Sociological Methods
and Research

The Impact of Survey Mode Design and Questionnaire Length on Measurement Quality

SCHOLARONE™
Manuscripts

Introduction

Face-to-face and telephone surveys are currently being challenged by decreasing response rates (Beullens et al., 2018; Brick & Williams, 2013; Leeuw et al., 2018) and a corresponding rise in costs due to intensified fieldwork efforts (Wolf et al., 2021). These forces have prompted many social surveys to consider moving away from traditional interviewer-administered single-mode designs to mixed-mode designs involving a combination of interviewer- and self-administered (e.g. web) modes, or solely self-administered using either paper-based mail or online modes. For example, longitudinal studies such as the Health and Retirement Study, the UK Household Longitudinal Study (*Understanding Society*), and the National Longitudinal Study of Adolescent to Adult Health (Add Health) have transitioned from fully face-to-face or telephone survey designs to mixed-mode designs involving online and/or mail-based data collection (Bianchi et al., 2017; Biemer et al., 2021; Cernat et al., 2016). Similarly, repeated cross-sectional and cross-national studies, such as the European Social Survey and the European Values Study have recently experimented with online and mail-based data collection as an alternative to face-to-face data collection in some countries (Cernat & Revilla, 2020; Luijkx et al., 2020). Most recently, the COVID-19 pandemic has forced many surveys to switch from face-to-face data collection to telephone and self-administered modes (e.g. Gummer, Schmiedeberg, et al., 2020; Sakshaug et al., 2020).

Yet, while adopting mixed-mode designs with greater use of self-administration have been shown to be cost-effective (Bianchi et al., 2017; Kappelhof, 2015; Wagner et al., 2014), such a design change can have non-trivial effects on data quality. This is due to the fact that different interview modes have inherently different properties (e.g. level of interviewer presence,

aural vs. visual stimuli) that can influence respondents' answers. These properties are particularly distinct between interviewer- and self-administered modes, and there is a broad literature showing that the same groups of respondents can provide different answers to the same items depending on which mode type is used (de Leeuw, 2005). The length of the questionnaire also has data quality implications, especially when a long questionnaire designed for interviewer-administration is administered via self-completion, which may lead to greater respondent burden and poorer data quality for questions positioned later in the questionnaire (Galesic & Bosnjak, 2009). Dropping questions or administering subsets of the questionnaire to different sets of respondents minimizes this risk (Raghunathan & Grizzle, 1995), but may result in a lack of measurement comparability with the full-length version. Thus, there is a need to understand the implications of these design decisions on data quality and data comparability.

Against this backdrop, it is important for practitioners to understand the measurement implications of changing from an interviewer-administered mode design with a long questionnaire to a self-administered mode design with a shorter questionnaire. In this article, we shed light on these issues by investigating differences in measurement quality in the German sample of the European Values Study (EVS). Sampled individuals of the general population were randomized to be interviewed using the traditional face-to-face mode design, a mixed-mode paper/web design, or a mixed-mode paper/web design with a shorter questionnaire implemented as part of a matrix sampling design. Using this experimental setup, it is possible to assess the extent to which measurement quality differs between the mode designs and questionnaire lengths. These comparisons will reveal potential tradeoffs of shifting to a self-administered mode design either as a supplement or as a replacement for traditional face-to-face data collection. The following research questions (RQ) are addressed:

RQ1) What is the impact of mode design and questionnaire length on point estimates and

distributions of survey items?

RQ2) To what extent does mode design and questionnaire length affect data quality

indicators (e.g. item nonresponse, response style indicators)?

RQ3) Does measurement equivalence hold for attitudinal scales measured under different

mode designs and questionnaire lengths?

Previous Research

The choice of data collection mode(s) is a crucial decision for any survey as it affects how

respondents receive and answer questions. Different modes have inherently distinct features that

affect the presentation of the questions and influences the answers that respondents provide.

Consequently, different modes can elicit different answers to the same questions posed to the

same respondents (de Leeuw, 2005). These "measurement mode effects" can arise when some

respondents are interviewed in one mode, while other respondents are interviewed in an

alternative mode. Measurement mode effects are problematic as they can bias comparisons

between respondents who are interviewed in different modes. Such effects can occur when

comparing respondents in a mixed-mode survey or in cross-national settings where different

countries administer the same questionnaire in different modes. Mode effects tend to be more

severe when mixing interviewer-administered (e.g. face-to-face, telephone) and self-

administered (e.g. mail, web) modes, rather than mixing within them (Cernat et al., 2016;

Klausch et al., 2013). These two mode types differ with respect to two key features which have

been attributed to their potential for mode effects: the communication channel (aural or visual)

and the presence/absence of an interviewer. Below we describe each feature and its potential

impact on measurement mode effects in turn.

The communication channel refers to whether the survey questions are presented visually

or orally to respondents. Face-to-face and telephone are primarily aural modes as interviewers

read the questions out loud and respondents process the information aurally. In contrast, mail and

web are primarily visual modes as respondents read the questions and process the information

using visual cues. There are some exceptions. For instance, face-to-face interviews are known to

use showcards to supplement the oral presentation (Lynn et al., 2012) and web surveys may use

audio components to supplement the visual presentation. Both communication channels are

suggested to affect the cognitive process and memory capacity in different ways (Krosnick &

Alwin, 1987; Schwarz et al., 1991). Because of this, different communication channels can lead

to different response behaviors. For example, aural modes are often associated with recency

effects, defined as the tendency to select the last spoken answer categories, which are most easily

remembered compared to the earlier spoken answer categories (Smyth et al., 1987). Visual

modes are usually associated with primacy effects, or the tendency to endorse the first answer

categories, which is in line with the notion that respondents select the first acceptable answer

category they see (Krosnick & Alwin, 1987). Though the empirical evidence on primacy and

recency effects in different modes is inconsistent (E. D. de Leeuw, 2005). Other response

behaviors, including the level of nondifferentiation (or straightlining) to attitudinal item batteries

and the frequency of selecting extreme answers to rating scales, have also been shown to vary

between self- and interviewer-administered modes more so than within them (Dillman et al.,

2009; Kim et al., 2019).

Regarding the presence/absence of an interviewer, this mode feature can have a strong influence on the answers that respondents provide. Interviewers are a well-known source of variance inflation (or "interviewer effects") in survey estimates (West & Blom, 2017), but they also have an important effect on social desirability bias. A consistent research finding is that interviewer-administered modes produce higher levels of socially desirable responding compared to self-administered modes (Cernat et al., 2016; Heerwegh, 2009; Kreuter et al., 2008; Tourangeau & Yan, 2007). However, interviewer presence may reduce the frequency of other suboptimal response behaviors, such as nondifferentiation and item nonresponse (Hope et al., 2014), by keeping respondents motivated and focused on the response task, which is more difficult to do in self-completion modes.

With respect to mode effects in attitudinal items and Likert scales, Klausch et al. (2013) report higher category thresholds for attitudes about police and traffic in the Netherlands in face-to-face and telephone modes compared to mail and web modes, an indication of stronger socially desirable responding in the former modes. Heerwegh & Loosveldt (2011) report evidence of socially desirable responding among telephone responses in a crime victimization survey in Belgium, reflected in more favorable attitudes toward the police, compared to mail responses. Cernat et al. (2016) identified higher levels of depression reported online compared to telephone and face-to-face in the US Health and Retirement Study. Cernat & Revilla (2020) investigated differences in measurement quality between the face-to-face ESS Round 8 and the CROss-National Online Survey (CRONOS) panel. They find higher item nonresponse and higher levels of primacy effects in the CRONOS panel, but similar levels of nondifferentiation with the ESS. Using equivalence testing they find that metric and scalar equivalence holds for four out of the five scales tested.

In addition to mode effects, the length of the questionnaire can also have data quality implications when transitioning from interviewer-administration to self-completion. While it is common to use interviewer modes to administer long questionnaires to respondents, administering long questionnaires via self-completion is likely to require greater effort from the respondent. In the case of web surveys, researchers have suggested the ideal length should not exceed 20 minutes (Callegaro et al., 2015; Revilla & Ochoa, 2017). Extending the questionnaire beyond this length bears the risk that respondents will engage in undesirable response behaviors toward the end of the interview as fatigue accumulates and they lose focus and motivation to invest the necessary effort to provide high-quality answers. Galesic & Bosnjak (2009) found that items positioned later in the questionnaire of a web survey were answered more quickly, with shorter answers given to open-ended questions, less variability in answers given to grid questions, and slightly more item nonresponse, compared to the same items positioned earlier in the questionnaire. Peytchev & Peytcheva (2017) also found strong evidence of greater measurement error when items on diet and exercise appeared later in a web survey instrument. Similar findings have been reported in self-administered paper questionnaires (Herzog & Bachman, 1981; Sahlqvist et al., 2011).

One strategy to minimize respondent burden is to shorten the questionnaire by dropping items. However, this strategy may be undesirable from a research perspective or impossible for a country that is participating in a cross-national survey with a fixed questionnaire. An alternative strategy is to implement a split or matrix questionnaire design by partitioning the questionnaire into shorter modules and administering subsets of the modules to different respondents (Raghunathan & Grizzle, 1995). The items that are not answered by all respondents are then imputed to compensate for the data which are missing by design and produce a full rectangular

dataset available for multivariate analyses – which is, however, difficult to achieve in general population survey datasets that include hundreds of variables. In an experimental evaluation of the matrix questionnaire design on measurement error reduction, Peytchev & Peytcheva (2017) found that the matrix design performed better than the full questionnaire design on several metrics (e.g. estimates of means, mean squared error, associations) that are susceptible to measurement error. Yet, additional research is needed to understand the effects of using a matrix questionnaire design on measurement quality in multi-item scales, where the measurement properties (e.g. context effects) can vary due to the different questionnaire structures.

