finally, we test for structural similarity across age groups. The analyses confirm the equivalence of the measurement of national identity across different age groups.

---

7 We ran our analyses with the lavaan-package (version 0.5-15) in R (version 3.0.2) (Rosseel, 2012). Following values suggest a good fit: TLI > 0.95, CFI > 0.95, RMSEA < 0.08, SRMR < 0.08 (Hu & Bentler, 1999; Brown, 2006, pp. 86f.).

8 We also conducted confirmatory factor analyses for a one-dimensional model for each national and ethnic identity. The results indicate a poor fit of the model in case of both national ($\chi^2 = 504.08$, df = 9, TLI = 0.79, CFI = 0.87, RMSEA = 0.19, SRMR = 0.07, AIC = 26881.43) and ethnic identity ($\chi^2 = 157.18$, df = 9, TLI = 0.72, CFI = 0.83, RMSEA = 0.13, SRMR = 0.06, AIC = 13338.95). The lower values of the AIC also show the comparatively better fit of the two-dimensional model.

9 In the multiple-group confirmatory factor analysis a negative $\Delta$ CFI value lower than -0.01 indicates a lack of measurement invariance (Cheung & Rensvold, 2002; Dimitrov, 2010).

*Table 4* Items for National and Ethnic Identity and Descriptive Statistics

| | | *M* | *SD* | Skewness | Kurtosis | N |
|---|---|---|---|---|---|---|
| National Identity | | | | | | |
| N1 | I am satisfied to belong to Germany. | 4.09 | 1.05 | -1.20 | 3.95 | 1,629 |
| N2 | I am glad to belong to Germany. | 3.94 | 1.12 | -0.97 | 3.26 | 1,632 |
| N3 | It bothers me if somebody speaks ill about Germany. | 2.96 | 1.36 | -0.05 | 1.80 | 1,635 |
| N4 | Germany is dear to me. | 3.16 | 1.28 | -0.23 | 2.01 | 1,631 |
| N5 | I feel strongly attached to Germans. | 3.07 | 1.27 | -0.15 | 2.01 | 1,629 |
| N6 | I feel like I am part of Germany. | 3.30 | 1.30 | -0.38 | 2.07 | 1,633 |
| Ethnic Identity | | | | | | |
| E1 | I am satisfied to belong to my family's country of origin. | 4.56 | 0.82 | -2.26 | 8.57 | 999 |
| E2 | I am glad to belong to my family's country of origin. | 4.55 | 0.83 | -2.17 | 7.86 | 1,000 |
| E3 | It bothers me if somebody speaks ill about my family's country of origin. | 4.38 | 1.08 | -1.91 | 5.87 | 1,006 |
| E4 | My family's country of origin is dear to me. | 4.48 | 0.85 | -1.91 | 6.82 | 998 |
| E5 | I feel strongly attached to people from my family's country of origin. | 4.19 | 1.09 | -1.38 | 4.18 | 995 |
| E6 | I feel like I am part of my family's country of origin. | 4.35 | 0.99 | -1.73 | 5.67 | 994 |

We also test measurement invariance of the ethnic identity (see table A4 in the appendix). The analyses suggest an equivalent measurement for different generations of immigrants. As in the case of national identity, the analyses also confirm the equivalence of the measurement across different age groups.[10]

10 For both national and ethnic identity, we also tested invariance across two broader age groups, namely 9 to 12 and 13 to 17 years. The results were similar as in the analyses presented here.

*Table 5* Reliability and Descriptive Statistics of the Scales for National and for Ethnic Identity

| | | | Natives vs. Immigrants | | Immigrant Generation | | |
|---|---|---|---|---|---|---|---|
| | | Total | Natives | Immig. | 1st gen. | 2nd gen. | 3rd gen. |
| National Identity | | | | | | | |
| Private Regard | *M* | 4.01 | 4.31 | 3.85 | 3.46 | 3.86 | 4.05 |
| | $\alpha$ | 0.85 | 0.81 | 0.86 | 0.88 | 0.86 | 0.76 |
| Attachment | *M* | 3.13 | 3.51 | 2.92 | 2.77 | 2.88 | 3.20 |
| | $\alpha$ | 0.85 | 0.84 | 0.84 | 0.82 | 0.84 | 0.81 |
| Ethnic Identity | | | | | | | |
| Private Regard | *M* | | | 4.56 | 4.60 | 4.58 | 4.25 |
| | $\alpha$ | | | 0.88 | 0.85 | 0.88 | 0.88 |
| Attachment | *M* | | | 4.36 | 4.48 | 4.38 | 3.91 |
| | $\alpha$ | | | 0.80 | 0.81 | 0.77 | 0.86 |

## 4.3 Reliability and Construct Validity

Since our analyses clearly support the model with two dimensions for both national and ethnic identity we construct subscales for *private regard* and *attachment* separately for national and for ethnic identity. The scales are constructed by the mean of the corresponding items. We examine the reliability for the total sample as well as differentiated by natives and immigrants, and by immigrant generation. The subscales are highly reliable, with Cronbach's α values greater than 0.75 for all groups (see table 5 for more details about the reliability and the descriptive statistics). Substantively, the order of mean values is as expected.

To check the construct validity of the measurement we inspected the correlation between the dimensions of national and of ethnic identity (see table 6). As expected, private regard and attachment in terms of each ethnic and each national identity are strongly correlated (see Jackson, 2002; Zander & Hannover, 2013). While previous research typically shows a negative relationship between national and ethnic identities of German immigrants (see Berry, Phinney, Sam, & Vedder, 2006; Chryssochoou & Lyons, 2011; Verkuyten & Yildiz, 2007; Zander & Hannover, 2013), in our sample national and ethnic identities are not related to each other at all.

*Table 6* Correlations among the Dimensions of National and of Ethnic Identity

| | NI-Private Regard | NI-Attachment | EI-Private Regard | EI-Attachment |
|---|---|---|---|---|
| NI-Private Regard | 1.00 | | | |
| NI-Attachment | 0.59*** | 1.00 | | |
| EI-Private Regard | 0.02 | -0.02 | 1.00 | |
| EI-Attachment | 0.01 | 0.07* | 0.63*** | 1.00 |

*Note*: *p<0.05, **p<0.01, ***p<0.001, N=950; NI = National Identity; EI = Ethnic Identity

*Table 7* Correlations of the Dimensions of National and of Ethnic Identity with Indicators of Social and of Cultural Integration

| | Friends | | Language Skills | |
|---|---|---|---|---|
| | German | FCO | German | FCO |
| NI-Private Regard | 0.18*** | -0.03 | 0.16*** | -0.08* |
| NI-Attachment | 0.15*** | -0.09** | 0.09** | -0.11* |
| EI-Private Regard | -0.05 | 0.19*** | -0.07* | 0.26*** |
| EI-Attachment | -0.09** | 0.23*** | -0.09** | 0.28*** |

*Note*: *p<0.05, **p<0.01, ***p<0.001; FCO = family's country of origin; NI = National Identity; EI = Ethnic Identity

To further assess the construct validity we examined the relationship between national and ethnic identities and indicators of the social and the cultural integration of immigrant children and youths. As presented in table 7, the results are in line with theoretical expectations as well as with the results of previous research (see, e.g., Agirdag, Van Houtte, & Van Avermaet, 2011; Leszczensky, 2013; Phinney et al., 2006; Sabatier, 2008; Zander & Hannover, 2013).

To sum up, the subscales for the evaluative and the emotional dimensions of national and of ethnic identity are reliable and capture expected substantive differences between immigrants and natives as well as across immigrant generations. The relation between private regard and attachment as well as their association with indicators of social and of cultural integration correspond with theoretical expectations. Taken together, our findings therefore suggest that the proposed measure is valid.

# 5 Discussion

As previous research on young immigrants' identification suffers from the lack of an adequate measure of ethnic and of national identity, our aim was to propose and test such a measure. For this purpose, we first specified the concept of ethnic and of national identity by referring to a common definition that distinguishes three key dimensions of social identity. Arguing that, for a variety of reasons, established measures of ethnic identity are of rather limited use in integration research, we introduced our own measure of ethnic and of national identity. In particular, in contrast to established measures of ethnic identity, our measure captures the cognitive, evaluative, and emotional dimensions of ethnic as well as of national identity.

We tested our proposed measure using data from immigrant and native youths. Our results clearly confirmed the supposed dimensionality of ethnic and of national identity. We further ascertained the invariance of our measurement between native and non-native students, across immigrant generations, and across age-groups. Measurement invariance is not only required to meaningfully compare groups but also to analyze individual developments over time. The constructed subscales are reliable, and correlations between the subscales and indicators of social and of cultural integration are in line with both theoretical expectations and findings of earlier studies.

Our proposed measure thus offers an adequate instrument that captures crucial dimensions of youths' ethnic and national identities. Our measurement is ready for use in school surveys (see Leszczensky & Pink 2015) and available for researchers via ZIS/GESIS (Leszczensky & Gräbs Santiago 2014b), as the items are understood by children and adolescents and can be answered relatively quickly in written questionnaires. Since our measure is invariant across age groups, it may be especially useful for panel studies that are interested in the intra-individual development and change in ethnic and national identities. Finally, as the different dimensions of ethnic and national identities are measured separately, our measurement is flexible enough to address different types of research questions. For example, our measure allows examination of the consequences of the emotional dimension of ethnic and of national identity, which may influence in-group bias (Ellemers et al., 1999) or friendship selection (Leszczensky, 2013). The flexibility of our measure also extends to the presence of dual identity, which is assessed directly, but which may also be constructed by combining subscales of ethnic and of national identity.

Besides addressing substantial questions regarding the causes and consequences of ethnic and of national identities with the help of our measure, future studies may provide further tests of our measure. For instance, while our sample did not allow conducting immigrant-group-specific analyses, it would be important to check whether our measurement is invariant across different ethnic groups as well. Similarly, while our sample included a relatively wide age range, future stud-

ies may test whether our measure can be understood by even younger children and whether it is still applicable to emerging adults.

## References

Agirdag, O., Van Houtte, M., & Van Avermaet, P. (2011). Ethnic school context and the national and subnational identifications of pupils. *Ethnic and Racial Studies, 34*, 357-378.

Altschul, I., Oyserman, D., & Bybee, D. (2006). Racial-ethnic identity in mid-adolescence: content and change as predictors of academic achievement. *Child Development, 77*, 1155-1169.

Ashmore, R. D., Deaux, K., & McLaughlin-Volpe, T. (2004). An organizing framework for collective identity: articulation and significance of multidimensionality. *Psychological Bulletin, 130*, 80-114.

Berry, J. W. (2001). A psychology of immigration. *Journal of Social Issues, 57*, 615-631.

Berry, J. W., Phinney, J., Sam, D., & Vedder, P. (2006). Immigrant youth: acculturation, identity, and adaptation. *Applied Psychology: An International Review, 55*, 303-332.

Brown, T. M. (2006). *Confirmatory factor analyses for applied research*. New York: Guilford.

Casey, T., & Dustmann, C. (2010). Immigrants' identity, economic outcomes and the transmission of identity across generations. *The Economic Journal, 120*, F31-F51.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255.

Chryssochoou, X., & Lyons, E. (2011). Perceptions of (in)compatibility between identities and participation in the national polity of people belonging to ethnic minorities. In A. E. Azzi, X. Chryssochoou, B. Klandermans, & B. Simon (Eds), *Identity and Participation in Culturally Diverse Societies. A Multidisciplinary Perspective* (pp. 69-88). West Sussex: Wiley-Blackwell.

Cokley, K. (2007). Critical issues in the measurement of ethnic and racial identity: a referendum on the state of the field. *Journal of Counseling Psychology, 54*, 224-234.

Diehl, C., & Schnell, R. (2006). 'Reactive ethnicity' or 'assimilation'? Statements, arguments, and first empirical evidence for labor migrants in Germany. *International Migration Review, 40*, 786-816.

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43*, 121-149.

Ellemers, N., Kortekaas, P., & Ouwerkerk, J. W. (1999). Self-categorisation, commitment to the group and group self-esteem as related but distinct aspects of social identity. *European Journal of Social Psychology, 29*, 371-389.

Erikson, E. (1968). *Identity: Youth and Crisis*. New York: Norton.

Ersanilli, E., & Saharso, S. (2011). The settlement country and ethnic identification of children of Turkish immigrants in Germany, France, and the Netherlands: what role do national integration policies play? *International Migration Review, 45*, 907-937.

Gordon, M. M. (1964). *Assimilation in American life*. Oxford University Press, New York

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Jackson, J. W. (2002). Intergroup attitudes as a function of different dimensions of group identification and perceived intergroup conflict. *Self and Identity, 1*, 11-33.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36 (4),* 409-426.

Kalter, F. (2008). Stand, Herausforderungen und Perspektiven der Migrationsforschung. In F. Kalter (Ed.), *Migration und Integration. Sonderheft 48 der Kölner Zeitschrift für Soziologie und Sozialpsychologie* (pp. 11-36). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kline, R. B. (2011). *Principles and practice of structural equation modeling.* New York: Guilford.

Leszczensky, L. (2012). *Dokumentation des kognitiven Pretests im Rahmen des Projektes "Soziale Netzwerke und ethnische Identifikationen von jugendlichen Migranten".* Mannheimer Zentrum für Europäische Sozialforschung.

Leszczensky, L. (2013). Do national identification and interethnic friendships affect one another? A longitudinal test with adolescents of Turkish origin in Germany. *Social Science Research*, *42*, 775-788.

Leszczensky, L., & Gräbs Santiago, A. (2014a). *Die Messung ethnischer und nationaler Identität von Kindern und Jugendlichen.* Mannheimer Zentrum für Europäische Sozialforschung, Working Paper, 155.

Leszczensky, L., & Gräbs Santiago, A. (2014b). Ethnische und nationale Identität von Kindern und Jugendlichen. In D. Danner, & A., Glöckner-Rist (Eds), *Zusammenstellung sozialwissenschaftlicher Items und Skalen*. doi: 10.6102/zis158

Leszczensky, L., & Pink, S. (2013). *Dokumentation des Instrumententests im Rahmen des Projektes "Soziale Netzwerke und ethnische Identifikationen von jugendlichen Migranten".* Mannheimer Zentrum für Europäische Sozialforschung.