In this study, we evaluate the measurement effects of shifting from an interviewer-administered (face-to-face) survey to a fully self-administered mixed-mode (web/mail) survey and from a long questionnaire to a matrix questionnaire design by comparing the quality of respondent answers in the German sample of the EVS. Motivated by the above literature, we investigate whether the shift in mode design and questionnaire length affects estimates of means, distributional properties, and indicators of data quality (e.g. item nonresponse, primacy, recency, middle answers, extreme response style, straightlining) for 140 items and 25 scales measured in the general population. Additionally, we apply Confirmatory Factor Analysis to test the assumption of measurement equivalence for the multi-item scales across the different mode designs and questionnaire lengths.

Data and Experiment

The European Values Study (EVS) is a cross-national repeated cross-sectional survey research program examining public opinion on topics like the environment, national identity, perception

of life, politics and society, religion and morale, and work. The EVS was first conducted in 1981 with subsequent cross-sectional surveys being fielded every nine years. In the most recent data collection, the EVS allowed participating countries to test the use of self-administered data collection modes; for more details, see (Luijkx et al., 2020). In this study we use the EVS data collected in Germany (EVS, 2017a, 2017b).

Data Collection Experiment in the EVS Germany

The EVS 2017/18 in Germany featured an interviewer-administered face-to-face (computer-assisted personal interviewing; CAPI) survey and two additional self-administered mixed-mode (web and mail) surveys (Wolf et al., 2021). A probabilistic sample of people residing in Germany aged 18 or older at the time of fieldwork was drawn from population registers. For this purpose, a two-stage sampling design was used. In the first step, municipalities were selected, and primary sampling units generated. In the second step, addresses for randomly selected persons within the primary sampling units were drawn from population registers. Each sampled unit had the same inclusion probability (i.e., the sample was self-weighting). To allow for an experimental comparison of the data collection modes, postal addresses from the drawn sample were randomly split into three experimental groups (the group sample sizes varied due to project requirements).

The first experimental group (drawn sample size = 5,338) consisted of the traditional face-to-face survey using the full EVS-questionnaire with an average length of 59 minutes necessary for completion. The fieldwork of the face-to-face survey was organized in two phases, to which the drawn sample was randomly allocated before the beginning of the fieldwork (phase one: N=3,762 eligible addresses, phase two: N=1,576 eligible addresses). In the first phase, a 10€ postpaid incentive was announced in the initial cover letter (N=814 realized interviews). For the

second phase (starting in January 2018), a 5€ prepaid incentive was used to increase survey

participation in this group (N=570 realized interviews). After an evaluation of the first phase, it

was decided to offer a 20€ postpaid incentive to the remaining nonrespondents (N=110 realized

interviews).

The second group (drawn sample size = 1,913) consisted of a mixed-mode (web/mail)

survey using the same full-length EVS questionnaire which – compared to the face-to-survey –

yielded a comparable, but slightly shorter, average interview duration of 55 minutes (for the web

interview). The third group (drawn sample size = 8,973) consisted of an additional mixed-mode

(web/mail) survey using a matrix questionnaire design, implemented to reduce the overall

response burden (Peytchev & Peytcheva, 2017; Raghunathan & Grizzle, 1995). The use of this

matrix design reduced the average questionnaire length for the self-administered mixed-mode

survey to 38 minutes (for the web interview).

The fieldwork of the face-to-face survey took place between October 2017 and April

2018. Participation in the face-to-face survey was comparatively low with a AAPOR response

rate 6 (The American Association for Public Opinion Research, 2016) of 28%, even though

respondents were offered monetary incentives. The fieldwork of the mixed-mode matrix survey

was conducted between November 2017 and March 2018 and was implemented in a responsive

design with two phases (Wolf et al., 2021). Here, the drawn sample was randomly allocated to

the two phases. In a first phase, different incentive strategies (5€ prepaid, 10€ postpaid) and

mode choice sequences (simultaneous, sequential) were experimentally implemented. Based on

an evaluation of the phase one results, in January 2018 the second phase of the matrix survey

started for which respondents were provided with a 5€ prepaid incentive in a concurrent mode

choice sequence (i.e. offering both the mail questionnaire and the web questionnaire right from

the beginning). The response rate of the matrix survey was 36.1%. On average, 73% of the

respondents participated via mail mode (27% via web).

Encouraged by this outcome, it was decided to also field the full-length EVS-

questionnaire in a self-administered mixed-mode survey with a 5€ prepaid incentive and

concurrent mode choice sequence. The fieldwork period for the full mixed-mode survey was

between September and November 2018 and resulted in a response rate of 35.3%. In this survey,

83% of the respondents participated via mail mode (17% via web).

Questionnaire Design

When designing the questionnaires for the data collection experiment, the general aim

has been to make the questionnaires as comparable as possible while making the necessary

mode-specific adaptions.[1] While the formulations of the questions and answer scales remained

largely unchanged, question/interviewer instructions were sometimes reformulated to fit the self-

administered context. As the European Values Study lacks complex filtering, question order was

the same across designs. Yet, there are also several inherent differences in the modes such as

"don't know"-options or using grid questions (de Leeuw, 2018, pp. 82-83). The EVS experiment

optimized the usability of the survey for each mode, so question batteries were presented/read

out item-by-item in the face-to-face mode and for mobile devices, while grids were used in the

mail mode and on large screens in the web mode.

The second major difference relates to the usage of "don't know" answer categories.

Following the best practice (in Germany), "don't know" options were not mentioned by the

---

[1] All variants of the German questionnaires are available for download in the country-specific
documentation of the EVS.

interviewers and were not included in the show cards. Yet, "don't know" was still introduced as

a response alternative for the interviewers if respondents mentioned this option spontaneously.

For the web and mail modes, explicit "don't-know" options were provided for most questions.

Attitudinal Scales

We analyzed 140 questions from 21 item batteries with three or more questions, covering all

major topics of the EVS. Out of these 21 item batteries we identified 25 scales through

Confirmatory Factor Analysis (see section on equivalence testing), which we analyze throughout

this study. We define a scale here as a group of items that measure the same latent concept. A

short description of the scales and the years when they were introduced in the EVS is provided in

Table 1. A longer description of the scales, including number of items, scale points, and the use

of showcards is given in Table A1 of the Appendix.

Table 1. Description of scales used in the analysis

Scale	Description	Est. in EVS
Action	Unconventional political participation	1981
Belong	Closeness to the world vis-à-vis their municipality	2017*
Childhood	Political sophistication of parents when they were 14 years old	2008
Concern	Compassion for foreigners	1999
Concern_grp	Compassion for vulnerable groups	1999
Democracy1	Democratic attitudes	2017*
Democracy2	Authoritarian attitudes	2017*
Elections	Assessment of the integrity of elections	2017*
Environment	Importance of environmental protection	1981-2017
European	Meaning of being European	2017
Immigration	Attitudes toward immigration	2008
Importance	Important in life	1990
Marriage	Conditions for a successful marriage	1981
National	Meaning of being German	2008
Norms1	Attitudes towards cheating and probity	1981-2017
Norms2	Liberal vs conservative values	1981-2017
Pol_system	Support for democracy	1999

Pol_watch	Frequency of news consumption	2017
Policy	Attitudes on redistribution and the welfare state	1990
Society	Importance of societal provisions	1990, 2017
Surveillance	Attitudes about state surveillance	2017
Traditional_family	Attitudes towards traditional family roles	2008-2017
Trust	Confidence in public and political institutions	1981-2017
Trust_pl	Interpersonal trust	2017*
Work	Importance of work in society	2008

Adopted from the World Values Study (WVS) in 2017.

Weighting for selection

We use weights to account for possible differences between the face-to-face, mixed-mode full-

length, and mixed-mode matrix samples due to selection. Iterative proportional fitting (Deming

& Stephan, 1940) was used to adjust the distribution of the sample to the target population's

distributions with respect to sex, age, education, household size, citizenship, and religion.

Reference distributions for the target population were provided by the German Federal Statistical

Office.

Methods

Data quality indicators

We computed several indicators to assess data quality in each sample – both at the item level and

at the scale level. These indicators capture response behaviors that could be the result of not

completing every step of the cognitive response process (Tourangeau et al., 2000) or due to

respondents attempting to reduce their perceived response burden (Krosnick, 1991, 1999).

To measure item nonresponse, we calculated "don't know" and "no answer" as the

proportion of all eligible items for which respondents have chosen one of these residual

categories. To measure response style, we estimated indicators for extreme response style,

primacy, recency, mid-point response style, and straightlining. Primacy refers to the tendency to

endorse the first answer categories, while recency is defined as the tendency to select the last

spoken/shown answer categories. We define an extreme response style as the tendency to use the

highest or lowest categories of a response scale – regardless of the content of the question (Van

Vaerenbergh & Thomas, 2013), while a mid-point response style is defined as the tendency to

select a middle alternative (Narayan & Krosnick, 1996). Straightlining refers to minimal

differentiation between responses within item batteries (Roßmann et al., 2018). We employed a

simple non-differentiation measure (Kim et al., 2019), which reflects the proportion of questions

where only a single response category was chosen for all respective items, ranging from zero to

one. Not every response style indicator could be estimated for each scale depending on the

number of scale points (3, 4, 5, 10, 11; see Table A2 in the Appendix).

Turning to the outcomes of the three data collection modes, we start our investigation by

comparing the distribution of the question and the scales across experimental groups. For the

calculation of the scales, we estimated average indices. For comparability, we rescaled all

questions and scales to a range from 0 to 1. For the analysis we compare mean values and

standard deviations. We performed unpaired t-tests and applied Bonferroni corrections to

counteract the problem of the large number of comparisons. If a difference reaches the $p<0.05$

level we plot this as an additional information alongside the comparisons of means. In a similar

manner, we performed variance-ratio tests for the quality of variances, also applying Bonferroni

corrections. To gauge the effect size of mean differences we calculated Cohen's D for each

question and scale and plotted the results in histograms.

Equivalence testing

In addition to comparing data quality using the indicators mentioned above we also use equivalence testing to identify potential measurement differences across mode designs and questionnaire lengths. Measurement equivalence is useful for two main reasons: first, it can tell us whether we can validly compare results collected in different groups. This is especially useful when a traditional face-to-face survey is switching to a new design and comparisons over time are important. It is also useful when data are collected from different groups using different designs and comparisons or combinations of these data are needed (such as in cross-cultural research). Second, it can indicate differences in data quality and potential causes for these differences.

In order to carry out the equivalence testing we identified items that are part of rating scales with at least three items and that have ordinal response categories. We initially identified 144 items covering 14 topics. We then ran a series of Confirmatory Factor Analysis (CFA) on groups of items that measure the same concept in order to evaluate their overall fit and ensure they can be used for further testing. In total 8 items were excluded at this stage either because they had small loadings or correlations with the other items in the scale (belong/national identify: v168; parents: v270, v274; concern_grp: v219; national: v190; importance: v1, v5, v6). Thus, in the final CFA models we were left with 132 items divided into 25 scales. The full list of items and the scales they belong to are presented in Table A1 of the Appendix, while the scales and their overall model fits can be seen in Table 2.

Table 2. Fit statistics for the scales analyzed based on a simple confirmatory factor analysis using the entire sample.

Scale	# response categories	Chi^2	df	P-value	CFI	RMSEA
Policy	10	9.2	5	0.10	1.00	0.02

	4	4598.4	135	0.00	0.93	0.08
Trust	4	4598.4	135	0.00	0.93	0.08
Democracy1	10	150.4	5	0.00	0.94	0.09
Democracy2	10	5.1	2	0.08	1.00	0.02
Pol_system	4	10.6	2	0.01	1.00	0.03
Norms1	10	82.4	5	0.00	0.97	0.06
Norms2	10	1210.7	35	0.00	0.88	0.08
Belong	4	629.4	2	0.00	0.92	0.29
Elections	4	985.0	20	0.00	0.93	0.12
Importance*	4	0.00	0	0.00	1.00	0.00
Immigration	10	5.4	2	0.07	1.00	0.02
National	4	199.0	2	0.00	0.98	0.16
European	4	96.5	2	0.00	0.99	0.11
Environment	5	134.0	5	0.00	0.96	0.08
Surveillance*	4	0.0	0	0.00	1.00	0.00
Pol_watch	5	34.1	2	0.00	0.98	0.07
Concern	5	29.03	3	0.00	0.998	0.048
Concern_grp*	5	0.0	0	0.00	1.00	0.00
Society	4	80.7	2	0.00	0.92	0.10
Childhood	4	1089.2	9	0.00	0.89	0.18
Trust_pl	4	1021.6	9	0.00	0.99	0.17
Work	5	34.7	5	0.00	0.99	0.04
Marriage	3	226.0	9	0.00	0.97	0.08
Traditional_family*	5	0.0	0	0.00	1.00	0.00
Action	3	72.5	2	0.00	0.99	0.10

models just identified and fit cannot be estimated

In order to test whether the measurement of these concepts is equivalent across groups

(single-mode vs. mixed-mode and mixed-mode long vs. mixed-mode matrix), we compare a

series of three multi-group CFA models with increasing restrictions (Baumgartner & Steenkamp,

2006; Davidov et al., 2014):

- **Configural model**. The structure of the CFA is the same between groups but all

  coefficients are allowed to be different. This model is used as the reference for

  subsequent models (an example is shown in Figure 1).

- **Metric model**. The loadings (λ in Figure 1) are restricted to be the same across groups. If this model is not significantly worse compared to the previous (configural) model then the variances of the latent variable can be compared across groups.

- **Scalar model**. The intercepts/thresholds (τ in Figure 1) are restricted to be the same across groups. If this model is not significantly worse than the previous one it implies that the means of the latent variables are comparable across groups.

The metric and the scalar models give information about different types of measurement error. The metric model refers to the covariance between the questions and the latent variable. In general, a higher loading (λ in Figure 1) highlights a stronger relationship with the latent variable and can be an indication of higher reliability (Bollen, 1989). As a result, lack of metric equivalence could indicate differences in reliability across groups. The scalar model, on the other hand, can be informative regarding systematic shifts in the responses, as it tests if the conditional average of each question is the same across groups. If scalar equivalence is not met it can indicate that the averages of the questions across groups (even after controlling for the scores on the latent variable) are different. This could be caused by systematic response styles like social desirability, acquiescence, and primacy or recency.

When equivalence is not found (metric and scalar models fit worse than the less restrictive model) it is also possible to investigate partial equivalence (Byrne et al., 1989). This implies running a series of models to find which variables can be consistently measured across groups. If at least two variables are found to be equivalent it might still be possible to compare the means and the variances of the latent variables (Byrne et al., 1989).

For all scales we run simple CFA models for equivalence testing with one latent variable explaining the observed variables (similar to Figure 1). We treat variables with five or more categories as continuous and those with less than five as categorical. All analyses are weighted (see description of selection weights above). For the continuous variables we use Maximum Likelihood Robust (MLR) estimation while for the categorical ones we use the Weighted Least Squares Means and Variance adjusted estimator with Theta parametrization. We fix the means of the latent variables to 0 to aid estimation and the loading of the first variable to 1.
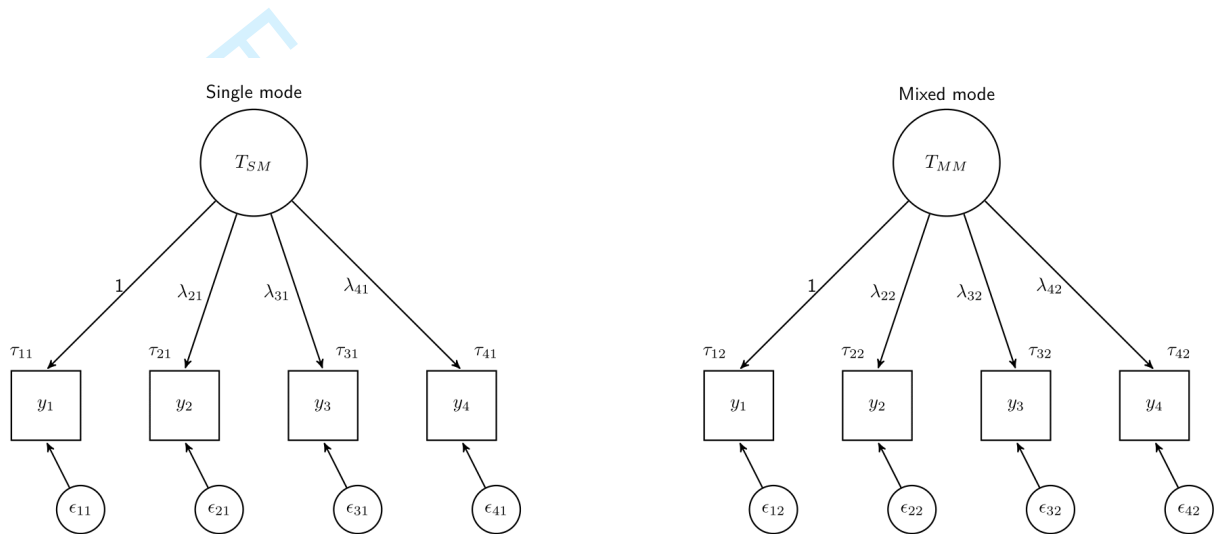
Figure 1. Visual representation of equivalence testing across two groups (single-mode vs. mixed-mode). The configural model is represented where the factor structure is the same but the coefficients are allowed to be different across groups. The large circle represents the overall concept or latent variable, the squares are the observed variables, and the small circles are the residuals. λ denotes the loadings or regression coefficients of the relationship between the latent and observed variables, while τ denotes the intercept/thresholds of these relationships.

One model had issues during equivalence testing. For the "traditional family" scale MLR with weights leads to estimation issues, so we use ML with no weights instead. Thus, in total we investigate 150 models (25 scales * 6 models; 3 models for single- vs. mixed-mode and 3 models for mixed-mode long vs. mixed-mode matrix) for equivalence testing. To compare model fit we use the cut-off value of 0.01 for the change in the Comparative Fit Index (ΔCFI) to define a

model as lacking equivalence (Chen, 2007). The data were cleaned in Stata and R, with the data

quality indicators created in Stata and the equivalence testing performed in Mplus.

Results

Comparison of means and distributions

We start by investigating differences in the estimated means and standard deviations of all 140

items initially used in the CFA models (RQ1), excluding only the scale religion for reasons of

consistency with the following analysis. The first row in Figure 2 shows scatterplots of the

standardized means for single vs. mixed-mode and mixed-mode long vs. mixed-mode matrix

(short) designs. We observe some significant differences especially in the single- vs. mixed-

mode comparison. This is confirmed in the histogram of Cohen's *d* (third row in Figure 2). Here

we see that while many of the differences are in the +/- .2 range (indicating small effects) there is

a group of variables that have larger averages in face-to-face compared to the mixed-mode

(mail/web) design and have moderate effect sizes between 0.2 and 0.6. This seems to be much

less the case when comparing the mixed-mode data by questionnaire length.

The comparison of standard deviations (second row in Figure 2) shows fewer significant

differences although a few variables seem to have significantly higher standard deviations in the

mixed-mode design compared to the face-to-face one. Again, the differences are less pronounced

between the mixed-mode long and the mixed-mode matrix surveys. The same patterns emerge

when we replicate the analysis at the scale level for all 25 scales in Figure 3. There are some

significant differences between the face-to-face mode and the mixed-mode long survey and only

few differences between the mixed-mode long and the mixed-mode matrix (short) survey.