Leszczensky, L., Pink, S., & Kalter, F. (2014). *Freundschaft und Identität in der Schule: Feldbericht zu Welle 1 und Welle 2 (Technical Report).* Mannheimer Zentrum für Europäische Sozialforschung, Working Paper, 157.

Leszczensky L., & Pink, S. (2015). Ethnic segregation of friendship networks in school: Testing a rational-choice argument of differences in ethnic homophily between classroom- and grade-level networks. *Social Networks, 42*, 18-26. doi:10.1016/j.socnet.2015.02.002

Marcia, J. E. (1980). Identity in adolescence. In J. Adelson (Ed.), *Handbook of adolescent psychology* (Vol. 9, pp. 159-187). New York: Wiley & Sons.

Meeus, W. (2011). The study of adolescent identity formation 2000-2010: a review of longitudinal research. *Journal of Research on Adolescence*, *21*, 75-94.

Nandi, A., & Platt, L. (2012). Developing ethnic identity questions for Understanding Society, the UK Household Longitudinal Study. *Understanding Society Working Paper Series, 3*, 80-100.

Nauck, B. (2001). Intercultural contact and intergenerational transmission in immigrant families. *Journal of Cross-Cultural Psychology, 32*, 159-173.

Nekby, L., & Rödin, M. (2010). Acculturation identity and employment among second and middle generation immigrants. *Journal of Economic Psychology, 31*, 35-50.

Phinney, J. S. (1990). Ethnic identity in adolescents and adults: review of research. *Psychological Bulletin, 108*, 499-514.

Phinney, J. S. (1992). The Multigroup Ethnic Identity Measure. *Journal of Adolescent Research, 7*, 156-176.

Phinney, J. S., Berry, J. W., Vedder, P., & Liebkind, K. (2006). The acculturation experience: attitudes, identities, and behaviors of immigrant youth. In J. W. Berry (Ed.), *Immigrant youth in cultural transition: acculturation, identity, and adaptation across national contexts* (pp. 71-116). Mahwah, NJ: Erlbaum.

Phinney, J. S., & Ong, A. D. (2007). Conceptualization and measurement of ethnic identity: current status and future directions. *Journal of Counseling Psychology, 54*, 271-281.

Quintana, S. M. (1999). Children's developmental understanding of ethnicity and race. *Applied and Preventive Psychology, 7*, 27-45.

Quintana, S. M. (2007). Racial and ethnic identity: developmental perspectives and research. *Journal of Counseling Psychology, 54*, 259-270.

Roberts, R. E., Phinney, J. S., Masse, L. C., Chen, Y., Roberts, C. R., & Romero, A. (1999). The structure of ethnic identity of young adolescents from diverse ethnocultural groups. *The Journal of Early Adolescence, 19*, 301-322.

Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *Journal of Statistical Software, 48 (2),* 1-36.

Rutland, A., Cameron, L., Jugert, P., Nigbur, D., Brown, R., Waters, C., Hossain, R., Landau, A., & Le Touze, D. (2012). Group identity and peer relations: a longitudinal study of group identity, perceived peer acceptance, and friendships amongst ethnic minority English children. *British Journal of Developmental Psychology, 2*, 282-302.

Sabatier, C. (2008). Ethnic and national identity among secondgeneration immigrant adolescents in France: the role of social context and family. *Journal of Adolescence, 31*, 185-205.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507-514.

Schulz, B., & Leszczensky, L. (2015). Native friends and host country identification among adolescent immigrants in Germany: The role of ethnic boundaries. International Migration Review. doi: 10.1111/imre.12163

Schwartz, S. J., Park, I. J. K., Huynh, Q.-L., Zamboanga, B. L., Umaña-Taylor, A. J., Lee, R. M., Rodriguez, L., Kim, S. Y., Whitbourne, S. K., Castillo, L. G., Weisskirch, R. S., Vazsonyi, A. T., Williams, M. K., & Agocha, V. B. (2012). The American identity measure: development and validation across ethnic group and immigrant generation. *Identity, 12*, 93-128.

Schwartz, S. J., Syed, M., Yip, T., Knight, G. P., Umaña-Taylor, A. J., Rivas-Drake, D., & Lee, R. M. (2014). Methodological issues in ethnic and racial identity research with ethnic minority populations: theoretical precision, measurement Issues, and research designs. *Child Development, 85*, 58-76.

Sellers, R. M., Smith, M. A., Shelton, J., Rowley, S. A., & Chavous, T. M. (1998). Multidimensional model of racialidentity: a reconceptualization of African American racialidentity. *Personality and Social Psychology Review, 2*, 18-39.

Simon, B., & Grabow, O. (2010). The politicization of migrants: further evidence that politicized collective identity is a dual identity. *Political Psychology, 31*, 717-738.

Simon, B., & Ruhs, D. (2008). Identity and politicization among Turkish migrants in Germany: the role of dual identification. *Journal of Personality and Social Psychology, 95*, 1354-1366.

Steinbach, A. (2004). *Soziale Distanz: Ethnische Grenzziehung und die Eingliederung von Zuwanderern in Deutschland*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Tajfel, H. (1978). Social categorization, social identity, and social comparison. In H. Tajfel (Ed.), *Differentiation between social groups* (pp. 61-76). London: Academic Press.

Umaña-Taylor, A. J., Yazedjian, A., & Bámaca-Gómez, M. (2004). Developing the Ethnic Identity Scale using Eriksonian and social identity perspectives. *Identity: An International Journal of Theory and Research, 4*, 9-38.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70.

Verkuyten, M., & Martinovic, B. (2012). Immigrants' national identification: meanings, determinants, and consequences. *Social Issues and Policy Review, 6*, 82-112.

Verkuyten, M., & Yildiz, A. (2007). National (dis)identification and ethnic and religious identity: a study among Turkish-Dutch Muslims. *Personality and Social Psychology Bulletin, 33*, 1448-1462.

Yoon, E. (2011). Measuring ethnic identity in the Ethnic Identity Scale and the Multigroup Ethnic Identity Measure-Revised. *Cultural Diversity and Ethnic Minority Psychology, 17*, 144-155.

Zander, L., & Hannover, B. (2013). Die Bedeutung der Identifikation mit der Herkunftskultur und mit der Aufnahmekultur Deutschland für die soziale Integration Jugendlicher mit Migrationshintergrund in ihrer Schulklasse. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *45*, 142-160.

# Appendix

*Table A1* Original Wording of the Proposed Measure of National and of Ethnic Identity by Dimensions

| Dimension | Item | Response Categories |
|---|---|---|
| *Cognitive* | | |
| Self-Categorization | Als was siehst du dich selbst? | „Nur als Deutschen“<br>„Mehr als Deutschen“<br>„Als Beides gleichermaßen“<br>„Mehr als Menschen aus dem Herkunftsland meiner Familie“<br>„Nur als Menschen aus dem Herkunftsland meiner Familie“ |
| Dual Identity | Manche Menschen sehen sich als deutsch an, andere zum Beispiel als türkisch und wieder andere als deutsch-türkisch. Wie ist das bei dir? Als was siehst du dich? | Halboffen: Deutsch; Türkisch; Deutsch-Türkisch; Kurdisch; Deutsch-Kurdisch; Italienisch; Deutsch-Italienisch; Polnisch; Deutsch-Polnisch; Etwas anderes, und zwar: … |
| *Evaluative* | | |
| Private Regard | Ich bin zufrieden damit, zu Deutschland/ zum Herkunftsland meiner Familie zu gehören. | 5er-Skala (Trifft-zu) |
| | Ich bin froh, zu Deutschland/ zum Herkunftsland meiner Familie zu gehören. | 5er-Skala (Trifft-zu) |
| *Emotional* | | |
| Attachment | Es stört mich, wenn jemand schlecht über Deutschland/ das Herkunftsland meiner Familie spricht. | 5er-Skala (Trifft-zu) |
| | Deutschland/Das Herkunftsland meiner Familie liegt mir sehr am Herzen. | 5er-Skala (Trifft-zu) |
| | Ich fühle mich eng verbunden mit den Deutschen/ Menschen aus dem Herkunftsland meiner Familie. | 5er-Skala (Trifft-zu) |
| | Ich fühle mich als Teil von Deutschland/ des Herkunftslandes meiner Familie. | 5er-Skala (Trifft-zu) |

*Table A2* Intercorrelations among the Items for Each National and Each Ethnic Identity

| National Identity (N=1,596) | | | | | | |
|---|---|---|---|---|---|---|
| | N1 | N2 | N3 | N4 | N5 | N6 |
| N1 | 1.00 | | | | | |
| N2 | 0.75*** | 1.00 | | | | |
| N3 | 0.35*** | 0.37*** | 1.00 | | | |
| N4 | 0.50*** | 0.53*** | 0.59*** | 1.00 | | |
| N5 | 0.43*** | 0.47*** | 0.49*** | 0.69*** | 1.00 | |
| N6 | 0.51*** | 0.55*** | 0.50*** | 0.64*** | 0.63*** | 1.00 |
| Ethnic Identity (N=978) | | | | | | |
| | E1 | E2 | E3 | E4 | E5 | E6 |
| E1 | 1.00 | | | | | |
| E2 | 0.79*** | 1.00 | | | | |
| E3 | 0.28*** | 0.32*** | 1.00 | | | |
| E4 | 0.57*** | 0.61*** | 0.45*** | 1.00 | | |
| E5 | 0.46*** | 0.47*** | 0.39*** | 0.60*** | 1.00 | |
| E6 | 0.54*** | 0.56*** | 0.39*** | 0.63*** | 0.63*** | 1.00 |

*Note*: *p<0.05, **p<0.01, ***p<0.001; for the items, see table 4

*Table A3* Measurement Invariance of National Identity

| *Native vs. Immigrants (N=1,596)* | $\chi 2$ | df | $\Delta \chi 2$ | $\Delta$ df | TLI | CFI | $\Delta$ CFI | RMSEA | SRMR |
|---|---|---|---|---|---|---|---|---|---|
| Native (n=561) | 18.37 | 8 | | | 0.98 | 0.99 | | 0.05 | 0.02 |
| Immigrants (n=1,035) | 48.04 | 8 | | | 0.97 | 0.99 | | 0.07 | 0.03 |
| Configural invariance | 67.83 | 16 | | | 0.97 | 0.99 | | 0.06 | 0.02 |
| Metric invariance | 78.83 | 20 | 10.50 | 4 | 0.97 | 0.98 | 0.00 | 0.06 | 0.03 |
| Scalar invariance | 92.76 | 24 | 13.61 | 4 | 0.98 | 0.98 | 0.00 | 0.06 | 0.04 |
| *Immigrant Generation (N=926)* | $\chi 2$ | df | $\Delta \chi 2$ | $\Delta$ df | TLI | CFI | $\Delta$ CFI | RMSEA | SRMR |
| 1st generation (n=57) | 3.96 | 8 | | | 1.05 | 1.00 | | 0.00 | 0.04 |
| 2nd generation (n=718) | 30.89 | 8 | | | 0.98 | 0.99 | | 0.06 | 0.03 |
| 3nd generation (n=151) | 22.42 | 8 | | | 0.89 | 0.94 | | 0.11 | 0.05 |
| Configural invariance | 54.18 | 24 | | | 0.98 | 0.99 | | 0.06 | 0.03 |
| Metric invariance | 64.92 | 32 | 6.97 | 8 | 0.98 | 0.99 | 0.00 | 0.06 | 0.03 |
| Scalar invariance | 87.31 | 40 | 25.49 | 8 | 0.98 | 0.98 | 0.00 | 0.06 | 0.04 |
| *Age (N=1,558)* | $\chi 2$ | df | $\Delta \chi 2$ | $\Delta$ df | TLI | CFI | $\Delta$ CFI | RMSEA | SRMR |
| 9 to 11 years (n=235) | 21.95 | 8 | | | 0.96 | 0.98 | | 0.09 | 0.03 |
| 12 to 13 years (n=912) | 39.39 | 8 | | | 0.97 | 0.98 | | 0.07 | 0.02 |
| 14 to 17 years (n=411) | 18.22 | 8 | | | 0.98 | 0.99 | | 0.06 | 0.02 |
| Configural invariance | 80.97 | 24 | | | 0.97 | 0.99 | | 0.07 | 0.03 |
| Metric invariance | 89.87 | 32 | 4.51 | 8 | 0.98 | 0.99 | 0.00 | 0.06 | 0.03 |
| Scalar invariance | 101.95 | 40 | 10.43 | 8 | 0.98 | 0.98 | 0.00 | 0.06 | 0.03 |

*Note*: MLM-estimator with robust standard errors and Satorra-Bentler scaled test statistic

*Table A4* Measurement Invariance of Ethnic Identity

| *Immigrant Generation (N=878)* | $\chi 2$ | df | $\Delta \chi 2$ | $\Delta$ df | TLI | CFI | $\Delta$ CFI | RMSEA | SRMR |
|---|---|---|---|---|---|---|---|---|---|
| 1st generation (n=58) | 15.14 | 8 | | | 0.52 | 0.75 | | 0.12 | 0.06 |
| 2nd generation (n=712) | 28.72 | 8 | | | 0.93 | 0.96 | | 0.06 | 0.03 |
| 3nd generation (n=108) | 2.42 | 8 | | | 1.00 | 1.05 | | 0.00 | 0.02 |
| Configural invariance | 41.92 | 24 | | | 0.95 | 0.98 | | 0.05 | 0.03 |
| Metric invariance | 49.27 | 32 | 4.85 | 8 | 0.97 | 0.98 | 0.00 | 0.04 | 0.03 |
| Scalar invariance | 58.49 | 40 | 7.08 | 8 | 0.97 | 0.97 | 0.00 | 0.04 | 0.04 |
| *Age (N=946)* | $\chi 2$ | df | $\Delta \chi 2$ | $\Delta$ df | TLI | CFI | $\Delta$ CFI | RMSEA | SRMR |
| 9 to 11 years (n=149) | 8.77 | 8 | | | 0.99 | 0.99 | | 0.03 | 0.03 |
| 12 to 13 years (n=544) | 14.86 | 8 | | | 0.98 | 0.99 | | 0.04 | 0.02 |
| 14 to 17 years (n=253) | 11.34 | 8 | | | 0.97 | 0.99 | | 0.04 | 0.03 |
| Configural invariance | 34.43 | 24 | | | 0.98 | 0.99 | | 0.04 | 0.03 |
| Metric invariance | 42.75 | 32 | 7.81 | 8 | 0.98 | 0.99 | 0.00 | 0.03 | 0.04 |
| Scalar invariance | 52.86 | 40 | 9.64 | 8 | 0.98 | 0.98 | 0.00 | 0.03 | 0.04 |

*Note*: MLM-estimator with robust standard errors and Satorra-Bentler scaled test statistic

methods, data, analyses | Vol. 9(1), 2015, pp. 111-136 DOI: 10.12758/mda.2015.004

# How to Measure Participation of Pupils at School. Analysis of Unfolding Data Based on Hart's Ladder of Participation

*Daniela Wetzelhütter & Johann Bacher*
*Johannes Kepler University (JKU) of Linz*

**Abstract**

A renaissance of the scientific study of participation has fortunately taken place in recent years. As a consequence of the Convention on the Rights of the Child, forms of participation of children were implemented in different fields of society, especially in the community and at the school level. These programs and attempts are described by science. However, little attention has been paid to the measurement of participation. Therefore, this paper proposes and analyses a measurement instrument for the frequently cited Ladder of Participation by Hart (1992). His model assumes that different levels (degrees) of participation exist, whereby the extent increases with each level (stage).