We also compare the means and standard deviations for the different survey designs by question topic and scale type (Table A3 in the Appendix), and find some indications of differences. Topics such as national identity, politics and society, and religion and morale, and 4-point scales have more differences in the single- vs. mixed-mode comparison, but not for the mixed-mode long vs. mixed-mode matrix comparison. However, we need to be careful when interpreting these findings because of the relatively small sample size at the item level and the clustering within scales.
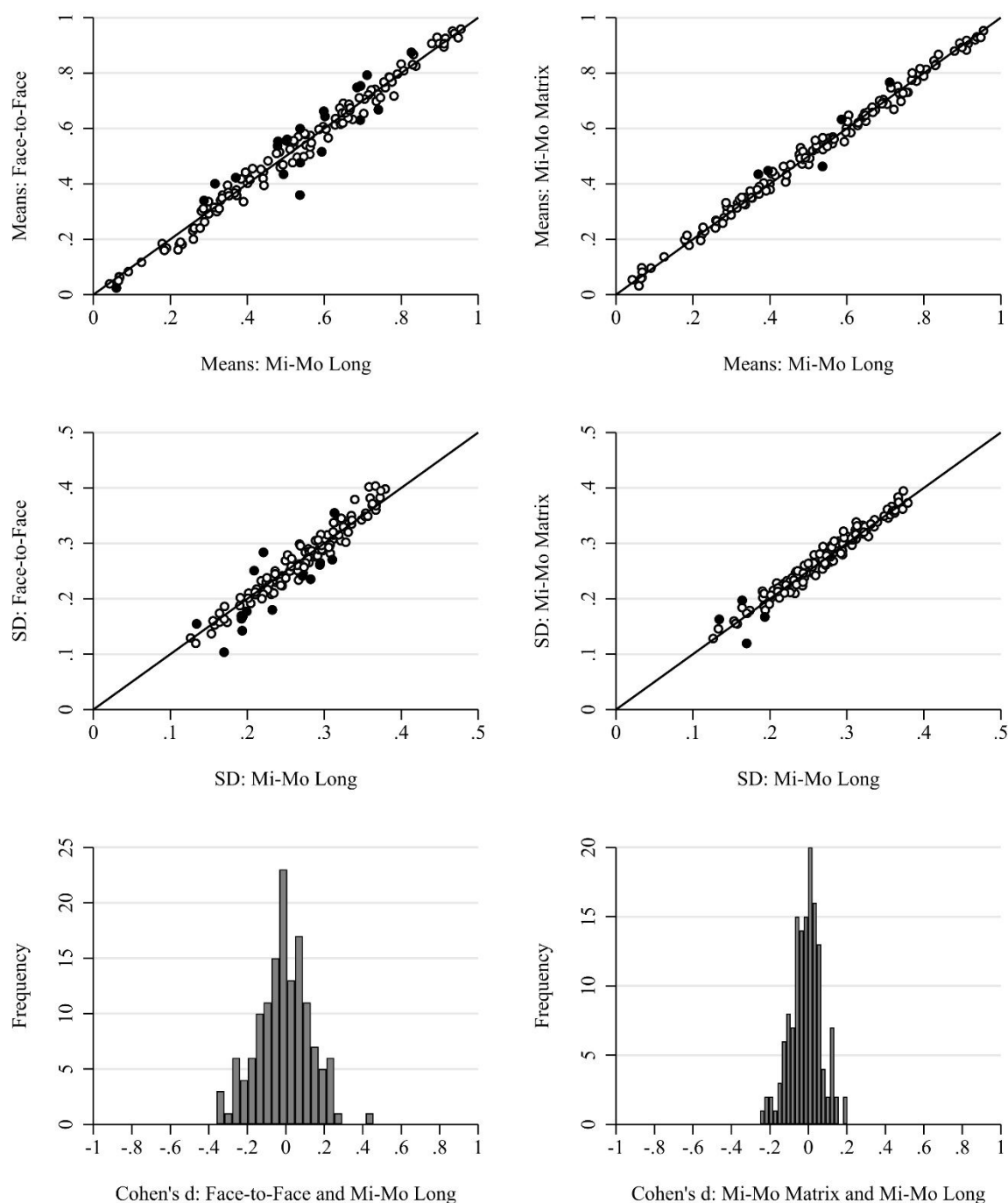
Figure 2. Item level scatterplots for means and standard deviations by survey design: face-to-face single mode, mixed-mode long, and mixed-mode matrix (short). Filled points are statistically significantly different after Bonferroni correction. Bottom row: histogram of Cohen's d for differences between groups. The diagonal lines are used to gauge whether values are higher or lower in one group or another.
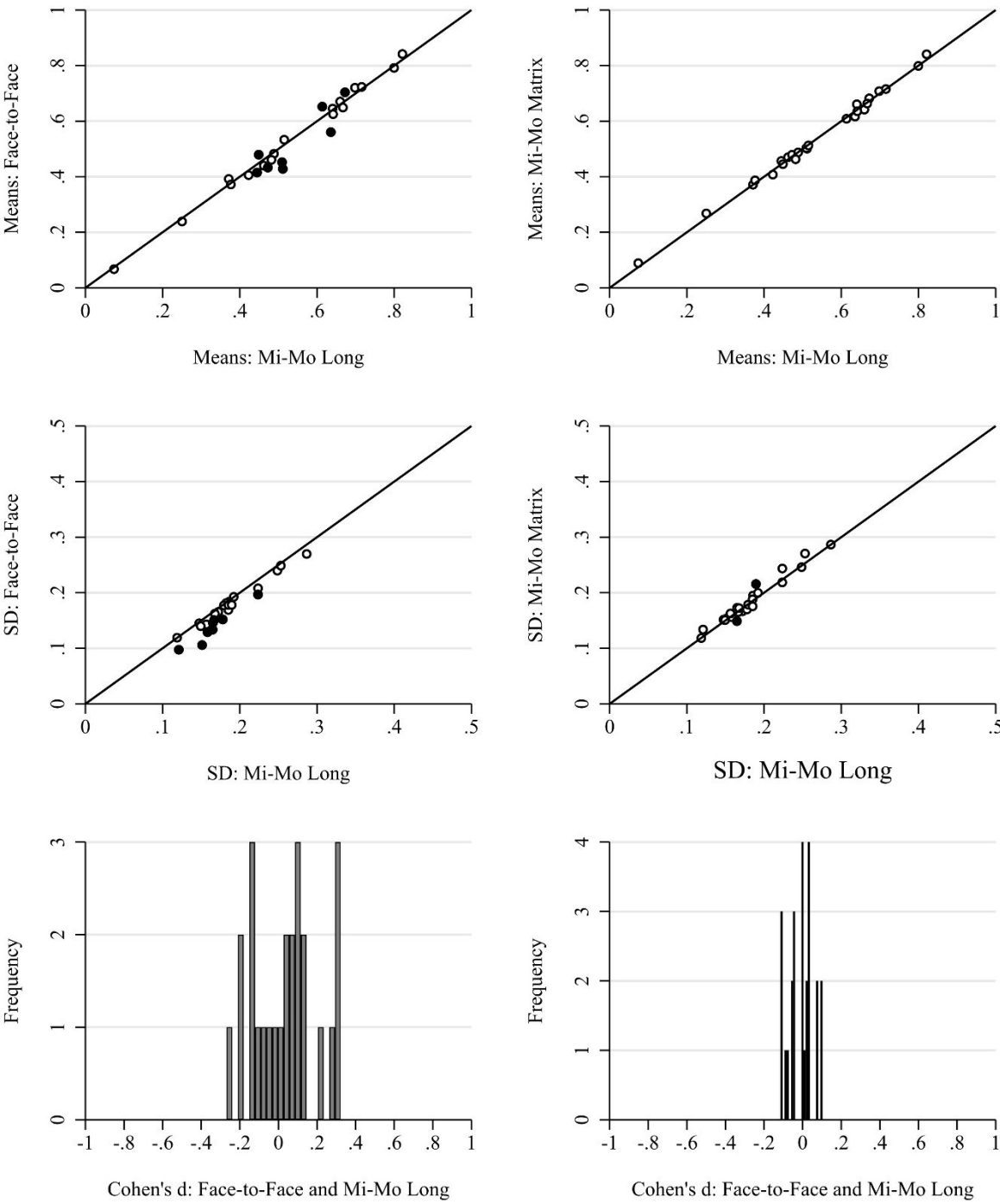
Figure 3. Scale level scatterplots for means and standard deviations by survey design: face-to-face single mode, mixed-mode long, and mixed-mode matrix (short). Filled points are statistically significantly different after Bonferroni correction. Bottom row: histogram of Cohen's d for differences between groups.

Figure 4. Item level average quality indicators with 95%-confidence intervals by group: face-to-face single mode, mixed-mode long, and mixed-mode matrix (short).

Comparison of data quality indicators

We next estimate multiple data quality indicators (RQ2). Figure 4 shows the average of each data quality indicator as well as the confidence interval based on all the items separately by the three groups of interest: face-to-face, mixed-mode long, and mixed-mode matrix. Starting with item nonresponse, we see that the rate is significantly higher in the mixed-mode designs (long and short) compared to the face-to-face design (i.e. confidence intervals do not overlap). This is true both for the "no answer" and "don't know" responses. The difference between the mixed-mode long and mixed-mode matrix designs is not significant for the "no answer" response. The "don't

know" responses appear to be less frequent in the mixed-mode matrix survey than in the mixed-mode long survey. For the other data quality indicators (primacy, recency, middle answer and extreme response style), there are only small differences between the face-to-face, mixed-mode long, and mixed-mode matrix surveys. While there appear to be slightly higher levels of recency for the interviewer-administered survey then for the self-administered mixed-mode survey, the picture is reverted for the use of middle answers with lower levels for the face-to-face survey and somewhat higher levels for the mixed-mode surveys.