*Keywords*: Multidimensional Scaling, Correspondence Analysis, CATPCA, Unfolding, Ladder of Participation, Degree of Participation

# 1 Introduction

Lewin and Lewin (1982 [1941]) demonstrated in their essay "Democracy and School" that practical experience of democracy is important for learning democratic attitudes. They dealt with the participation in school lessons and pointed out that such participation is potential in primary children. However, the participation task should be defined manageably. According to the authors, democracy in school leads to a more amiable climate in groups and improves individual academic performance.

Childhood participation became the focus of social science research in the 1990s. At that time it was related to the emergence of childhood research (James, Jenks & Prout, 1998; Zinnecker, 1999; Wilk & Bacher, 1994; Bacher, Gerich, Lehner, Straßmair & Wilk, 1999). In this connection, scientists intensively discussed questions such as the relationship between project-based participation and formal democratic structures.

In the past few years (cf. Wetzelhütter, Paseka & Bacher, 2013) the topic has once again been addressed, but this time from the perspective of school development research. One such research focus was the pupil's perspective of involvement or rather participation in school. In this connection, research on participation often refers to the Ladder of Participation developed by Hart (1992) or comparable models from Oser and Biedermann (2006) or Arnstein (2011[1969]). Regarding Hart's construct, the intensity of participation is the core element, where at each "level" the extent of participation of children increases (see Chapter 2). The model suggests a one-dimensional (degree of participation) unfolding scale, whose items are characterized by different intensities of participation. It is assumed that the indicators neither mutually exclude each other (this means that they may occur concurrently) nor mutually depend on each other (this means that reaching a stage does not require the presence of another step – as would be the case for a Guttman scale).

However, Hart makes no clear statement about the dimensional structure of his concept. Therefore, the paper addresses the following question:

- Can participation in school be measured one-dimensionally as assumed by Hart?

In part 2, the paper describes the theoretical background of Hart's Ladder of Participation. The methodical approach (operationalization, databases and data analysis)

*Direct correspondence to*
Daniela Wetzelhütter, Johannes Kepler University (JKU) of Linz, Department of Empirical Social Research, Austria.
E-mail: Daniela.Wetzelhuetter@jku.at

*Acknowledgment*: We would like to thank the two anonymous reviewers for their important suggestions.

is discussed in part 3. The results (descriptive statistics, examination of dimensionality, and validation of the outcome) are presented in part 4. Finally, part 5 provides a summary followed by conclusions.

## 2 Theoretical Framework (Measurement Model): Hart's "Ladder of Participation"

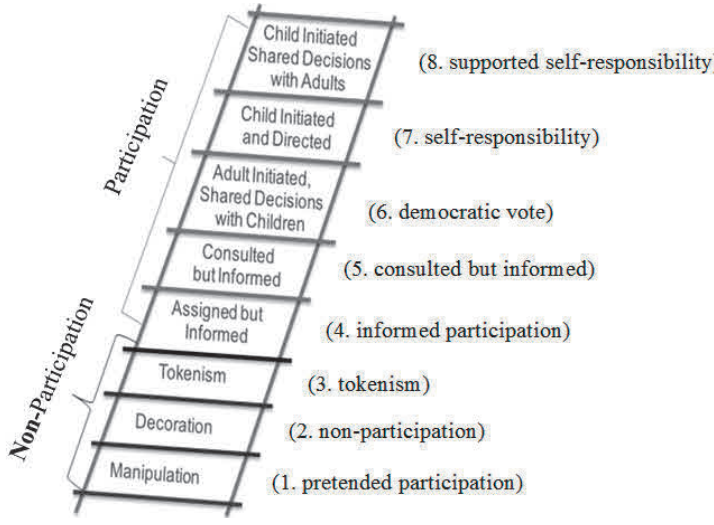
Hart (1992) defines participation as a "process of sharing decisions which affect one's life and the life of the community in which one lives" (p. 5). Therefore, participation is (for him) a *process* of collective decision making. Based on an investigation of participation projects sponsored by UNICEF (programs for street children or prevention programs), Hart (1992) developed an eight-stage Ladder of Participation, based on a model by Arnstein published in 1969 (Hart, 1992). This model generally differentiates between non-participation and participation (Figure 1), though every stage describes a different "process" of children's involvement. Hart further divides non-participation into:

1. "Manipulation" (pretend participation),
2. "Decoration" (children are "used," but participation is not pretend),
3. "Tokenism" (token policy – e.g. when children are given a voice, but are badly/not prepared for the topic on which they are voting).

Participation begins with the fourth stage – the intensity is divided into:

4. "Assigned but Informed" ("just" informed participation of children),
5. "Consulted but Informed" (informed participation, where the opinions of children are treated seriously),
6. "Adult Initiated, Shared Decisions with Children" (adults initiate projects but the decision making is shared by children and adults),
7. "Child Initiated and Directed" (initiated and directed by children), and
8. "Child Initiated, Shared Decisions with Adults" (projects which are initiated by children, but the decision making is shared by children and adults).

"Real participation" is placed at the top position. The lower the degree of participation, the lower the scope and the involvement. At this point it has to be mentioned that the ladder metaphor is an unfortunate choice, since it could be assumed that reaching the top step implies climbing the previous stages. This would imply that the scale is constructed in the logic of a Guttman scale. As Hart pointed out in a later paper (2008), despite an increasing degree of participation for each level, the ladder does not stand for a developmental process: "In fact the ladder is primarily about the degree to which adults and institutions afford or enable children to participate" (p. 23). This means each stage measures (as outlined above) the degree of




According to: Hart (1992, p. 8) – subsequently used definitions in brackets

*Figure 1* Hart's Ladder of Participation

participation. Depending on the project/topic and participant, the level may be of different degrees of involvement. For instance, a project may directly start at level 7 ("Child Initiated and Directed") without passing the previous stages of 1 to 6.

This paper examines the question of whether the degree of participation can be measured one-dimensionally using a simple battery of eight questions, outlined in more detail in the next chapter.

Scientifically, such an instrument would be important to develop and test theories concerning the influence of participation on the individual, class and school level. According to Lewin (e.g. Lewin, Lippitt & White, 1939) it would be interesting to analyze if participation results in higher abilities and more effectiveness as well as in less dissocial (aggressive) behavior of school children.

# 3 Methodical Approach

## 3.1 Operationalization: Degree of Participation

In the "(Do) students develop school!?" project (Altrichter, Bacher, Langer, Gamsjäger & Wetzelhütter, 2012), a scale was developed to measure the degree of participation in a school as indented by Hart's Ladder of Participation. The final version – a rating in the form of a five-point Likert scale – can be seen in Table 1.

*Table 1* Rating scale to measure participation of students in school

| **What applies to your school and how much?** | fully applies | largely applies | uncertain | does not apply much | does not apply at all | concrete example in a school |
|---|---|---|---|---|---|---|
| We may propose our own projects and implement them independently. | Level 7 "self-responsibility" (V7) | | | | | Students propose to publish an online school newspaper; they can use the infrastructure and produce the newspaper alone |
| We may propose our own projects and teachers/directors help us with the implementation. | Level 8 "supported self-responsibility" (V8) | | | | | As above, but teachers and the director support them |

*Table 1 continued*

| **How are choices made at your school?** | fully applies | largely applies | uncertain | does not apply much | does not apply at all | concrete example in a school |
|---|---|---|---|---|---|---|
| Although we are asked about our opinion on important decisions, it will not be considered. | Level 1 "pretend participation" (V1) | | | | | The school board wants to renovate the school garden. The students are asked for their ideas, but these ideas are ignored in the final decision. |
| We don't get enough information about important decisions, but we are allowed to vote. | Level 3 "tokenism" (V3) | | | | | The students were informed that two different textbooks are available and that they should decide which of them they want to use. No further information was given to them. |
| We get informed about important decisions but we are not allowed to vote. | Level 4 "informed" (V4) | | | | | The school informed their students that the school garden will be renovated. |
| We get informed about important decisions and our opinion is asked for. | Level 5 "consulted but informed" (V5) | | | | | The school informed their students that the school will renovate the school garden and asked the students for their ideas. |
| When it comes to important decisions, several options are presented to us that we then vote on. | Level 6 "democratic vote" (V6) | | | | | The school intends to renovate the school garden. Different concepts were presented to the students. Each student could vote for one. |
| Important decisions are made without our involvement. | Level 2 "non-participation" (V2) | | | | | The school got new computers in the holidays. The students were not asked. |

*Note*. Translated version based on the original German version.

Each item (every statement) represents one level of Hart's ladder. For instance, the statement "When it comes to important decisions, several options are presented to us that we then vote on" corresponds to stage 6 (democratic vote). The five-point Likert scale provided the following potential responses: "fully applies," "largely applies," "uncertain," "does not apply much," and "does not apply at all."

This scale belongs to the group of subject-oriented scaling procedures (see Likert, 1932). In our case, the subjects are schools and school classes. The scale consists of eight items (objects) which may be ranked on one dimension (degree of participation) at which each subject has an ideal point (perceived degree of participation) as suggested in Coombs' (1964) unfolding model (for details see Section 3.3).

## 3.2 Data Sets and Sampling Design

The data analysis is based on two data sets (see Table 2). The first data set was generated as a pre-test study for calibration (calibration sample) and the second as a nationwide sample for validation (validation sample). The sample is representative for vocational colleges ("Berufsbildende mittlere und höhere Schulen")[1] and upper level grammar schools ("Allgemeinbildende höhere Schulen, Oberstufe") with respect to sex and language spoken at home. In both cases, the selection of the pupils followed a stratified cluster sampling (see for example Bacher, 2009 or Sturgis, 2004). During the first stage, of the calibration sample, entire classes were selected per school grade. This procedure resulted in 22 (64.7%) out of 34 school classes, with n=382 (=86.8% return rate) surveyed pupils out of 440 selected pupils.

During the first stage of the validation sample, 282 schools (=36.2%, including 65 alternatively drawn schools) of several types of schools[2] were selected in each province of Austria (see Table 2). Forty-five percent of these schools were willing to participate (n=127). Within these schools, two classes (altogether 254 classes) were randomly defined based on the Kish selection grid (see Kish 1965). This approach was chosen to avoid a conscious choice by headmasters, for example of "laptop classes" or classes with a specialization in IT topics. Ultimately, 127 Austrian schools (=45% of the gross sample) with 253 classes (=99.6% of the selected classes) and 4,101 pupils (=67.8% of the drawn pupils) took part from all over the country. Due to the fact that participation in the survey was acquired exclusively via telephone and conducted online, the return rate is classified as good.

1 The OeAD (2014) provides information about the Austrian educational system.

2 AHS (upper level of grammar schools) and BMHS (vocational colleges) including A&F (schools and colleges for agriculture and forestry) and BAKIP (training institutions for kindergarten teachers), see OeAD (2014)

*Table 2* Total population with drawn and realized samples

| Sample | Stage | Total Population N | Drawn Sample | | Realized Sample | |
|---|---|---|---|---|---|---|
| | | | N | %[a] | n | %[b] |
| Calibration Sample | 1 = classes | 34 | 22 | 64.7 | 22 | 100.0 |
| | 2 = pupils | 698 | 440 | 63.0 | 382 | 86.8 |
| Validation Sample | 1 = schools | 779 | 282 | 36.2 | 127 | 45.0 |
| | 2 = classes | 1577[c] | 254 | 16.1 | 253 | 99.6 |
| | 3 = pupils | 37494[d] | 6045[e] | 16.1 | 4101 | 67.8 |

*Note*. (a) Proportion of the drawn sample in terms of the basic population.
(b) Proportion of the realized sample in terms of the drawn sample.
(c) Total population of classes of the drawn sample of schools (secondary education level).
(d) Total population of pupils of the drawn sample of classes (secondary education level).
(e) Estimated number of pupils based on the average number of pupils per class of the individual school.

## 3.3 Data Analysis

### 3.3.1 Model Specification

As mentioned, the Hart ladder does not constitute a Guttman scale. To reach a certain degree of participation in a school, it is not necessary to pass stages with lower levels of participation. A school may immediately start, for example, with informed participation (stage 4) without passing stages 1 (pretend participation) to 3 (tokenism). According to Coombs (1964), the suitable scaling procedure for our scale is an "unfolding model." Figure 2 illustrates the idea.

This model presumes that one dimension is measured at which the items are sorted according to one criterion and every person (stimulus scaling) or every object (object scaling) has one ideal point on the scale. The ideal point (cf. for example de Leeuw, 2005) represents the point of maximum preferential choice of one person or the point that best represents the "required" object.