We next estimated quality indicators for the 25 scales. Figure 5 mirrors the analysis performed in Figure 4, but also adds an indicator for straightlining, which can only be computed at the scale level. Results are very similar to those presented at Figure 4: item nonresponse is significantly more frequent in both mixed-mode designs than in the face-to-face mode, and the mixed-mode matrix shows a smaller extent of "don't know" responses than the mixed-mode long. For the remaining data quality indicators differences between the surveys are small, although the face-to-face mode appears to perform slightly better than both mixed-mode designs with respect to straightlining and middle response style, while the mixed-mode surveys perform slightly better with respect to recency.
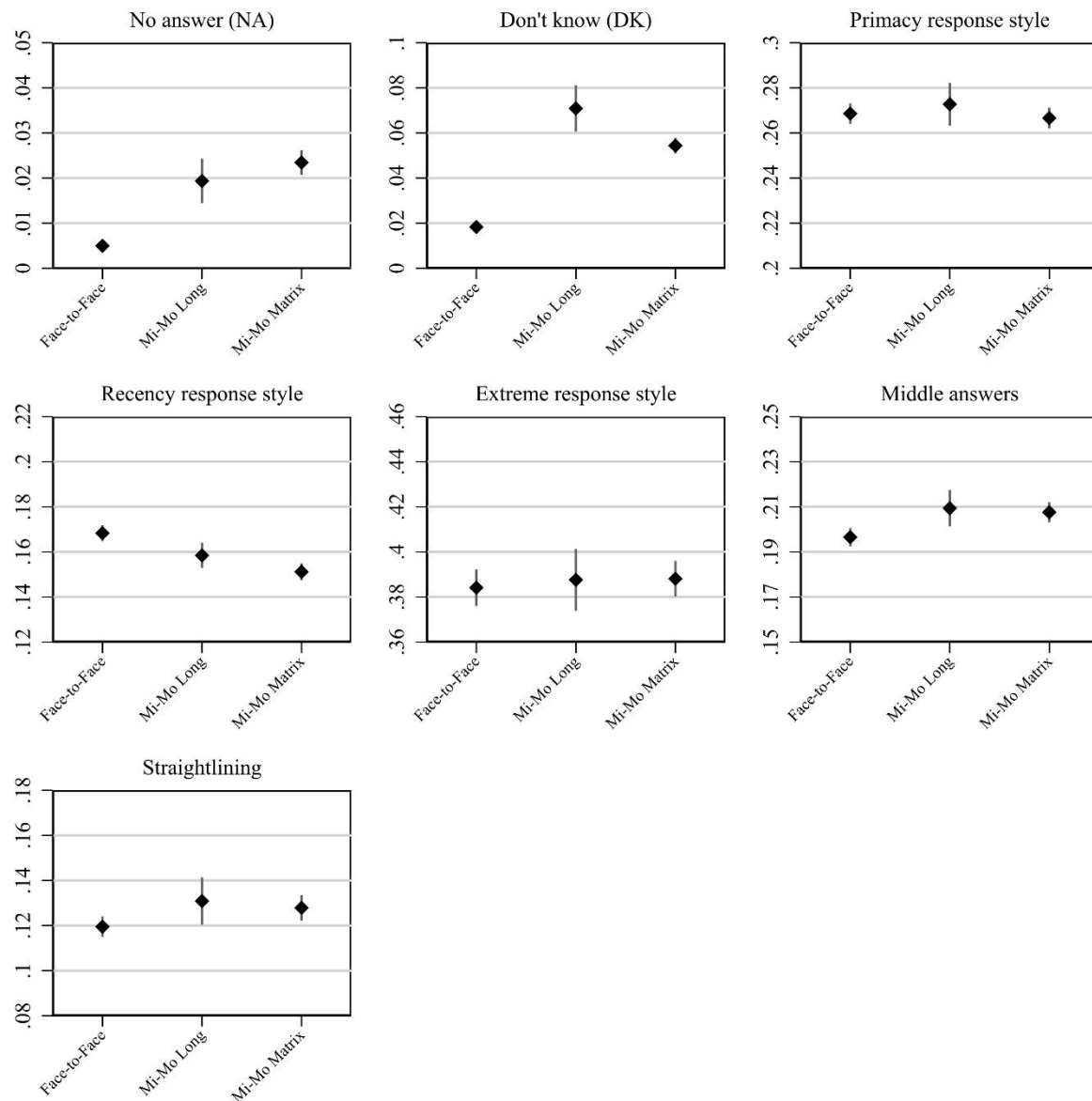
Figure 5. Scale level average quality indicators with 95%-confidence intervals by group: face-to-face single mode, mixed-mode long, and mixed-mode matrix (short).

Lastly, Figure 6 shows the data quality indicators separately for each scale. When turning

to this lower level of aggregation we see the results of the previous analysis reconfirmed: the

mixed-mode designs consistently show greater proportions of "no answer" and "don't know"

responses across all scales. These proportions also vary between the scales and the magnitude of

the difference is quite large for several scales. Yet, these differences in item nonresponse seem to

be only weakly related to differences in the scale averages (top-left row). Again, we find no

indication for a difference in extreme response style between the surveys (bottom-left row).

Finally, the mode designs seem to be only weakly and inconsistently related to the use of middle

answers, primacy, recency and straightlining; thus, the results remain consistent between the

analyses presented in Figures 5 and 6.

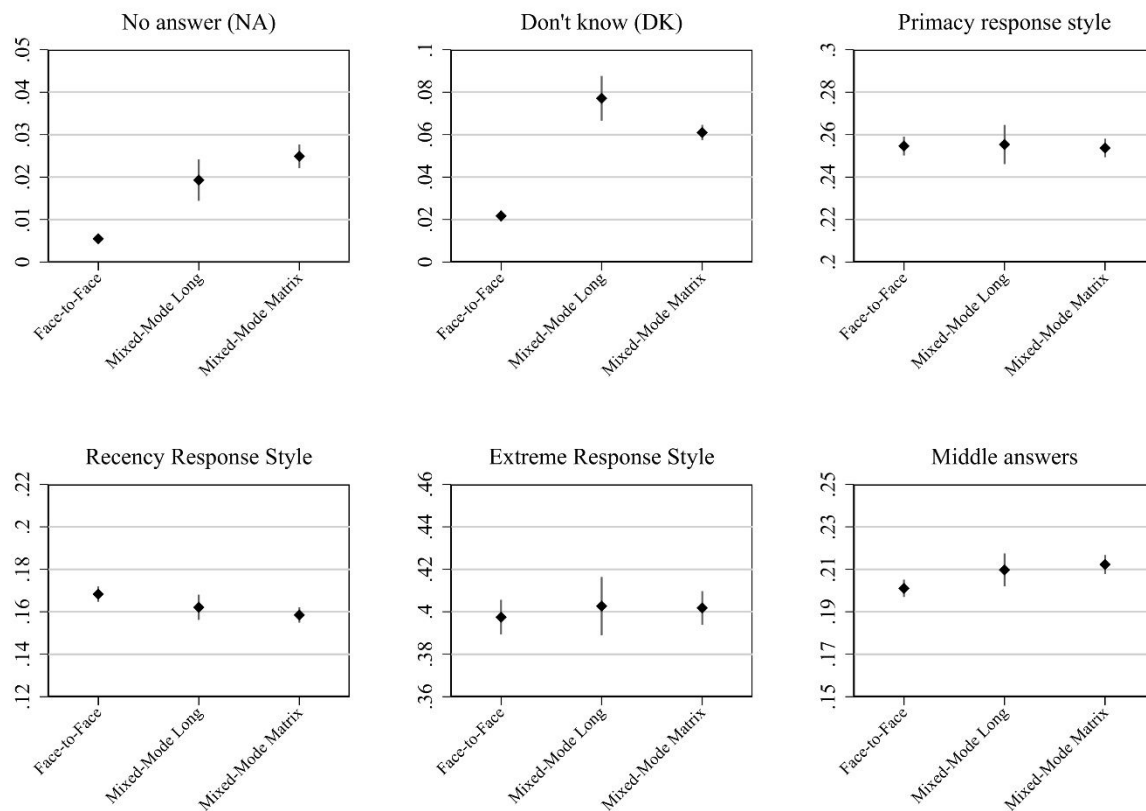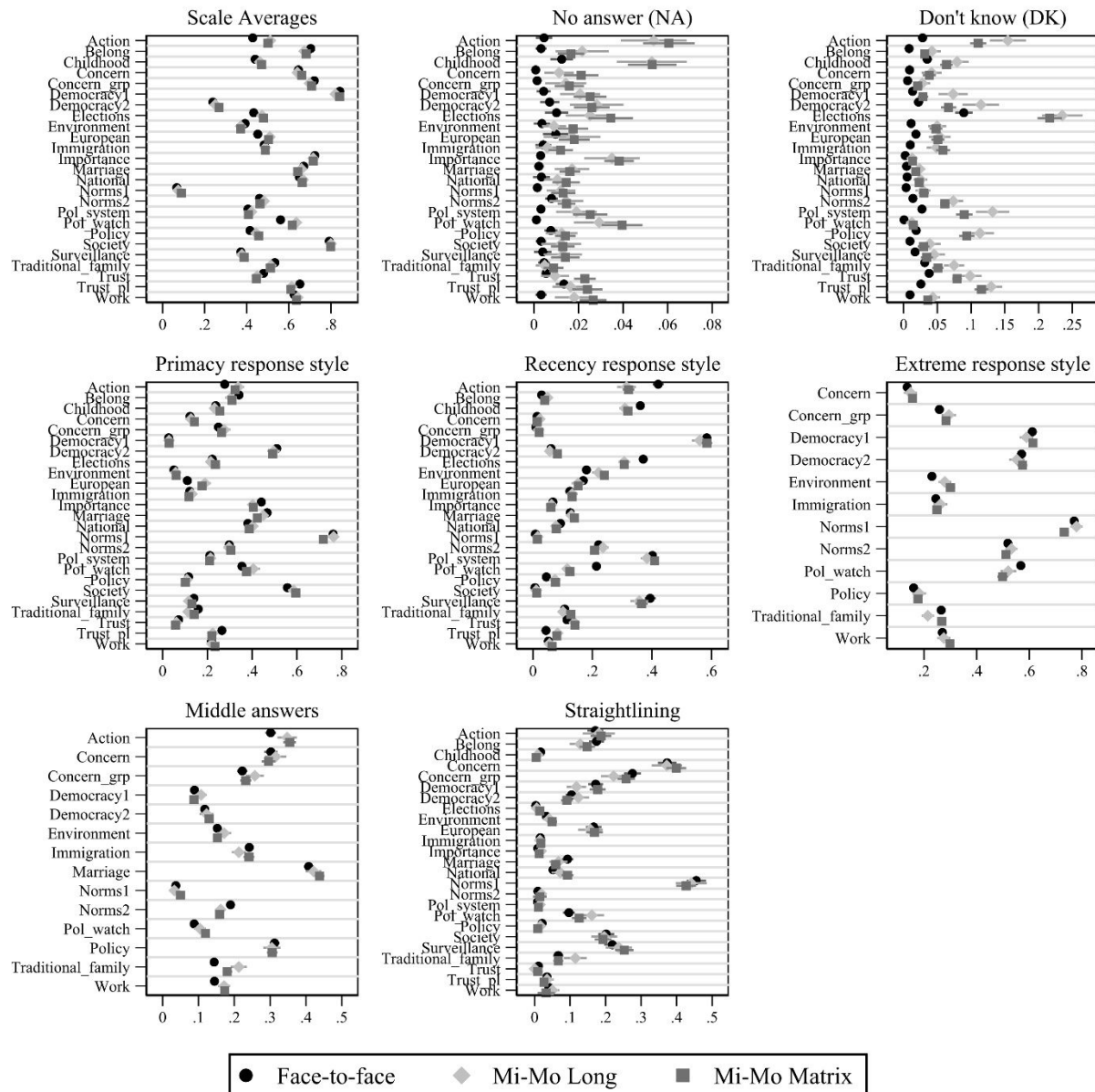*Figure 6. Scale level quality indicators with 95%-confidence intervals by group: face-to-face single mode, mixed-mode long, and mixed-mode matrix (short).*