In the present study, the appropriate model may be realized as follows. In every school or in each class a certain degree of participation (ideal point) exists, which is judged by the surveyed pupils. The level that comes closest to this (perceived) degree receives the highest agreement, the level that is second closest the second largest agreement, and so on. Based on these individual rank orders (so the assumption in regard to Hart's ladder), a common solution (order of the participation items) can be determined. This solution reflects the extent of participation

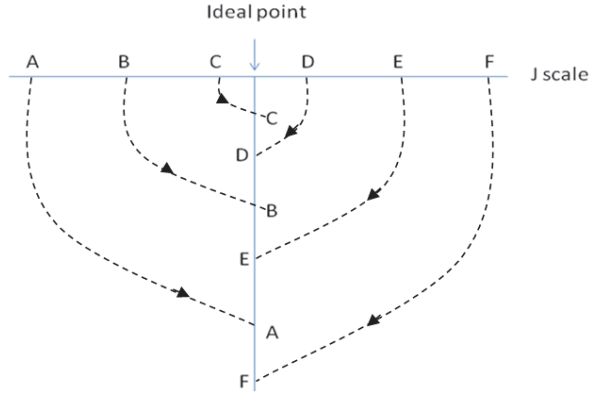



*Figure 2* Unfolding model according to Coombs (1964, p. 80)

along one dimension. Practically, an unfolding analysis can be performed within multidimensional scaling (MDS) procedures. Taking these considerations into account, and referring to the narrow definition of multidimensional scaling of Cox and Cox (2001), it is intended to locate the items within a one-dimensional or at least a low-dimensional space that best matches the original (dis)similarities and distances between the items (objects). For the present paper, PROXSCAL (Meulman & Heiser, 2011) was used to perform the multidimensional scaling[3]. This procedure enables the user to generate the distances and dissimilarities (cf. Borg, Groenen & Mair, 2013) of the items for several groups as a so-called "Three-Way Model" (cf. for example Kruskal & Wish, 1978). This means the analyses are based on distance matrices generated per "school class" (calibration sample) or per "school" (validation sample). For PROXSCAL (Meulman & Heiser, 2011; IBM-SPSS, 2011b; Borg et al., 2013), the following specification was chosen:

- *Proximities*: Euclidean distance
- *Starting configuration*: classical (Torgerson)
- *Condition*: matrix (distances are exclusively comparable within one matrix)
- *Transformation*: ordinal (ties: keep ties)
- *Model*: Generalized Euclidean: each school or class has an individual space, which is a rotation of the common space with subsequent weighting of the dimensions

3 For this analysis cases were excluded with missing values and/or without variation in the responses.

The specification follows the recommendation in the literature (IBM-SPSS, 2011b; Borg & Groenen, 2005; Borg et al., 2013). We selected ordinal transformation for ordinal data and kept ties for formal reasons, because a rating scale with a small number of levels may result in the same proximity values for some items (Borg et al., 2013). We chose Torgerson as the starting configuration because "this option tends to give better quality solutions" (Borg & Groenen, 2005) and defined "matrix" as the condition for indicating that a separate transformation has to be found for each matrix (ibid). We chose with "Generalized Euclidean" a dimension-weighting model, to account for individual differences (see for example Borg & Groenen, 2005) between classes (calibration sample) and schools (validation sample) as implied by unfolding data.

However, the specifications are arbitrary to a certain degree as is the case for all applications of MDS. Therefore, we used correspondence analysis, specifically CATPCA (categorical principal component analysis, see Meulman & Heiser, 2011), as an additional method (see below). The specification of an ordinal transformation implies a non-metric MDS. Non-metric MDS requires a specific number of objects (in our case items) for each dimension (Bacher, Pöge & Wenzig, 2010). In this regard, the ratio of 5:1 (items : dimensions) for a complete dissimilarity matrix (ibid) is recommended, or Kruskal and Wish's (1978) rule of thumb of at least more than fourfold the number of items per dimension.

Due to the problem of choosing an appropriate specification for MDS and due to the small number (eight) of analyzed items, the dimensionality is also tested by a correspondence analysis (CATPCA). This procedure tolerates a smaller number of items. Two items per dimension are sufficient. In addition, the user does not have to make specifications concerning starting configuration, distance measures, transformation, handling of ties etc. All these decisions follow from the statistical model of correspondence analysis (of course with the disadvantage of being less flexible).

The correspondence analysis is similar to MDS (e.g. Borg & Groenen, 2005; Greenacre, 2013). It enables the interpretation of distances (Chi² and weighted Euclidean distances are calculated) between variables (in this context items) as well as objects (in this context pupils). In the literature, the similarity of these two methods is mentioned frequently (cf. Hoffman & de Leeuw 1992; Cherkassky & Mulier, 2007; Young, 1985). The multiple correspondence analysis is applied in equivalent situations to the MDS (Meulman, van der Kooij & Heiser 2004) or in combination with MDS (e.g. Green, 2010).

In this paper, CATPCA was performed. This procedure (cf. Blasius & Thiessen, 2012) is an appropriate technique for analyzing ordered categorical variables, and was performed with the option of optimizing the relationship between variables (Normalization: VPRINCIPAL). The level of analysis was pupils; the order of the variables was interpretable and the representation comparable with the MDS solutions. The data analysis was performed with IBM SPSS Version 22 (IBM-SPSS,

2011b). Applying factor analysis would not be appropriate from a theoretical point of view because a bipolar concept often results in two interpretable but unrelated factors (see van Schuur & Kiers, 1994) which would not correspond to the assumed theoretical concept (see Chapter 2). Nonetheless, explorative factor analysis can be applied to our data.[4]

### 3.3.2 Analyzing Procedure

Data analysis was being performed in three steps:

*Firstly*: In order to test whether our measurement instrument of Hart's ladder is one-dimensional, one- and two-dimensional solutions were computed based on the calibration sample, using MDS (PROXSCAL) and CATPCA.

For the decision regarding which configuration fits better, the following criteria were applied (for a justification, see below):

- MDS: stress level: Stress 1 <0.2; Stress 2<0.4; congruence coefficient: >0.9
- CATPCA: Eigenvalue >1 and Cronbach's alpha >0.7
- MDS and CATPCA: substantive interpretation

Concerning the stress level, it has to be mentioned that Borg et al. (2013) already emphasized that "evaluating a given Stress value is a complex matter" (p. 23). This means that several considerations have to be made. Stress decreases, for instance, with the number of dimensions but increases, for instance, with the number of items (points/objects) or proportion of error components (noise) in the data (Borg et al., 2013). Therefore, the frequently cited benchmark of Kruskal (1964) or Kruskal and Wish (1978), which states that Stress 1 has to be lower than 0.2, and the benchmark of Fahrmeir et al. quoted in Gediga (1998), which says that Stress 2 should be lower than 0.4, are both applied conditionally. The congruence coefficient (which, in regard to Borg et al., (2013), can be interpreted as a correlation coefficient) should be close to the value of 1. Concerning CATPCA, it can be noted that the eigenvalue(s) should be considerably higher than 1 (according to Kaiser's criterion, cf. for example Hardy & Bryman, 2009) and Cronbach's alpha should be higher than 0.7 (cf. for example Nunnally, 1978). Additionally, the correlation of the distance matrices (see CPCC below) obtained by the MDS and CATPCA, based on the one- and two-dimensional solution, is calculated in order to verify the findings.

4 In our case, two factors (Eigenvalue >1) can be extracted for each sample (validation and calibration). Based on the validation sample, variables representing the first four steps of the ladder (V1–V4) load on the first factor, the next two variables (V5, V6) on both factors, and the last two (V7, V8) on the second factor. The outcome for the calibration sample is slightly different. V3 loads on both factors, while V5 loads on the first instead of both factors. Neither result corresponds to the theoretical concept (Chapter 2). The result partially reproduces the grouping of items in three clusters (see Chapter 4).

Lastly, substantive interpretation will be considered, which means that a solution is supported if the arrangement of the items is explicable.

*Secondly*: The validity of the scale was examined. Initially it was tested to which extent the results of the calibration sample are replicated in the Austria-wide validation sample. Afterwards, the cophenetic correlation coefficient (CPCC) was computed in order to judge how well the validation sample reproduces the calibration sample (see Romesberg, 2004). The CPCC is defined as (linear correlation coefficient):

$$CPCC = \frac{\sum_{i<j}(d_x(i,j)-\bar{d}_x)(\mathrm{d}_y(i,j)-\bar{d}_y)}{\sqrt{\left[\sum_{i<j}(d_x(i,j)-\bar{d}_x)^2\right]\left[\sum_{i<j}(d_y(i,j)-\bar{d}_y)^2\right]}}$$

$d_x(i,j)$ = Euclidean distance between the i$^{th}$ and j$^{th}$ items of the calibration sample
$d_y(i,j)$ = Euclidean distance between the i$^{th}$ and j$^{th}$ items of the validation sample
$\bar{d}_x$ = average of $(d_x(i,j))$; $\bar{d}_y$ = average of $(d_y(i,j)$

The CPCC measures the similarity between two dissimilarity matrices – in our case between the computed dissimilarity matrix for the calibration and the computed dissimilarity matrix for the nationwide validation sample. Both matrices were calculated based on representation in a one- or two-dimensional space. According to Romesberg (2004), a value from 0.8 upwards is acceptable, since "the distortion is not great" (p. 27).

*Thirdly*: Two content variables were included in the scaling process. Validity was examined by two criteria: the location of the items in the dimensional space and the stability of the obtained scaling model by comparison of the model with and without content variables. The content variables are "co-decision to decorate the classroom" and "co-decision of school rules." Regarding decisions about rules at school, they are mostly made by laws, teachers, or school heads. Merging the results of the mentioned sparkling-science project (Altrichter et al., 2012) shows that on unique occasions or with informal participation possibilities, pupils often do not get enough background information or time to develop/determine their position. Therefore, it is reasonably assumed, if pupils are involved in decision making, that they become informed just before the vote is carried out in order to simplify the procedure. Accordingly, the hypothesis is:

H1: The item "co-decision of school rules" is located next to the level "tokenism" (level 3, V3), and the obtained structure (without the criteria variable) doesn't change.

Furthermore, the sparkling-science project (Altrichter et al., 2012) showed that participation most likely takes place in "peripheral" issues. For instance, the "co-

decision to decorate the classroom" is usually based on a joint decision of teachers and pupils. Accordingly, the hypothesis is:

H2: The item "co-decision to decorate the classroom" is located next to the item "consulted but informed" (level 5, V5), and the obtained structure of the model (without the criteria variable) doesn't change.

The results of the described data analysis will now be presented in Chapter 4, with descriptive results provided in the appendix.

# 4 Results

## 4.1 Dimensionality of the scale in the calibration sample

In order to determine the dimensionality of the scale, one- and two-dimensional solutions were computed by MDS (PROXSCAL) and by CATPCA. Table 3 summarizes the results of the calibration sample.

The one-dimensional MDS solution doesn't fulfill the defined stress criteria of 0.20 (Stress 1) or 0.40 (Stress 2), which indicates a serious proportion of noise (errors) in the data. As expected, a two-dimensional model decreases the stress level, but the stress criteria are not met. However, stress 1 is only slightly higher than the threshold of 0.20 and in both cases the congruence coefficient is close to a perfect solution (value of 1). The results of CATPCA support a one-dimensional solution more clearly. The second eigenvalue is sufficiently smaller than the first eigenvalue. The consistency of the second dimension is poor (Cronbach's alpha=0.206), whereas the Cronbach's alpha of the first dimension is 0.773. Furthermore, the correlation between the distances (r=0.930) obtained by MDS (PROXSCAL) and CATPCA also supports the one-dimensional solution. The poor correlation (r=0.501) based on the two-dimensional models indicates that the respective solutions are less stable and robust. They depend on the applied method (see also Figure 3 A–D).

From a substantive point of view, both MDS solutions are acceptable (see Figure 3 A and C). In either case, the items can be divided into two groups of items at the zero point of the x-axis, which may be interpreted as the scope of action. This result is similar to Hart's outline – only item V4 is allocated in the "wrong group." The left group can be named "effective/adequate participation" and the right group may be identified as "ineffective/inadequate participation." However, the two-dimensional solution is not interpretable as a "ladder of participation." Based on Figure 3C, the scope of action is further distinguishable due to the initiative of involvement in the dependence of students (upper half) vs. teachers (lower half). Nonetheless, the rank order of the items (reading from left to right)

*Table 3* Summarized results for determining dimensionality using PROXSCAL and CATPCA (calibration sample)

| | PROXSCAL | | | CATPCA | | |
|---|---|---|---|---|---|---|
| Model | Stress 1 | Stress 2 | Congruence coeff. | Eigenvalue (Variance) | Cronbach's alpha[(b)] | CPCC[(a)] |
| One-dim. solution | 0.316 | 0.616 | 0.95 | 3.1 (38.65%) | **0.773** | **0.930** (p=0.000) |
| Two-dim. solution | **0.203** | 0.556 | 0.98 | 1.2 (15.26%) | 0.206 | 0.501 (p=0.007) |

*Note*. (a) n=28 distances (=8 times 7/2); CPCC between distance matrices obtained by the coordinates produced by PROXSCAL and CATPCA. A correlation of 1 would imply that the distances between items computed by PROXSCAL perfectly fit (in a linear sense) with distances between items computed by CATPCA. For the one-dimensional solution, the value of the CPCC is near 0.930, whereas the CPCC is 0.501 for the two-dimensional solution. Hence, the two-dimensional solution is less robust, whereas the one-dimensional solution is more stable and robust.

(b) Cronbach's alpha (see Cronbach, 1951) is based on all items. Pupils are the level of analysis. Cronbach's alpha is defined as (see Meulman et al., 2004):

$\alpha = \frac{m}{m-1} \cdot \frac{\sum \lambda - 1}{\sum \lambda}$ : where the number of categories is *m* and the eigenvalues are $\lambda^2$

deviates in both cases from the order that Hart recommends. The levels "consulted but informed" (V5) and "democratic vote" (V6) depict the highest degree of participation, while (supported) self-determination (V7 and V8) shows a slightly lower degree. In addition, the level "pretend participation" (V1) should depict the lowest degree. This result contradicts the theoretical concept but is comprehensible in terms of "involvement" (as synonym for participation). For instance pupils are at least involved even if participation is pretend (V1) in contrast to simply being informed (V4) or not participating at all (V2).

The one-dimensional solution of CATPCA (Figure 3B) refers to the existence of three clearly identifiable groups (see below): "sufficient participation" (V5, V6, V7, V8), "symbolic participation" (V3, pupils can vote, but are not prepared or well informed, it can be assumed that their votes are ignored, too) and "deficient participation" (V1, V2, V4). The items of each group are located close together. This explains the poor fit of MDS (PROXSCAL), which seems to require that the objects are distributed in the whole space. If objects are located close together, this results in errors in MDS, whereas this fact improves the results of CATPCA. CATPCA can be seen as a factor analysis for categorical scaled items. Two or more measures with the same properties (the same location in the space) improve the scaling.

*Figure 3A*

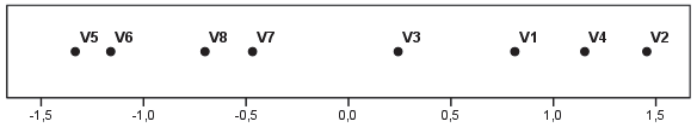



*Figure 3B*

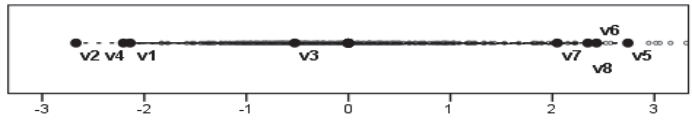



*Figure 3C*

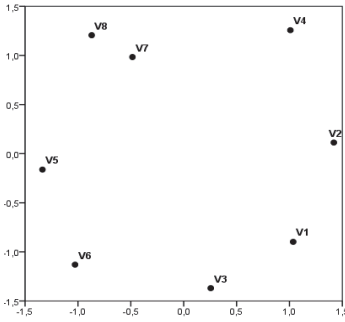
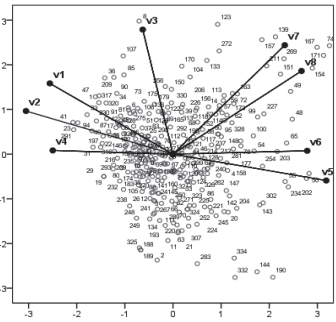



*Figure 3D*




*Note*. V1 = pretend participation, V2 = non-participation, V3 = tokenism, V4 = informed, V5 = consulted but informed, V6 = democratic vote, V7 = self-responsibility, V8 = supported self-responsibility

*Figure 3 A–D* One- and two-dimensional solutions (PROXSCAL (3A and 3C) vs. CATPCA (3B and 3D)) – calibration sample

A further reason for the poor stress is the fact that PROXSCAL uses (in this case) individual distance measures, whereas CATPCA is applied for an aggregated similarity matrix. Hence, individual differences influence the overall fit of MDS less. The substantive reason for these differences is presumably personalized participation (see Wetzelhütter et al., 2013), which depends on different perceptions of participation by pupils.

Apart from that, the (more or less ambiguous) label "uncertain" of the mid response category of the five-option rating scale might have been ambiguous for the respondents and might have caused errors, resulting in measuring a different dimension (see for example Rost, Carstensen & von Davier, 1999).

As an interim conclusion, it can be noted that the results, especially those of CATPCA, support the one-dimensional model rather than the two-dimensional. Arguments for this solution are:

- The one-dimensional model is replicable by different analytical methods (MDS/PROXSCAL and CATPCA) as can be seen, for example, by a high CPCC (which is not the case for the two-dimensional solution).
- The eigenvalue and Cronbach's alpha for the first dimension are acceptable (which is not the case for the second dimension).
- The one-dimensional solution is, in accordance with Hart's concept of a "ladder of participation," substantively interpretable as a degree of participation. Hence, the one-dimensional solution has a theoretical basis.

However, the assumed eight different levels of participation cannot be differentiated. The findings suggest that it is only possible to distinguish three groups of participation.

## 4.2 Model Validation: Reproduction of Dimensional Space in the Validation Sample

The model validation is carried out by two approaches. First, it is tested as to what extent the above-described solutions are replicable using the Austria-wide data. In a second and final step, relevant content variables are added as criteria variables in order to test content related hypotheses (see below).

### 4.2.1 Replication of the Model

As before, one- and two-dimensional solutions were computed using MDS (PROXSCAL) and CATPCA in order to determine the dimensionality of the developed scale. Table 4 shows that the results of the validation sample confirm the previous outcome of the calibration sample (cf. Table 3). Moreover, the stress values of the MDS are smaller compared with the calibration sample. Stress 1 fulfills the threshold in the two-dimensional case but Stress 2 still has a value higher than the threshold. The fact that the stress values are higher than the threshold indicates that the MDS approach has problems. Again the reason is that there are groups of items and the items of each group are located close together, which results in errors. CATPCA supports the one-dimensional solution. Once more, the CPCC (CPCC r=0.910) obtained by MDS (PROXSCAL) and CATPCA also supports the one-dimensional solution. Hence, the one-dimensional solution is more robust against the method applied.

From a substantive point of view (see Figure 4A-D), as before, the one- and two-dimensional MDS solutions show a possible distinction at the continuum of the scope of action (from left to right on the x-axis). Therefore, "effective/adequate participation" (left) and "ineffective/inadequate participation" (right) may, regarding the two-dimensional model, be further distinguishable regarding the initiative of involvement (y-axis). The two-dimensional model differs from the earlier one

*Table 4* Summarized results for determining dimensionality using PROXSCAL and CATPCA (validation sample)

| | PROXSCAL | | | CATPCA | | |
|---|---|---|---|---|---|---|
| Model | Stress 1 | Stress 2 | Congruence coeff. | Eigenvalue (Variance) | Cronbach's alpha[(b)] | CPCC[(a)] |
| One-dim. solution | 0.289 | 0.557 | 0.96 | 3.2 (40.1%) | **0.787** | **0.910** (p=0.000) |
| Two-dim. solution | **0.174** | 0.440 | 0.98 | 1.3 (16.07%) | 0.254 | 0.606 (p=0.001) |

*Note*. (a) n=28; Correlation between distance matrices obtained by the coordinates produced by PROXSCAL and CATPCA. (b) Cronbach's alpha is based on all items (formula for calculation, see Table 3). Pupils are the level of analysis.

*Figure 4A*

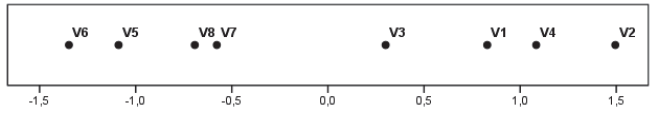



*Figure 4B*

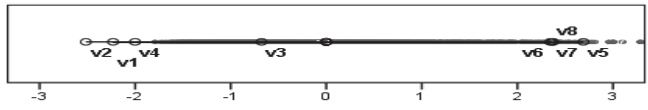



*Figure 4C*

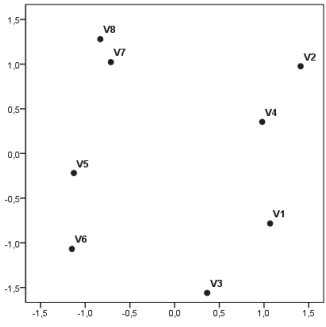
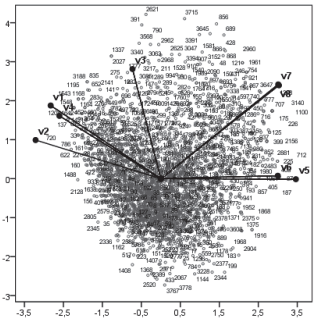



*Figure 4D*




*Note*. V1 = pretend participation, V2 = non-participation, V3 = tokenism, V4 = informed, V5 = consulted but informed, V6 = democratic vote, V7 = self-responsibility, V8 = supported self-responsibility

*Figure 4 A–D* one- and two-dimensional solutions (PROXSCAL (4A and 4B) vs. CATPCA (4C and 4D)) measuring the degree of participation (validation sample)

*Table 5* CPCC for one- and two-dimensional distance matrices

| Calibration with Validation Sample | | | | |
|---|---|---|---|---|
| Procedure | Dimensionality | CPCC[(a)] | p | n |
| PROXSCAL | One-dimensional solution | **0.978** | 0.000 | 28 |
| | Two-dimensional solution | 0.838 | 0.000 | 28 |
| CATPCA | One-dimensional solution | 0.995 | 0.000 | 28 |
| | Two-dimensional solution | **0.945** | 0.000 | 28 |

*Note*. (a) Correlation between distance matrices of the calibration sample with the validation sample obtained by the coordinates produced by PROXSCAL and CATPCA

presented in two points. Firstly, the items v5 and v6 can hardly be differentiated concerning the scope of action. Secondly, the items v4 and v2 are switched.

Once again, CATPCA differentiates between three groups: "sufficient participation" (V5, V6, V7, V8), "symbolic participation" (V3), and "deficient participation" (V1, V2, V4).

As described in Section 3.2.2, the CPCC was computed in order to measure the similarity between the calibration and the validation sample. In our case, distance matrices were calculated for each sample (calibration and validation) based on the representation in a one- and two-dimensional space. Table 5 shows that the distance matrix of the validation sample represents the similarity structure of the calibration sample almost perfectly (MDS: CPCC=0.978; CATPCA: CPCC=0.995) in the one-dimensional case. For the two-dimensional case, the CPCC is good (MDS: CPCC=0.838; CATPCA: CPCC=0.945), but the coefficients are smaller. Consequently, this result also reinforces the one-dimensional solution.

To summarize, the one-dimensional solution is supported by:

- Substantive interpretation: the results suggest that three levels of participation can be distinguished
- Model robustness: the one-dimensional MDS (PROXSCAL) solution is replicated by a different analytical method (CATPCA)
- Model replication: the one-dimensional solution can be reproduced with the validation sample. The CPCC is nearly perfect.

Therefore, the outstanding analyses are based on the one-dimensional solution.

### 4.2.2 Criteria Validation

In a second and final step, relevant content variables are added as criteria variables. It is assumed that they are connected with specific levels of participation. It is expected that the predictors will be positioned next to those items which are influenced or rather determined by their occurrence. Concurrently, the basic (one-dimensional) model should not significantly change due to this strategy.

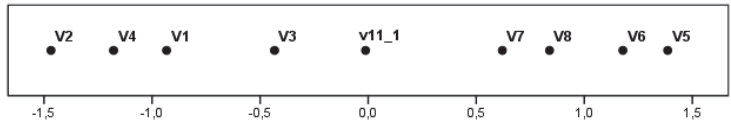
Figure 5 A–B shows the results for the first hypothesis. As expected, co-decision of school rules (v11_1) is located next to "Tokenism," which means that pupils are allowed to vote but are badly prepared concerning the topic. This may be caused for example by the aim to simplify such procedures. The structure remains basically unchanged, so the grouping of the items in the three groups is not affected.

In accordance with the second hypothesis, "co-decision to decorate the classroom" (V11_2) is placed next to the item which implies an (nearly) "equal" involvement in decision making (V5). In fact, the item V11_2 is, independently of the position of V5, placed next to it. This result reinforces the stability of the solution (see Figure 6 A–B). Only the items V4 and V1 are interchanged in the validation sample – the structure remains basically unchanged. The grouping of the items in the three groups is not affected.

In addition, we computed the mean of the two criteria variables in dependence of the scale values (object quantifiers, obtained by CATPCA) of the pupils. If the scale measures different degrees of participation, a higher scale value should correspond to a stronger agreement in the presence of the two criteria variables at school. The results are shown in Table 6. For significance testing, the multilevel structure was taken into account using the SPSS module MIXED (IBM-SPSS, 2011a). The results confirm our interpretation. A higher degree of participation on the scale corresponds to a more positive evaluation of presence of the two criteria.

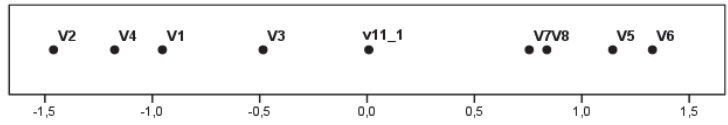
To sum up, the structure of the model remained basically unchanged regardless of the used data set. The above-mentioned hypotheses were confirmed, as the included criteria variables are placed as expected next to related items. In addition, a clear correspondence exists between the derived scale values and the criteria variables.

*Figure 5A*




*Note*. Stress 1: 0.40; Stress 2: 0.76; congruence coefficient: 0.92

*Figure 5B*

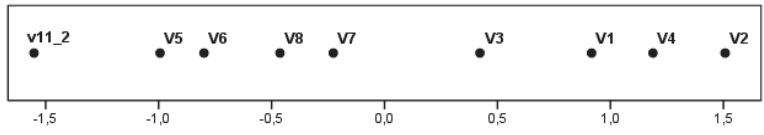



*Note*. Stress 1: 0.38; Stress 2: 0.73; congruence coefficient: 0.92

v11_1 = co-decision of school rules, V1 = pretend participation, V2 = non-participation, V3 = tokenism, V4 = informed, V5 = consulted but informed, V6 = democratic vote, V7 = self-responsibility, V8 = supported self-responsibility

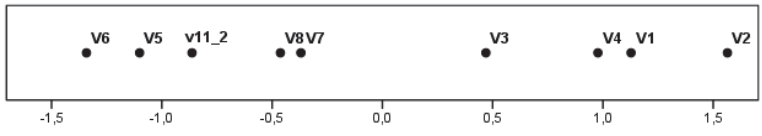
*Figure 5 A–B* One-dimensional solution (PROXSCAL) in connection with context variable I (calibration sample (5A) vs. validation sample (5B))

*Figure 6A*




*Note*. Stress 1: 0.34; Stress 2: 0.65; congruence coefficient: 0.94

*Figure 6B*




*Note*. Stress 1: 0.34; Stress 2: 0.64; congruence coefficient: 0.94

V11_2 = co-decision to decorate the classroom, V1 = pretend participation, V2 = non-participation, V3 = tokenism, V4 = informed, V5 = consulted but informed, V6 = democratic vote, V7 = self-responsibility, V8 = supported self-responsibility

*Figure 6 A–B* One-dimensional solution (PROXSCAL) in connection with context variable II (calibration sample (6A) vs. validation sample (6B))

*Table 6* Means of criteria variables in dependence of scale values of pupils (obtained by CATPCA)

| | V11_1 | | V11_2 | |
|---|---|---|---|---|
| Degree of participation | Mean | p | Mean | p |
| low (scale value below -1) | 4.80 (n=604) | Reference Group | 3.44 (n=604) | Reference Group |
| middle (scale value between -1 and 1) | 4.42 (n=2539) | 0.000 | 2.85 (n=2539) | 0.000 |
| high (scale value larger than 1) | 4.04 (n=606) | 0.000 | 2.37 (n=606) | 0.000 |

*Note*. V11_1 = co-decision of school rules, V11_2 = co-decision to decorate the classroom, Scale values from 1 = always (strongly present) to 5 = never (not present), Significance from using multilevel model (MIXED), random effects for schools were assumed

# 5 Summary and Conclusions

In the past few years, a stronger scientific focus on childhood, youth, and educational research on participation has taken place, and Hart's Ladder of Participation (1992, 2008) is frequently used in this context. Hart distinguishes eight levels of participation, each of which is characterized by a different degree of involvement and self-responsibility of children. No developmental process where a previous level must be successfully passed before the next level can be approached is assumed. From a methodological point of view, Hart assumes a one-dimensional unfolding model.