Equivalence testing: Mode design comparison

We next investigate whether the 25 scales identified have comparable measurement models (i.e. measurement equivalence) across face-to-face and mixed-mode designs and mixed-mode long and mixed-mode matrix designs (RQ3). We start by testing whether the three levels of

measurement equivalence (configural, metric, and scalar) hold across the single-mode face-to-face survey and the mixed-mode mail/web survey, for each of the 25 scales. Measurement equivalence is assessed by inspecting differences in the CFI values between models. Full details of the model fit statistics are provided in the online supplementary materials. The models for one of the scales ("importance") did not converge when analyzing it using the multi-group approach. The following results will refer to the remaining 24 scales.

Out of 24 scales analyzed the vast majority (21 out of 24) achieve metric equivalence based on our criterion (see Table 3, "Metric" column). This suggests that valid comparisons of unstandardized relationships between the interviewer- and self-administered mode groups can be obtained for most scales. An alternative interpretation of equal loadings is that reliabilities are similar across the groups (Bollen, 1989).

The three scales that do not reach metric equivalence are Norms1, Pol_watch, and Traditional_family. For Norms1, the largest difference appears for the loadings of items v152 (if it is justified: "Cheating on tax if you have the chance") where the loading is larger in the mixed-mode design and v162 (if police violence is justified) where the reverse is true. For Pol_watch, the loading for v211 (how often you follow politics on social media) has a much smaller loading in the mixed-mode compared to the single-mode design (0.40 vs. 0.07). Similarly, the loadings are less strong in the mixed-mode data compared to the face-to-face data for v83 ("It is a duty towards society to have children") and v84 ("Adult children have the duty to provide long-term care for their parents"). This leads to a mixed result for these scales although overall the loadings are generally stronger in the face-to-face compared to the mixed-mode design, indicating slightly higher reliability in the interviewer-administered mode for these items.

Scalar equivalence is also achieved for the majority (15 out of 24) of the scales (see Table 3, "Scalar" column), indicating that it is valid to compare the latent means between mode designs. Out of the nine scalar non-equivalent scales, two of them overlap with the metric non-equivalent scales: Pol_watch and Traditional_family. These two scales are particularly problematic as they have different loadings and different intercepts and therefore are not measured in the same way across the mode designs. The other scalar non-equivalent scales include: Concern_grp, both Democracy scales, Environment, Norms2, Policy, and Work. We dissect some of the causes of scalar non-equivalence in these scales below.

Scalar equivalence is essential for ensuring that averages of the latent variables are comparable across groups. Additionally, differences in intercepts/thresholds can be informative as they may be caused by systematic biases such as social desirability or acquiescence. The differences found in the nine scales that are not scalar equivalent show mixed patterns, with larger intercept/threshold values in both mode design groups for particular items. For example, for items v136 ("it's essential for democracy for people to receive state aid" where larger numbers mean more essential) and v211 ("how often you follow politics on social media" where larger numbers mean less often) the intercepts are larger in the mixed-mode design. In contrast, for items v156 ("Euthansia can be justified" where larger numbers mean can be justified) and v154 ("Abortion can be justified" where larger numbers mean can be justified) the conditional mean is larger in the face-to-face single-mode design. At this point we can only speculate regarding possible causes for such differences, such as social desirability, leading to systematic shifts in the observed averages of these variables. What is clear is that systematic differences in the intercepts can happen across mode designs.

Table 3. Equivalence Testing Conclusions Based on CFI Differences – Mode Design

Comparison. Problematic items are denoted in parentheses.

Scale	Metric	Scalar
Action	✓	✓
Belong	✓	✓
Childhood	✓	✓
Concern	✓	✓
Concern_grp	✓	✗ (v217, v218, v220)
Democracy1	✓	✗ (v133, v135, v136, v138, v141)
Democracy2	✓	✗ (v134, v137, v139, v140)
Elections	✓	✓
Environment	✓	✗ (v199-v203)
European	✓	✓
Immigration	✓	✓
Marriage	✓	✓
National	✓	✓
Norms1	✗ (v150, v152, v159, v162)	✓
Norms2	✓	✗ (v151, v153-v158, v160, v161, v163)
Pol_system	✓	✓
Pol_watch	✗ (v209-v211)	✗ (v208-v211)
Policy	✓	✗ (v103-v107)
Society	✓	✓
Surveillance	✓	✓
Traditional_family	✗ (v83, v84)	✗ (v82-v84)
Trust	✓	✓
Trust_pl	✓	✓
Work	✓	✗ (v46-v50)
TOTAL	21/24	15/24

Equivalence testing: Questionnaire length comparison

Next, we test for measurement equivalence across the long- and short-questionnaire groups used

in the mixed-mode (web/mail) surveys. The same model evaluation criterion (i.e. change in CFI

values) is used. Full model details can be found in the online supplementary materials.

Metric equivalence is achieved for all but two scales (Policy and Traditional_family) (see

Table 4, "Metric" column). Again, this means that one can be relatively assured that

unstandardized relationships can be validly compared between both long- and short-questionnaire versions. The full-length questionnaire version shows stronger loadings for items v83 ("It is a duty towards society to have children"), v84 ("Adult children have the duty to provide long-term care for their parents"), v107 ("Private ownership of business and industry should be increased"), and v105 ("Competition is good"), while the shorter questionnaire version has stronger loadings for item v106 ("Incomes should be made more equal"). As mentioned above, this indicates that for some items the full-length questionnaire has higher reliability compared to the shorter length version, although this difference appears only for a small proportion of items.

Scalar equivalence is established for all but five scales (see Table 4, "Scalar" column), supporting the possibility of making valid comparisons of latent means between different questionnaire lengths. Among the non-equivalent scales, Policy, also lacks metric equivalence, suggesting that this scale cannot be measured in the same way in both long- and short-questionnaire forms. The other scalar non-equivalent scales include Democracy1, Norms1, Pol_watch, and Work.

Looking at the largest differences in the intercepts for these non-equivalent scales we find that items v136 ("it's essential for democracy for people to receive state aid" where larger numbers mean essential), v104 ("People who are unemployed should have the right to refuse a job they do not want" where larger numbers mean more support for statement), v211 ("how often you follow politics on social media" where larger numbers mean less often), v133 ("Governments tax the rich and subsidize the poor" where higher numbers mean essential for democracy), v149 ("Claiming state benefits which you are not entitled to" where higher numbers mean always justified), and v138 ("Civil rights protect people from state oppression" where

higher numbers mean essential for democracy) have higher intercepts in the shorter questionnaire compared to the longer one. Once again, a number of different types of systematic biases could lead to these patterns such as social desirability or recency effects.

Table 4. Equivalence Testing Conclusions Based on CFI Differences – Questionnaire Length Comparison. Problematic items are denoted in parentheses.

Scale	Metric	Scalar
Action	✓	✓
Belong	✓	✓
Childhood	✓	✓
Concern	✓	✓
Concern_grp	✓	✓
Democracy1	✓	✗ (v133, v135, v136, v138, v141)
Democracy2	✓	✓
Elections	✓	✓
Environment	✓	✓
European	✓	✓
Immigration	✓	✓
Marriage	✓	✓
National	✓	✓
Norms1	✓	✗ (v149, v150, v152, v159, v162)
Norms2	✓	✓
Pol_system	✓	✓
Pol_watch	✓	✗ (v208-v211)
Policy	✗ (v104-v107)	✗ (v103-v107)
Society	✓	✓
Surveillance	✓	✓
Traditional_family	✗ (v83, v84)	✓
Trust	✓	✓
Trust_pl	✓	✓
Work	✓	✗ (v46-v50)
TOTAL	22/24	19/24

Discussion

In this study, we investigated two important aspects of modern survey data collection: differences in measurement quality between a single-mode face-to-face design and a fully self-administered mixed-mode (mail/web) design, and differences in measurement quality between a long and short questionnaire implemented in a self-administered mixed-mode (mail/web) design. Overall, we found small differences in the distributions of variables across these different designs (RQ1), although we observed larger differences in estimated means and lower variation for items measured in the single-mode face-to-face design compared to the mixed-mode (mail/web) designs. We also observed more item missing data in the mixed-mode designs compared to the single-mode design but few systematic differences in data quality at the item level (RQ2). Moreover, there were only few differences at the item level when comparing the mixed-mode surveys with the different questionnaire lengths.