This paper addresses the question of whether participation in school can be measured one-dimensionally, as assumed by Hart. In order to answer the question, a rating scale was developed measuring those eight steps of the ladder from a low (pretend participation) to a high (supported self-responsibility) degree of participation and implemented in an Austrian study on school participation. MDS procedure (PROXSCAL) and CATPCA were applied for data analysis.

The results suggest that participation in school can be measured with a one-dimensional scale in accordance with Hart. The one-dimensional solution reveals higher robustness towards different data analysis and higher consistency as measured by Cronbach's Alpha than a two-dimensional solution. In addition, the one-dimensional model is reproduced with the validation sample, and remained unchanged when validation criteria are included and tested hypotheses with criteria variables are confirmed.

However, in contrast to Hart (1992) we have not been successful in distinguishing eight different levels of participation. The results reveal three groups of participation: "sufficient participation," "symbolic participation," and "deficient participation."

Two primary reasons may be considered as explanations for this outcome. Firstly, the result may reflect the fact that democracy in school is less evident than assumed (see for example Wetzelhütter et al., 2013), so that only three groups can be distinguished. Secondly, the measurement instrument may have been deficient and too insensitive, not able to differentiate the assumed different degrees of participation.

In addition, the content validity of several items can be placed into question. Hart's ladder (2008), for example, positions "pretend participation" at the bottom (Level 1) and "supported self-responsibility" at the top (Level 8). Empirically, "pretend participation" (Level 1) is placed above, for example, "informed participation" (Level 4) and "non-participation" (Level 2), while "supported self-responsibility" (Level 8) is placed below, for example, "democratic vote" (Level 6). This suggests that, from a pupil's point of view, any form of participation, even pretense (Level 1), implies a higher degree of participation than just being informed (Level 4) or not participating at all (Level 2).

Hence, a first conclusion of the study is to try to improve the instrument in order to be able to distinguish between the two mentioned explanations (only three groups of participation are seen by school children versus the instrument lacks sensitivity).

With reference to data analysis, CATPCA (categorical principal component analysis) seems to be more appropriate than MDS. In addition, the former has the advantage that the user must specify fewer parameters to which no clear guidelines exist.

Despite the above-described weakness of the proposed instrument, it can already be used in order to test theories about the influence of participation in school on the individual, class, and school level and to improve school practice in order to enable the productive development of schoolchildren.

# References

Arnstein, S. (2011[1969]). A ladder of citizen participation. In R. T. LeGates & F. Stout (Eds.), *The City Reader*. Fifth edition (pp. 238-250). New York, NY: Routledge.

Altrichter, H., Bacher, J., Langer, R., Gamsjäger, M. & Wetzelhütter, D. (2012). *Schüler/innen entwickeln Schule?! – Endbericht*. Linz: Johannes Kepler Universität.

Bacher, J. (2009). Analyse komplexer Stichproben. In M. Weichbold, J. Bacher, J. & C. Wolf (Eds.), *Umfrageforschung. Herausforderungen und Grenzen*. Österreichische Zeitschrift für Soziologie. Sonderheft 9/2009 (pp. 253-272). Wiesbaden: VS Verlag.

Bacher, J., Gerich, J., Lehner, R., Straßmair, M. & Wilk, L. (1999). Auswirkungen kommunaler Jugendbeteiligungsprojekte. In C. Friesl, M. Hahn, B. Heinzlmaier & C. Klein (Eds.), *Erlebniswelten + Gestaltungsräume* (pp. 213-238). Graz & Wien: Verlag Zeitpunkt.

Bacher, J., Pöge, A. & Wenzig, K. (2010). *Clusteranalyse. Anwendungsorientierte Einführung in Klassifikationsverfahren*. 3. Auflage. München: Oldenbourg Verlag.

Blasius, J. & Thiessen, V. (2012). *Assessing the Quality of Survey Data.* Chichester, UK: Sage.

Borg, I., Groenen, P. J. F. & Mair, P. (2013). *Applied Multidimensional Scaling.* Heidelberg: Springer Verlag.

Borg, I. & Groenen, P. J. F. (2005). *Modern Multidimensional Scaling. Theory and Application. Second edition*. New York, NY: Springer.

Bryman, A. & Cramer, D. (2009). Constructing variables. In M. Hardy & A. Bryman (Eds.), *Handbook of Data Analysis* (pp. 17-34). London: Sage Publications, Inc.

Cherkassky, V. & Mulier, F. (2007). *Learning from Data. Concepts, Theory, and Methods.* Second edition. Hoboken, NJ: Wiley & Sons, Inc.

Coombs C. H. (1964). *A Theory of Data*. New York, NY: Wiley & Sons, Inc.

Cox, T. F. & Cox, M. A. A. (2001). *Multidimensional Scaling.* Second edition. Boca Raton, FL: Chapman & Hall/CRC.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika,* 16 (3), 297-334.

De Leeuw, J. (2005). Multidimensional Unfolding. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science.* Vol. 3 (pp. 1289-1294). New York, NY: Wiley & Sons, Inc.

Gediga, G. (1998). *Skalierung: Eine Einführung in die Methodik zur Entwicklung von Test- und Messinstrumenten in den Verhaltenswissenschaften*. Münster: Lit Verlag.

Green, R. J. (2010). *Coastal Towns in Transition. Local Perceptions of Landscape Change.* Dordrecht: Springer.

Greenacre, M. J. (2013). Correspondence Analysis. In T. D. Little (Ed.), *The Oxford Handbook of Quantitative Methods. Statistical Analyses.* Vol. 2 (pp. 142-153). New York, NY: Oxford University Press.

Hart, R. A. (1992). *Children's participation. From tokenism to citizenship.* Innocenti essays, No. 4. Florence: UNICEF, United Nations Children's Fund.

Hart, R. A. (2008). Stepping back from 'the ladder': Reflections on a model of participatory work with children. In A. Reid, B.B. Nikel & V. Simovska, (Eds.), *Participation and Learning. Perspectives on Education and the Environment, Health and Sustainability* (pp. 19-31). Dordrecht: Springer.

Hoffman, D. L. & de Leeuw, J. (1992). Interpreting multiple correspondence analysis as a multidimensional scaling method. *Marketing Letters,* 3 (3), 259-272.

IBM-SPSS (2011a). *IBM-SPSS. Advanced Statistics 22*. Chicago: IBM-SPSS.

IBM-SPSS (2011b). *IBM-SPSS. Categories 22*. Chicago: IBM-SPSS.

James, A., Jenks, C. & Prout, A. (1998). *Theorizing Childhood*. Cambridge: Polity Press.

Kish, L. (1965). *Survey Sampling.* New York, NY: Wiley & Sons, Inc.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29 (2), 115-129.

Kruskal, J. B. & Wish, M. (1978). *Multidimensional scaling. Sage University paper series on quantitative applications in the social sciences.* Number 07-11. Beverly Hills: Sage Publications, Inc.

Lewin, K. & Lewin, G. (1982 [1941]). Democracy and school. *Understanding the child, 10,* 7-10. (KLW 6; 285-291, "Demokratie und Schule").

Lewin, K., Lippit, R., & White, R. K. (1939). Patterns of aggressive behavior in experimentally created "social climates." *Journal of Social Psychology, 10*, 217-299.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140 (22),* 5-55.

Linting, M., Meulman, J. J., Groenen, P. J. F. & van der Kooij, A. J. (2007). Nonlinear principal components analysis: Introduction and application. *Psychological Methods, 12 (3),* 336-358.

Meulman, J. J. & Heiser W. J. (2011). *IBM SPSS Categories 20.* IBM: https://www.csun.edu/sites/default/files/categories20-64bit.pdf

Meulman J. J., van der Kooij A. J. & Heiser W. J. (2004). Principal component analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In D. Kaplan (Ed.), *Handbook of Quantitative Methodology for the Social Sciences* (pp. 49-70). Thousand Oaks, CA: Sage Publications, Inc.

Nunnally, J. C. (1978). *Psychometric Theory.* Second edition. New York, NY: McGraw-Hill.

OeAD (2014). The Austrian Educational System. https://www.oead.at/welcome_to_austria/education_research/the_austrian_education_system/EN/

Oser, F., & Biedermann, H. (2006). Partizipation – ein Begriff, der ein Meister der Verwirrung ist. In C. Quesel & F. Oser (Eds.), *Die Mühen der Freiheit: Probleme und Chancen der Partizipation von Kindern und Jugendlichen* (pp. 17-37). Zürich: Ruegger.

Romesberg, H. C. (2004). *Cluster Analyses for Researchers.* Place: NC: LULU Press.

Rost, J., Carstensen, C. H. & von Davier, M. (1999). Sind die Big Five Rasch-skalierbar? Diagnostica, 45 (3), 119-127.

Sturgis, P. (2004). *Analysing Complex Survey Data: Clustering, Stratification and Weights.* Social Research Update, Issue 43. Guildford: University of Surrey.

Van Schuur, W. H. & Kiers, H. A. L. (1994). Why factor analysis often is the incorrect model for analysing bipolar concepts, and what model to use instead. *Applied Psychological Measurement, 18 (2)*, 97-110.

Wetzelhütter, D., Paseka, A. & Bacher, J. (2013). Partizipation in der Organisation Schule aus der Perspektive der Schülerinnen und Schüler. In S. Weber, M. Göhlich, A. Schröer & C. Fahrenwald, (Eds.), *Organisation und Partizipation. Beiträge der Kommission Organisationspädagogik* (pp. 157-166). Wiesbaden: VS Verlag.

Wilk, L. & Bacher, J. (Eds.). (1994). *Kindliche Lebenswelten. Eine sozialwissenschaftliche Annäherung.* Opladen: Leske + Budrich.

Young, F. W. (1985). Multidimensional scaling. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences*, Vol. 5 (pp. 649-659). New York, NY: Wiley & Sons, Inc.

Zinnecker, J. (1999). Forschen für Kinder – Forschen mit Kindern – Kinderforschung. Über die Verbindung von Kindheits- und Methodendiskurs in der neuen Kindheitsforschung zu Beginn und am Ende des 20. Jahrhunderts. In M. Honig, A. Lange, & H.-R. Leu (Eds.), *Aus der Perspektive von Kindern? Zur Methodologie der Kindheitsforschung* (pp. 69-80). Weinheim & München: Juventa.

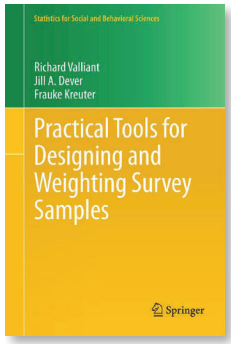
# Appendix A

*Table A1* descriptive results: participation-scale

| | fully applies | largely applies | uncertain | does not apply much | does not apply at all | n [(a)] |
|---|---|---|---|---|---|---|
| **Participation Item** (level at the scale) | | | Calibration sample | | | |
| V1: Pretend Participation (Level 1) | 6.0 | 13.7 | 33.7 | 34.9 | 11.6 | 335 |
| V2: Non-participation (Level 2) | 10.7 | 21.2 | 29.6 | 25.7 | 12.8 | 335 |
| V3: Tokenism (Level 3) | 3.9 | 23.3 | 36.4 | 27.8 | 8.7 | 335 |
| V4: Informed participation (Level 4) | 5.7 | 21.8 | 32.5 | 24.2 | 15.8 | 335 |
| V5: Consulted but informed (Level 5) | 11.9 | 33.7 | 33.1 | 14.9 | 6.3 | 335 |
| V6: Democratic vote (Level 6) | 12.2 | 34.3 | 26.0 | 17.6 | 9.9 | 335 |
| V7: Self-responsibility (Level 7) | 11.3 | 30.1 | 34.0 | 19.1 | 5.4 | 335 |
| V8: Supported self-responsibility (Level 8) | 13.1 | 36.7 | 28.4 | 14.3 | 7.5 | 335 |
| **Participation Item** (level at the scale) | | | Validation sample | | | |
| V1: Pretend Participation (Level 1) | 7.2 | 22.6 | 28.4 | 33.8 | 8.0 | 3779 |
| V2: Non-participation (Level 2) | 14.2 | 23.0 | 29.0 | 22.6 | 11.2 | 3779 |
| V3: Tokenism (Level 3) | 3.5 | 17.5 | 31.2 | 34.6 | 13.3 | 3779 |
| V4: Informed participation (Level 4) | 7.8 | 25.0 | 30.6 | 26.6 | 10.1 | 3779 |
| V5: Consulted but informed (Level 5) | 6.4 | 33.0 | 29.7 | 22.8 | 8.2 | 3779 |
| V6: Democratic vote (Level 6) | 8.3 | 27.0 | 27.9 | 22.1 | 14.7 | 3779 |
| V7: Self-responsibility (Level 7) | 9.2 | 30.5 | 30.7 | 21.5 | 8.1 | 3779 |
| V8: Supported self-responsibility (Level 8) | 11.9 | 32.8 | 27.1 | 19.2 | 9.1 | 3779 |

*Note*. (a) Exclusively cases with missing values or without variation in the response

# Book Reviews




Richard Valliant, Jill A. Dever, Frauke Kreuter (2013):
Practical Tools for Designing and Weighting Survey Samples
Springer (Statistics for Social and Behavioral Sciences; 51)
ISBN: 978-1-4614-6448-8
692 pages
€59.49

The preface of the book starts with: "Survey sampling is fundamentally an applied field. During this roundtable, we will discuss techniques long used by experienced survey statisticians with little or no references in the literature". And at the beginning of Chapter 1 one can read: "This is a practical book". This shows the basis of the book. The authors try to address three groups of readers:

1. *Students* seeking a more in-depth understanding of applied sampling either through a year-long course or by way of a supplementary reference.
2. *Survey statisticians* searching for practical guidance on how to apply concepts learned in theoretical or applied sampling courses.
3. *Social scientists* and other *survey practitioners* who desire insight into the statistical thinking taken to design, select, and weight random survey samples.

The book is organized into four parts with 18 chapters: *Designing Single-Stage Sample Surveys (Chapters 2-7)*, *Multistage Designs (Chapters 8-11)*, *Survey Weights and Analyses (Chapters 12-16)* and *Other Topics (Chapters 17-18)*. At the beginning of the first three parts the authors introduce each part with a real life example of a survey project, give some useful examples and at the end a solution. At the end of the book more than 100 pages are reserved for *Notation Glossary*, *Data Sets*, *R Functions Used in this Book, References, Solutions to Selected Exercises, Author Index* and *Subject Index*.

In Chapter 1 the authors mention that designing a sample plan requires the consideration of the following factors:

1. Specifying the objective(s) of the study.
2. Translating a subject-matter problem into a survey problem.
3. Specifying the target population, units of analysis, key study variables, auxiliary variables (i.e. covariates related to study variables and for which population statistics may be available), and population parameters to be estimated.

4. Determining which sample frame(s) are available for selecting units.
5. Selecting an appropriate method of data collection.

The book also contains hints such as: How to write a technical report describing the sample design.

In Chapter 2 the first of several projects demonstrates the workflow of designing a single-stage personnel survey. This chapter includes specifications for the study, questions raised by sampling experts in response to the specifications, preliminary analyses and documentation.

Chapter 3 contains several commonly used probability sampling plans, e.g. simple random sampling, stratified simple random sampling including different types of allocations of the sample size to the strata, sampling (with and without replacement) with varying probabilities, systematic sampling and Poisson sampling. Not only sampling designs are considered but also different estimators for estimating population parameters, such as means, totals and portions. Already at this stage the authors introduce the concept of the design effect. Sampling rare populations and estimation for subgroups (domains) is mentioned. R and SAS are presented as statistical software packages usable for sample selection, along with some examples. A lot of exercises complete Chapter 3.

Chapter 4 deals with one- and two-sided tests and power calculation.

In Chapter 5 the reader learns the methods of multi-criteria optimization for single-stage designs using the Microsoft Excel Solver, the SAS PROCs NLP and OPTMODEL or the R package Alabama.

Chapter 6 introduces disposition codes and outcome rates following the standard definitions of AAPOR.

In Chapter 7 solutions to the multipurpose design of a survey introduced in Chapter 2 are suggested.

With Chapter 8, Part II of the book starts with a new project designing an area sample.

First the known formulae in the context of multistage samples are presented and then their use is illustrated by an example using a data set called Maryland area population. The authors elaborate the rules for designing primary sampling units. The Current Population Survey and other surveys serve as examples to explain the different concepts to the reader.

The important field of weighting starts with Chapter 13. Different reasons for the necessity of using weights are given, such as reducing the variance of estimators and to adjust for under-coverage by using appropriate auxiliary data. Different kinds of calibration techniques, including post-stratification, raking and the well-known GREG are introduced. Especially the GREG is illustrated by general examples.

Variance estimation serves to determine the precision of an estimator. Exact methods, linearization and replication methods are presented. The weighting part is completed by a detailed example.

Chapter 17 is the beginning of part IV "Other Topics" introducing multiphase designs that are helpful, i.e. in the case of non-respondent subsampling. In Chapter 18 process control and quality measures are discussed. Performance rates and several other indicators are evaluated as important quality control tools. Also, data editing and documentation as essential parts of a survey are mentioned.

The authors do not cover all topics in the context of sample surveys, e.g. handling of missing data or using multiple frame problems which nowadays are wide spread in telephone sampling.

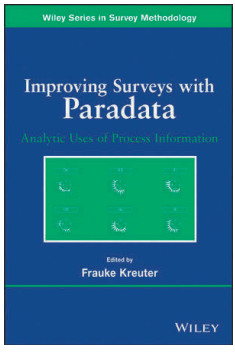
At URL *https://umd.app.box.com/s/9yvvibu4nz4q6rlw98ac* there is a folder containing files that accompany the book. "Examples.zip" contains R-code for 65 examples in the Practical Tools book (updated 2014.10.24). Changes and updates to the PracTools Package (updated 2015.01.21) are in the "NEWS changes and updates.pdf". In the R package PracTools (February 19, 2015) one can find functions for sample size calculation for survey samples using stratified, clustered, and or one-, two-, and three-stage sampling designs. Other functions compute variance components for multistage designs and sample sizes in two-phase designs. However, they are built under R Version 3.1.3 and will be hopefully soon updated to newer versions of R. The different projects mentioned in the book can be found in "Projects.zip".

A special offer is that one can buy single chapters of the book via
http://www.springer.com/de/book/9781461464488.

The book is very helpful for researchers, practitioners and all people designing a survey in practice, and does not just give the reader the formulae in the context of the sampling design and the estimators. The parts at the planning stage of a survey and the important quality control steps are often not mentioned in other textbooks. The abundance of examples helps the reader to understand the whole process. Thus, I can highly recommend this book.

*Siegfried Gabler*

DOI: 10.12758/mda.2015.005




Frauke Kreuter (Editor) (2013):
Improving Surveys with Paradata. Analytic Uses of Process Information
Wiley Series in Survey Methodology
ISBN: 978-0-470-90541-8
416 pages
€65.70

*Improving Surveys with Paradata. Analytic Uses of Process Information* by Frauke Kreuter is an edited volume containing timely contributions on the use of paradata in survey research, survey management and data analysis. Due to its wide scope and high quality, the book provides helpful information to survey methodologists, survey managers and analysts on how to make valuable use of this young data source.

Kreuter defines "paradata as additional data that can be captured during the process of producing a survey statistic. Those data can be captured at all stages of the survey process and with very different granularities." (Kreuter, 2013, p. 3). In other words, paradata are data collected during the survey data collection process either automatically through computerized systems or individually by interviewers. The book covers a great number of different types of paradata, the most central examples of which are keystrokes, time stamps, call record data and interviewer observations.

Different types of chapters make up this edited volume: reviews of existing studies using paradata to investigate survey errors, applications of paradata in the context of survey production and contributions on the measurement properties of paradata. Finally, the book draws attention to possible applications that still lack empirical evidence, thus formulating a future research agenda.

While some chapters are pitched at an introductory level looking into data structures of paradata and addressing data preparation concerns, other chapters target a statistically advanced readership and cover more complex models for the analysis of paradata. Although these complex models are illustrated by means of mathematical formulas and therefore demand corresponding reading abilities, these formulas are derived and explained step-by-step, thus enabling even less experienced readers to comprehend them.

The book comprises a total of 15 chapters, the content of which is briefly reviewed in the following. In Chapter 1 the editor provides an introduction by defining para-

data and relating them to metadata and auxiliary data as well as an outline on the structure of the book. The remaining chapters are divided into three parts each focusing on the use of paradata in a different setting.

The first part *Paradata and Survey Errors* includes four chapters that address the use of paradata to investigate and adjust for errors occurring at different stages of the survey process. More precisely, chapters 2, 3, 4 and 5 focus on three error components taken from the total survey error framework (see Kreuter, 2013): nonresponse, measurement and coverage error. For each of these error components the aforementioned chapters name relevant paradata and inform about previous research including these data.

In Chapter 2, Kreuter and Olson give a review of the kinds of paradata that have been used to analyze nonresponse error to date. Furthermore, this chapter discusses the possibilities of including paradata in nonresponse adjustments. Chapters 3 and 4 both deal with paradata and measurement error. In Chapter 3, Olson and Parkhurst describe which paradata are important for assessing measurement error and how they differ by mode of data collection. In Chapter 4, Yan and Olson report studies using these paradata to investigate measurement error, therefore supporting the findings of the previous chapter with empirical evidence. Finally, in chapter 5 Eckman focuses on coverage error. For this purpose, she gives an overview of existing research with paradata on frame errors and coverage bias, and points to additional possibilities of using paradata to study coverage error.

The second part *Paradata in Survey Production* consists of five chapters demonstrating how paradata may be used either after data collection to correct for survey errors or directly during survey data collection to guide changes in survey design to increase survey quality. Although these five chapters each consist of individual case studies, the findings can easily be transferred to other survey settings.

In chapter 6, Kirgis and Lepkowski give insight into design changes in the 2006-2010 National Survey of Family Growth based on paradata from a previous data collection. Paradata from the ongoing survey were then used to monitor these design changes and informed responsive design elements. In chapter 7, Wagner reports a number of studies carried out to optimize calling strategies. For this purpose, he includes paradata into models estimating contact probabilities. In chapter 8, Sakshaug summarizes the use of paradata to study nonresponse regarding additional requests within surveys, such as consent to record linkage. In addition, he points out how paradata may be used to intervene on respondents probably not giving consent to those requests. In chapter 9, Jans, Sirkis and Morgan show that paradata can play a decisive role in monitoring the quality of ongoing data collections; more specifically, they demonstrate how they are implemented in graphical control displays. In chapter 10, Schouten and Calinescu present original research

into paradata as predictors of nonresponse and measurement error using data from the Dutch Labour Force Survey.

The third part *Special Challenges* contains five chapters dealing with the measurement properties of paradata. The chapters in this part cover mode-specific paradata, demonstrate methods for analyzing paradata with particular data structures and focus on quality issues.

In Chapter 11, Callegaro provides a typology of paradata that can be collected in web surveys. Furthermore, he gives examples of how these paradata have been used thus far. Chapters 12 and 13 both deal with analytical challenges resulting from the data structure. In Chapter 12, Durrant, D'Arrigo and Müller describe the use of multilevel modeling to analyze hierarchically structured call record data. In Chapter 13, Schafer shows how penalized spline models can be employed to monitor quality indicators with means shifting over time. The two remaining chapters 14 and 15 both focus on the quality of paradata. In Chapter 14, West and Sinibaldi provide a literature review on the quality of various paradata and discuss potential error sources. In Chapter 15, West demonstrates by means of several simulation studies how erroneous paradata may affect nonresponse adjustments.

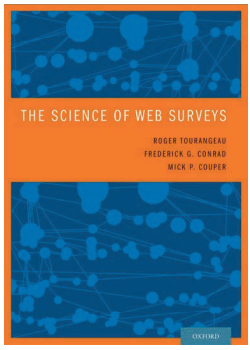
Each chapter of the book is self-contained. Thus, the reader may choose to read the book from front to back or just individual chapters without missing context. To the reader's convenience, there are many cross-references linking the individual chapters.

In conclusion, *Improving Surveys with Paradata. Analytic Uses of Process Information* provides a comprehensive overview of the possibilities for using paradata in survey methodological research and field management. The book illustrates a variety of areas paradata may contribute to, including advances in academic research on and with paradata as well as more survey practical settings, such as during fieldwork monitoring. As research into paradata is relatively new there is still a need for further analyses. In this way, the book valuably summarizes the current state of research. Whether researchers in academia or survey practitioners, whether experienced or just starting out, we can highly recommend this book to anyone interested in integrating paradata in their daily work.

## Literature

Kreuter, F. (2013). Improving Surveys with Paradata: Introduction. In F. Kreuter (Eds.), *Improving Surveys with Paradata. Analytic Uses of Process Information* (pp. 1-9). Hoboken, New Jersey: Wiley.

*Franziska Gebhard & Annelies G. Blom*
DOI: 10.12758/mda.2015.006




Roger Tourangeau, Frederick Conrad, Mick Couper
(2014):
The Science of Web Surveys
Oxford University Press
208 Pages
ISBN: 978-0-19-974704-7
$72.00

*The Science of Web Surveys* by Tourangeau, Conrad, and Couper is a summary of state-of-the-art methodological research on surveys using the web mode of data collection. It mainly addresses academic researchers, but the book can also be recommended for survey practitioners conducting web surveys. Unlike other publications in the area of web surveys (e.g. Callegaro et al. 2014, or Couper 2008), it has a relatively broad research scope.

Generally, the book is structured following a Total Survey Error perspective. However, although there are two chapters on errors of nonobservation (chapter two on sampling and coverage error and chapter three on nonresponse error), its main focus is on errors of observation. Therefore, this book can especially be recommended for researchers interested in measurement error in web surveys. For survey practitioners, it offers useful advice on visual design aspects and interactive features.

All chapters provide a rich collection, review, and comparison of recent studies found in the literature about web surveys. Studies of the same problem are displayed and compared in tables. This is helpful to gain an overview of key findings.

In the introductory chapter of their book, the authors discuss the evolution as well as general advantages and disadvantages of web surveys. They briefly explain the Total Survey Error logic and state the structure and intended audience of the book.

Chapters two and three are dedicated to the discussion of errors of nonobservation in web surveys. In chapter two, the authors schematically depict different types of web surveys. On the basis of this typology, they discuss the lack of a commonly shared sampling method for web surveys, the statistical consequences of nonprobability sampling, and the challenge of the "digital divide" of individuals with and without Internet access. At the end of the chapter, there is a short introduction to and comparison of commonly applied weighting procedures.

In chapter three, the authors provide definitions on nonresponse error as well as nonresponse rates. They then discuss the effect of well-established design features that affect web survey participation, such as prenotification letters and incentives.

In this chapter, they also discuss some special web survey issues, for example, mixed-mode design and survey break-offs. The distinction between probability-based and nonprobability surveys plays an important role throughout the chapter. In this context, the authors criticize the tendency of some survey data users to care more about the amount of data than its quality (p. 43).

The fourth chapter introduces the section on errors of observation, which expands over the rest of the book. This chapter is a primer on basic web survey design features, for example, the web page layout. The authors describe visual possibilities of web surveys, such as multimedia presentations and interactive features. They also point out that researchers need to be careful when using technological advances in order to reduce and not enhance measurement error (p. 59-61). For instance, researchers should be aware of respondents who have not installed JavaScript on their computers and therefore cannot see certain kinds of survey design features.

In chapter five on the web as a visual medium, common measurement issues with regard to visual presentation in web surveys are discussed, such as response scale interpretations and the impact of images in the questionnaire. The authors highlight differences between the web and other modes of data collection with regard to response behavior, for example, how having to use a cursor influences response distributions. They argue that the web, more than other modes, depends on "visibility," that is the degree to which visible features of a web page are noticed by respondents (p. 93). Among other findings, the authors present results from eye-tracking studies which show that in a vertical list of response options the time and number of respondent fixations on a row of text decreases from top to bottom (p. 96).

Chapter six is about the interactive possibilities inherent to web surveys. Examples include progress bars to show the proportion of answered and yet to be answered questions, interventions to stop respondents from "speeding" through the questionnaire, and even more elaborate and creative features such as animated interviewer faces (p. 122-126).

Chapter seven contains a meta-analysis comparing web surveys to alternative data collection modes with regard to the amount of measurement error found. One of the key findings is that measurement error due to socially desirable responding is lowest in web surveys, especially when compared to interviewer-administered surveys (p. 142). Based on these results, the authors reflect on the effect of virtual interviewers in web surveys on potential measurement error. They conclude that the more "lifelike" these interviewers are, the more they induce social desirability bias (p. 145).

In the final chapter of the book, there is a summary of findings from the previous chapters, as well as a brief list of practical recommendations for web survey practitioners. The chapter also contains a mathematical model for survey mode effects. This comes rather unexpectedly, mainly because the application of multiple modes

is not discussed in great detail in the previous chapters. Nevertheless, the proposed model is a comprehensible formalization that is helpful for assessing the consequences of a multimodal survey design.

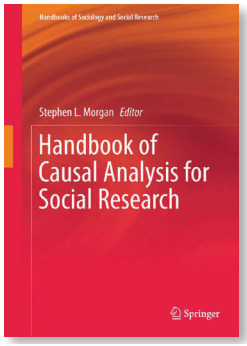
In conclusion, *The Science of Web Surveys* is an excellent review of state-of-the-art research in the area of web survey data collection. It provides comprehensive literature discussions, especially with regard to measurement error. This book informs readers about advances in understanding web survey phenomena. Additionally, it serves as a guideline to effective web survey design. The book is compact, comprehensible, and I can highly recommend it.

## Literature

Callegaro, M., Baker, Reg. P.,Bethlehem, J., Göritz, A. S., Krosnick, J. A., Lavrakas P. J. (2014): *Online panel research. A data quality perspective.* Chichester: Wiley (Wiley series in survey methodology).

Couper, M. (2008). *Designing Effective Web Surveys.* New York: Cambridge University Press.

*Carina Cornesse*
DOI: 10.12758/mda.2015.007




Stephen L. Morgan (Editor) (2013):
Handbook of Causal Analysis for Social Research
New York: Springer.
ISBN 978-94-007-6093-6
431 Pages
109 $ (Softcover)

Causality is not only a core topic in the philosophy of science and the hard sciences. Since their very beginning social scientists were deeply interested in the possibilities and pitfalls of causal inference. Controversies on the topic were and are particularly intense in the social sciences as the particularities of the field raise additional problems and often hamper the approximation to the heuristic ideal of the randomized controlled experiment. Over time different methods like correlational, path, regression, and event history analysis were in vogue. In recent decades the literature on causality is moving fast forward, receiving important input both from theorists interested in explanation and social mechanisms and methodologists developing on statistical concepts of causality and specific identification/estimation techniques.

The Handbook of Causal Analysis for Social Research, edited by Stephen Morgan and with chapters by leading experts in the field, both reflects this long history of causal inference in the social sciences and gives an overview of the current state of the art. The volume has 19 chapters divided into six parts.

In Part I "Background and Approaches to Analysis" Sondra N. Barringer, Scott R. Eliason, and Erin Leahey give a concise review of the theoretical, statistical, and econometric literature on causality linking the concept back to classical work (e.g., by Mill, Hume, and Weber) and presenting old and new variants of causal modeling. Jeremy Freese and J. Alex Kevern give background on types of causes, introduce several distinctions (manipulable/preventable, proximal/distal, necessary/sufficient; proximate/ultimate), and highlight that causes can differ in their importance.

In Part II "Design and Modeling Choices" Herbert L. Smith proposes a design-based identification approach "de-emphasizing computational statistical considerations in favor of observational frames" (p. 47), and explains why the experiment is commonly regarded as the "gold standard". However, Smith also emphasizes that "the experiment itself is more ideal—model—than method in the social sciences." (p. 57) In chapter 5 James Mahoney, Gary Goertz, and Charles C. Ragin com-

pare additive, linear causal models, commonly used by statistical researchers, and logic-based causal models, commonly used by set-theoretic researchers, seeking to clarify fundamental dissimilarities on the conceptual level rather than to unify both approaches. David J. Harding and Kristin S. Seefeldt pledge in Chapter 6 "Mixed Methods and Causal Analysis" for a combination of quantitative and qualitative methodologies to strengthen causal analyses, for example by elucidating selection processes, intervening mechanisms, and sources of heterogeneity.

Part III "Beyond Conventional Regression Models" zooms in on selected statistical methods. Glenn Firebaugh, Cody Warner, and Michael Massoglia give an overview of methods for longitudinal data analysis and emphasize the advantages of fixed effects and hybrid models. The contribution of Hui Zheng, Yang Yang, and Kenneth C. Land deals with heteroscedastic regression models which allow the researchers to directly model heteroscedasticity. Tim F. Liao presents methods to analyze group differences (as regards linear predictors and underlying distributions) in generalized linear models. Richard Breen and Kristian Bernt Karlson discuss the specific challenges of non-linear probability models in identification and estimation of causal effects as well as techniques to circumvent these problems. Jennie E. Brand and Juli Simon Thomas raise awareness for the well-known, but in practice often neglected problem of treatment-effect heterogeneity for causal inference and "encourage researchers to routinely examine treatment-effect heterogeneity with the same rigor they devote to pretreatment heterogeneity." (p. 189) Xiaolu Wang and Michael E. Sobel conclude this section of the book with a chapter on a topic gaining in prominence in recent years: causal mediation analysis. They show that identification of indirect effects often rests on implausible assumptions and sketch alternatives to the common practice.

Part IV "Systems of Causal Relationships" heavily draws on graphical causal models, which are introduced by Felix Elwert with a special focus on endogenous selection bias caused by collider variables. Carly R. Knight and Christopher Winship illustrate the analytical potential of directed acyclic graphs for the study of causal mechanisms. After discussing conceptions of causal mechanisms the authors explicate preconditions for a successful front-door analysis and give hints how identification issues can be solved. Finally, Kenneth A. Bollen and Judea Pearl aim at clearing persistent myths and misunderstandings about structural equation modeling and rehabilitating the method. They "conclude that the current capabilities of SEMs to formalize and implement causal inference tasks are indispensable; its potential to do more is even greater." (p. 301)

Part V "Influence and Interference" consists of two chapters on social interactions. The presence of social ties complicates identification of causal effects due to self-selection and violations of the stable unit treatment value assumption. In Chapter 16 Guanglei Hong and StephenW. Raudenbush present models which take the social

fact of interference among participants into account and help to relax the often overly restrictive and unrealistic SUTVA. In Chapter 17 Tyler J. VanderWeele and Weihua An explicate that the separation of social influence, homophily, and environmental confounding is one of the major tasks in social network analysis which can be tackled with longitudinal data.

The last part of the book, VI "Retreat from Effect Identification" begins with an overview by Markus Gangl of instrumental variables, sensitivity analysis, and non-parametric bounds on treatment effects. As the author shows these methods for partial identification help researchers to better understand and communicate the robustness of their results. Finally, Richard A. Berk et al. caption their chapter with the statement "What You Can Learn from Wrong Causal Models". They do not argue that wrong model assumption do not compromise causal inference, but claim that "causal thinking can help inform how a statistical approximation is specified, and […] can be instrumental when results need to be interpreted" (p. 423)

As this short summary illustrates the handbook covers a wide range of important topics of causal inference and surely is an invaluable resource for students and researchers interested in the topic. Thereby, the handbook complements the excellent and influential textbook *Counterfactuals and Causal Inference* by Morgan and Winship (2015, 2nd ed.). On the one hand, many of the contributors also draw heavily on the counterfactual model of causal inference (Rubin) and graphical models of causality (Pearl). On the other hand, the volume gives more room to unresolved discussions and conflicting positions than is possible in a textbook. This reflects that the "target audiences" of the handbook are "advanced graduate students and faculty researchers in sociology" (p. v). As always in edited volumes, although most chapters follow a common structure, the analytical rigor and the required formal background vary. Thus, probably not all chapters are equally accessible to graduate students and – to guard against misunderstandings – the handbook also gives no guidance how to conduct analyses with a specific statistical software package. Nonetheless, due to the exceptionally high quality, the clarity of presentation, and the many examples the handbook is well-suited for teaching methodology to advanced classes. Finally, I have no doubt that the volume will fulfil its main aim: it will bring the field of causal inference forward and raise the methodological rigor of social science research in general.

*Tobias Wolbring*
*DOI: 10.12758/mda.2015.008*

# New releases in ZIS

Open-access measurement instruments for the social sciences

ZIS offers more than 200 measurement instruments, freely available items, background information, and data sets. Publications are in German or English. The following instruments have recently been published:

**Political Efficacy Short Scale (PEKS)**

Constanze Beierlein and colleagues published a 3 item scale measuring internal and external political efficacy. Empirical results support the reliability and the convergent and divergent validity. The items are available in German and English. Reference statistics for a probability sample of the German population are provided.

Beierlein, C., Kemper, C. J., Kovaleva, A. & Rammstedt, B. (2014). *Political Efficacy Kurzskala (PEKS)*. In D. Danner & A. Glöckner-Rist (Eds.), Zusammenstellung sozialwissenschaftlicher Items und Skalen. doi: 10.6102/zis34

**Crystalized Intelligence Short Scale**

Stefan Schipolowski and colleagues published a 12 item scale that allows measuring crystallized intelligence. The short scale is based on the long version of the BEFKI intelligence test and allows a reliable and valid measurement of crystalized intelligence. The items are available in German. Reference statistics for a probability sample of the German population are provided.

Schipolowski, S., Wilhelm, O., Schroeders, U., Kovaleva, A., Kemper, C. J., & Rammstedt, B. (2014). *Kurzskala kristalline Intelligenz (BEFKI GC-K)*. In D. Danner & A. Glöckner-Rist (Eds.), *Zusammenstellung sozialwissenschaftlicher Items und Skalen*. doi: 10.6102/zis220

**Ethnic and national identity of adolescents**

Lars Leszczensky and colleagues published a 14 Item scale that assesses the ethnic and national identity of adolescents. The scale contains measures of cognitive, evaluative, and affective aspects of identity. Empirical results support the reliability, validity, and configural measurement invariance for different groups. The items are available in German.

Leszczensky, L. & Gräbs Santiago, A. (2014). *Ethnische und nationale Identität von Kindern und Jugendlichen*. In D. Danner & A. Glöckner-Rist (Eds.), *Zusammenstellung sozialwissenschaftlicher Items und Skalen.* doi: 10.6102/zis158

www.gesis.org/zis

# Information for Authors

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects.

Manuscripts that have already been published elsewhere or are simultaneously submitted to other journals will not be considered. As a rule we do not restrict authors' rights. All rights remain with the author, and articles in mda are published under the CC-BY open-access license.

Mda aims for a quick peer-review process. All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. The decision on publication is made by the editors based on the reviews. The editorial team will contact the authors by email with the result at the latest eight weeks after submission; if the reviews have not been received by then, we provide a status update with a new target date.

When preparing a paper for submission, please consider the following guidelines:

- Please submit your manuscript by e-mail to mda(at)GESIS(dot)org.
- The total length of the manuscript shall not exceed 10.000 words.
- Manuscripts should…
  - be written in English, using American English spelling. Please use correct grammar and punctuation. Non-native English speakers should consider a professional language editing prior to publication.
  - be typed in a 12 pt Roman font, double-spaced throughout.
  - start with a cover page containing the title of the paper and contact details / affiliations of the authors, but be anonymized for review otherwise.
- Please also send us an abstract of your paper (approx. 300 words), a brief biographical note (no longer than 250 words), and a list of 5-7 keywords for your paper.
- Acceptable formats for Graphics are
  - Tiff
  - Jpeg (uncompressed, high quality)
  - pdf
- Please ensure a resolution of at least 300 dpi and take care to send hiqh-quality graphics. Line art images should have a resolution of 500-1000 dpi. Please note that we cannot print color images.
- The type area of our journal is 11.5 cm (width) x 18.5 cm (height). Please consider this when producing tables or graphics.
- Footnotes should be used sparingly.
- By submitting a paper to mda the authors agree to make data and program routines available for purposes of replication.

Please follow the APA guidelines when preparing in-text references and the list of references.

**Entire Book:**

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: John Wiley & Sons.

**Journal Article (with DOI):**

Klimoski, R., & Palmer, S. (1993). The ADA and the hiring process in organizations. *Consulting Psychology Journal: Practice and Research*, 45(2), 10-36. doi:10.1037/1061-4087.45.2.10

**Journal Article (without DOI):**

Abraham, K. G., Helms, S., & Presser, S. (2009). How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the United States. *American Journal of Sociology*, 114(4), 1129-1165.

**Chapter in an Edited Book:**

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research*. Second Edition (pp. 593-630). Bingley: Emerald.

**Internet Source (without DOI):**

Lewis, O., & Redish, L. (2011). *Native American tribes of Wisconsin*. Retrieved April 19, 2012, from the Native Languages of the Americas website: www.native-languages.org/wisconsin.htm

For more information, please consult the Publication Manual of the American Psychological Association (Sixth ed.).

gesis
Leibniz Institute for the Social Sciences

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)