We next investigated measurement equivalence (RQ3), or the degree to which the different survey designs have the same measurement model, in 24 attitudinal scales (covering 129 items). When comparing mode designs, we found that 21 out of 24 scales achieve metric equivalence and 15 out of 24 achieve scalar equivalence. When comparing questionnaire lengths, we found that 22 out of 24 scales reach metric equivalence and 19 out of 24 reach scalar equivalence. Thus, the majority of scales allow for comparing means and variances of latent variables across the different survey designs.

We believe these findings carry a positive message that implies relatively similar data quality and comparable data between face-to-face and mail/web mode designs and between long and short self-completion questionnaire versions. In this context, we view the implementation of a different mode design, namely shifting from an interviewer-administered design to a self-

administered one, to be more problematic than shifting from a longer to a shorter questionnaire length. This is largely consistent with previous literature showing larger differences in measurement quality between interviewer and self-administered modes (Cernat et al., 2016; Klausch et al., 2013).

There were some indications that a single-mode face-to-face design has slightly higher reliabilities than a self-administered mixed-mode design, and that a full-length self-completion questionnaire has higher reliabilities compared to a matrix questionnaire for some items. We also found that systematic differences in the conditional means are often present. One can only speculate on their potential causes and implications for data analysis. This is a topic for future research. That being said, the differences between the two mode designs are relatively small, which may also be due to the extensive use of showcards in the face-to-face mode.

We point out some limitations of this study. Namely, we examined the EVS data for only a single country. Examining the measurement quality of cross-national comparisons under different mode and questionnaire designs is a worthwhile topic for future research. Furthermore, alternative approaches to testing measurement equivalence and data quality could be considered. In the present study, we found differences in response rates between the surveys with the face-to-face survey having the lowest response rate. For a more detailed description of differences in sample composition between the designs, we refer to the study by Wolf et al. (2021). Based on our research questions, we focused on analyses of differences on measurement and utilized weighting to control for selective participation. However, future research could go one step further and try to feature both the measurement and the nonresponse perspectives simultaneously. Such a combined perspective might help to better understand the total survey error properties for a specific survey design that are likely to arise by motivating specific strata to

participate (or not) and the strata-specific measurement error this might introduce. Further, the

EVS data did not include all the possible design options, for example a short face-to-face survey,

although we do believe the utility and prevalence of such a design would be limited in practice.

Finally, we have also implemented incentive experiments (5€ prepaid vs 10€ postpaid) in the

EVS-Germany and further research could explore measurement implications between different

monetary incentive types (conditional vs unconditional) as existing research usually compares

the use of incentives against no incentive.

Nonetheless, the present study undertook a comprehensive analysis of more than 100

items from more than 20 scales in one of the largest international surveys used in the social

sciences. Furthermore, the study investigated data quality in a number of different ways:

comparing distributional properties at the item and scale level, comparing various data quality

indicators at the item and scale level, and measurement equivalence testing. Moreover,

comparing both the mode design and questionnaire length brings important insights for survey

practice. Often, shifting from single-mode or interviewer-administered data collection to self-

completion, especially web based, necessitates further adaptations that can be important. For

example, web and mobile data collection may have questionnaire length or interview duration

constraints that can impact how the survey is adapted. As such, the switch from a long face-to-

face survey to a shorter web survey seems to be especially prevalent. Our findings can help

survey practitioners to assess which design alterations are likely to result in more measurement

differences compared to the previous design and which aspects of data quality may be affected.

Future research should aim to replicate our findings in other countries, other

questionnaires, and other survey designs. Of special interest would be to understand the extent of

the measurement differences between short and long questionnaires and to investigate potential

causes for such differences. Also of interest for data users would be to study the kinds of biases

they should expect if measurement equivalence (metric and/or scalar) does not hold across

survey designs. This is especially relevant for longitudinal surveys (panel or repeated cross-

sectional studies) that switch from face-to-face to self-completion or surveys where different

population groups are interviewed using different mode designs (e.g., different countries in

cross-cultural research).

References

Baumgartner, H., & Steenkamp, J.-B. E. M. (2006). An Extended Paradigm for Measurement

Analysis of Marketing Constructs Applicable to Panel Data. *Journal of Marketing*

Research (JMR), *43*(3), 431–442.

Beullens, K., Loosveldt, G., Vandenplas, C., & Stoop, I. (2018). Response Rates in the European

Social Survey: Increasing, Decreasing, or a Matter of Fieldwork Efforts? *Survey*

Methods: Insights from the Field (SMIF). https://doi.org/10.13094/SMIF-2018-00003

Bianchi, A., Biffignandi, S., & Lynn, P. (2017). Web-face-to-face mixed-mode design in a

longitudinal survey: Effects on participation rates, sample composition, and costs.

Journal of Official Statistics, *33*(2), 385–408.

Biemer, P. P., Harris, K. M., Liao, D., Burke, B. J., & Halpern, C. T. (2021). Modeling Mode

Effects for a Panel Survey in Transition. In *Measurement Error in Longitudinal Data* (A.

Cernat and J.W. Sakshaug (Eds.)). Oxford University Press.

Bollen, K. (1989). *Structural equations with latent variables*. Wiley-Interscience Publication.

Brick, J. M., & Williams, D. (2013). Explaining Rising Nonresponse Rates in Cross-Sectional

Surveys. *The ANNALS of the American Academy of Political and Social Science*, *645*(1),

36–59. https://doi.org/10.1177/0002716212456834

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor

covariance and mean structures: The issue of partial measurement invariance.

Psychological Bulletin, *105*(3), 456.

Callegaro, M., Manfreda, K. L., & Vehovar, V. (2015). *Web Survey Methodology*. SAGE.

Cernat, A., Couper, M. P., & Ofstedal, M. B. (2016). Estimation of Mode Effects in the Health

and Retirement Study Using Measurement Models. *Journal of Survey Statistics and*

Methodology, *4*(4), 501–524. https://doi.org/10.1093/jssam/smw021

Cernat, A., & Revilla, M. (2020). Moving from Face-to-Face to a Web Panel: Impacts on

    Measurement Quality. *Journal of Survey Statistics and Methodology*, *smaa007*.

    https://doi.org/10.1093/jssam/smaa007

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance.

    Structural Equation Modeling: A Multidisciplinary Journal, *14*(3), 464–504.

    https://doi.org/10.1080/10705510701301834

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement

    Equivalence in Cross-National Research. *Annual Review of Sociology*, *40*(1), 55–75.

    https://doi.org/10.1146/annurev-soc-071913-043137

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., Hox, J., & Luiten, A. (2018). International Nonresponse Trends across Countries

    and Years: An analysis of 36 years of Labour Force Survey data. *Survey Methods:*

    Insights from the Field, 1–11. https://doi.org/10.13094/SMIF-2018-00008

Deming, 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.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., & Messer, B. L. (2009).

    Response rate and measurement differences in mixed-mode surveys using mail,

    telephone, interactive voice response (IVR) and the Internet. *Social Science Research*,

    38(1), 1–18. https://doi.org/10.1016/j.ssresearch.2008.03.007

EVS. (2017a). *European Values Study 2017: Integrated Dataset – Matrix Design Data (EVS 2017 Matrix Design)*. GESIS Data Archive, Cologne. ZA7502 Data file Version 1.0.0, doi:10.4232/1.13092.

EVS. (2017b). *European Values Study 2017: Integrated Dataset (EVS 2017)*. GESIS Data Archive, Cologne. ZA7500 Data file Version 3.0.0, doi:10.4232/1.13511.

Galesic, M., & Bosnjak, M. (2009). Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey. *Public Opinion Quarterly*, *73*(2), 349–360. https://doi.org/10.1093/poq/nfp031

Gummer, T., Christmann, P., Verhoeven, S., & Wolf, C. (2021). Using Responsive Survey Design to Conduct Surveys Under Uncertainty. *Unpublished Manuscript*.

Gummer, T., Schmiedeberg, C., Bujard, M., Christmann, P., Hank, K., Kunz, T., Lück, D., & Neyer, F. J. (2020). The impact of Covid-19 on fieldwork efforts and planning in pairfam and FReDA-GGS. *Survey Research Methods*, *14*(2), 223–227. https://doi.org/10.18148/srm/2020.v14i2.7740

Heerwegh, D. (2009). Mode Differences Between Face-to-Face and Web Surveys: An Experimental Investigation of Data Quality and Social Desirability Effects. *International Journal of Public Opinion Research*, *21*(1), 111–121. https://doi.org/10.1093/ijpor/edn054

Heerwegh, D., & Loosveldt, G. (2011). Assessing mode effects in a national crime victimization survey using structural equation models: Social desirability bias and acquiescence. *Journal of Official Statistics*, *27*(1), 49.

Herzog, A. R., & Bachman, J. G. (1981). Effects of Questionnaire Length on Response Quality. *Public Opinion Quarterly*, *45*(4), 549–559. https://doi.org/10.1086/268687

Hope, S., Campanelli, P. C., Nicolaas, G., Lynn, P., & Annette, J. (2014). *The role of the interviewer in producing mode effects: Results from a mixed modes experiment comparing face-to-face, telephone and web administration* (ISER Working Paper Series No. 2014–20). Institute for Social and Economic Research. https://econpapers.repec.org/paper/eseiserwp/2014-20.htm

Kappelhof, J. W. (2015). Face-to-Face or sequential Mixed-Mode surveys among Non-Western minorities in The Netherlands: The effect of different survey designs on the possibility of nonresponse bias. *Journal of Official Statistics*, *31*(1), 1–30.

Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2019). Straightlining: Overview of Measurement, Comparison of Indicators, and Effects in Mail–Web Mixed-Mode Surveys. *Social Science Computer Review*, *37*(2), 214–233. https://doi.org/10.1177/0894439317752406

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. https://doi.org/10.1177/0049124113500480

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly*, *72*(5), 847–865. https://doi.org/10.1093/poq/nfn063

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*(1), 537–567.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order

effects in survey measurement. *Public Opinion Quarterly*, *51*(2), 201–219.

https://doi.org/10.1086/269029

Luijkx, R., Jónsdóttir, G. A., Gummer, T., Ernst Stähli, M., Frederiksen, M., Ketola, K.,

Reeskens, T., Brislinger, E., Christmann, P., Gunnarsson, S. Þ., Hjaltason, Á. B., Joye,

D., Lomazzi, V., Maineri, A. M., Milbert, P., Ochsner, M., Pollien, A., Sapin, M.,

Solanes, I., … Wolf, C. (2020). The European Values Study 2017: On the Way to the

Future Using Mixed-Modes. *European Sociological Review*, *jcaa049*.

https://doi.org/10.1093/esr/jcaa049

Lynn, P., Hope, S., Jäckle, A., Campanelli, P., & Nicolaas, G. (2012). *Effects of visual and aural

communication of categorical response options on answers to survey questions* (Working

Paper No. 2012–21). ISER Working Paper Series.

https://www.econstor.eu/handle/10419/65994

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude

measurement. *Public Opinion Quarterly*, *60*(1), 58–88.

Peytchev, A., & Peytcheva, E. (2017). Reduction of Measurement Error due to Survey Length:

Evaluation of the Split Questionnaire Design Approach. *Survey Research Methods*, *11*(4),

361–368. https://doi.org/10.18148/srm/2017.v11i4.7145

Raghunathan, T. E., & Grizzle, J. E. (1995). A Split Questionnaire Survey Design. *Journal of the

American Statistical Association*, *90*(429), 54–63. https://doi.org/10.2307/2291129

Revilla, M., & Ochoa, C. (2017). Ideal and Maximum Length for a Web Survey. *International

Journal of Market Research*, *59*(5), 557–565. https://doi.org/10.2501/IJMR-2017-039

Roßmann, J., Gummer, T., & Silber, H. (2018). Mitigating satisficing in cognitively demanding

grid questions: Evidence from two web-based experiments. *Journal of Survey Statistics*

and Methodology, *6*(3), 376–400.

Sahlqvist, S., Song, Y., Bull, F., Adams, E., Preston, J., & Ogilvie, D. (2011). Effect of

questionnaire length, personalisation and reminder type on response rate to a complex

postal survey: Randomised controlled trial. *BMC Medical Research Methodology*, *11*(1),

1–8.

Sakshaug, J. W., Beste, J., Coban, M., Fendel, T., Haas, G.-C., Hülle, S., Kosyakova, Y., König,

C., Kreuter, F., Küfner, B., Müller, B., Osiander, C., Schwanhäuser, S., Stephan, G.,

Vallizadeh, E., Volkert, M., Wenzig, C., Westermeier, C., Zabel, C., & Zins, S. (2020).

Impacts of the COVID-19 Pandemic on Labor Market Surveys at the German Institute

for Employment Research. *Survey Research Methods*, *14*(2), 229–233.

https://doi.org/10.18148/srm/2020.v14i2.7743

Schwarz, N., Strack, F., Hippler, H.-J., & Bishop, G. (1991). The impact of administration mode

on response effects in survey measurement. *Applied Cognitive Psychology*, *5*(3), 193–

212. https://doi.org/10.1002/acp.2350050304

Smyth, M. M., Morris, P. E., Levy, P., & Ellis, A. W. (1987). *Cognition in action* (pp. xiii, 346).

Lawrence Erlbaum Associates, Inc.

The American Association for Public Opinion Research. (2016). *Standard Definitions: Final*

Dispositions of Case Codes and Outcome Rates for Surveys. 9th edition. AAPOR.

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

Cambridge University Press.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217.

Wagner, J., Arrieta, J., Guyer, H., & Ofstedal, M. B. (2014). Does Sequence Matter in Multimode Surveys: Results from an Experiment. *Field Methods*, *26*(2), 141–155. https://doi.org/10.1177/1525822X13491863

West, B. T., & Blom, A. G. (2017). Explaining Interviewer Effects: A Research Synthesis. *Journal of Survey Statistics and Methodology*, *5*(2), 175–211. https://doi.org/10.1093/jssam/smw024

Wolf, C., Christmann, P., Gummer, T., Schnaudt, C., & Verhoeven, S. (2021). Conducting General Social Surveys as Self-Administered Mixed-Mode Surveys. *Public Opinion Quarterly*.

Appendix

Table A1. Overview of scales

Scale	Thematic block	Items	# items	# scale points	CAPI page	Mail page	Split module	Showcards
Action	Politics and Society	v98-v101	4	3	16	18	D	Yes
Belong	National Identity	v164-v167	4	4	25	30	B	Yes
Childhood	Respondent's parents	v267-v269, v271-v273	6	4	54	60	A	Yes
Concern	Politics and Society	v212-v216	5	5	32	39	C	Yes
Concern_grp	Politics and Society	v217, v218, v220	3	5	32	40	C	Yes
Democracy1	Politics and Society	v133, v135, v136, v138, v141	5	11	21	26	D	Yes
Democracy2	Politics and Society	v134, v137, v139, v140	4	11	21	26	D	Yes
Elections	Politics and Society	v176-v183	8	4	27	32	D	Yes
Environment	Environment	v199-v203	5	5	30	36	D	Yes
European	National Identity	v194-v197	4	4	29	35	B	Yes
Immigration	National Identity	v185-v188	4	10	28	33	C	Yes
Importance	Perception of Life	v2, v3, v4	3	4	3	4	Core	Yes
Marriage	Family	v65-v70	6	3	13	15	A	No

National	National Identity	v189, v191-v193	4	4	29	34	B	Yes
Norms1	Religion and Morale	v149, v150, v152, v159, v162	5	10	23	29	C	Yes
Norms2	Religion and Morale	v151, v153-v158, v160, v161, v163	10	10	23	29	B-C (B: v151, v153-v155, v157, v158, v160, v161; C: v156, v163)	Yes
Pol_system	Politics and Society	v145-v148	4	4	23	28	D	Yes
Pol_watch	Politics and Society	v208-v211	4	5	32	39	D	Yes
Policy	Politics and Society	v103-v107	5	10	16	19	C	Yes
Society	Politics and Society	v221-v224	4	4	33	40	D	Yes
Surveillance	Politics and Society	v205-v207	3	4	31	38	D	Yes
Traditional_family	Family	v82-v84	3	5	14	17	A	Yes
Trust	Politics and Society	v115-v132	18	4	19	25	Core-B-C (Core: v115, v118, v121, v128; B: v117, v120, v122, v124, v126, v129,	Yes

							v131; C: v116, v119, v123, v125, v127, v130, v132)	
Trust_pl	National Identity	v32-v37	6	4	6	7	C	Yes
Work	Work	v46-v50	5	5	8	9	A-B (A: v46, v48, v50; B: v47, v49)	Yes

Notes: Showcards were used for all scales in CAPI interviews. A total of 274 items were collected. Matrix questionnaires included core module + 2 out of 4 modules (A to D). Modules were randomized as follows: Core + A/B; Core + A/C; Core + A/D; Core + B/C; Core + B/D; Core + C/D

Table A2. Availability of Response Style Indicators by Scale

Scale	# scale points	Middle answers	Primacy	Recency	Extreme response style	Straightlining
Action	3	Included	Included	Included		Included
Belong	4		Included	Included		Included
Childhood	4		Included	Included		Included
Concern	5	Included	Included	Included	Included	Included
Concern_grp	5	Included	Included	Included	Included	Included
Democracy1	11	Included	Included	Included	Included	Included
Democracy2	11	Included	Included	Included	Included	Included
Elections	4		Included	Included		Included
Environment	5	Included	Included	Included	Included	Included
European	4		Included	Included		Included
Immigration	10	Included	Included	Included	Included	Included
Importance	4		Included	Included		Included
Marriage	3	Included	Included	Included		Included
National	4		Included	Included		Included
Norms1	10	Included	Included	Included	Included	Included
Norms2	10	Included	Included	Included	Included	Included
Pol_system	4		Included	Included		Included
Pol_watch	5	Included	Included	Included	Included	Included
Policy	10	Included	Included	Included	Included	Included
Society	4		Included	Included		Included
Surveillance	4		Included	Included		Included
Traditional_f amily	5	Included	Included	Included	Included	Included
Trust	4		Included	Included		Included
Trust_pl	4		Included	Included		Included
Work	5	Included	Included	Included	Included	Included

Table A3. Percentage of significant differences in average item level means and standard deviations (SDs) by survey design and by thematic block and scale type

	Face-to-Face vs. Mixed-Mode Long		Mixed-Mode Long vs. Mixed-Mode Matrix		
	Means: %-significant differences*	SD: %-significant differences[†]	Means: %-significant differences*	SD: %-significant differences[†]	N Items
By Thematic Block					
Environment	0.00%	0.00%	0.00%	0.00%	5
Family	0.00%	0.00%	0.00%	0.00%	9
National Identity	37.50%	4.17%	0.00%	0.00%	24
Perception of Life	0.00%	16.67%	0.00%	0.00%	6
Politics and Society	14.71%	14.71%	7.35%	2.94%	68
Religion and Morale	13.33%	6.67%	0.00%	13.33%	15
Respondent's parents	0.00%	12.50%	0.00%	0.00%	8
Work	0.00%	20.00%	0.00%	0.00%	5
By Scale Type					
3 Point Scale	0.00%	0.00%	0.00%	0.00%	10
4 Point Scale	19.72%	11.27%	2.82%	2.82%	71
5 Point Scale	11.54%	11.54%	7.69%	0.00%	26
10 Point Scale	12.12%	12.12%	3.03%	6.06%	33
Total	15.00%	10.71%	3.57%	2.86%	140

*Notes: Proportion of significant differences between the means by survey mode and question type. *Significance test of means based on unpaired t-tests with Bonferroni corrections.*

[†]Significance test of distributions based on variance-ratio tests with Bonferroni corrections.

Topic	group	model	Filename	ChiSq	ChiSqDF	ChiSqP	CFI	RMSEA
action	Length	config	inv_data3_	50.525	5	0	0.987	0.091
action	Length	metric	inv_data3_	38.641	8	0	0.991	0.059
action	Length	scalar	inv_data3_	35.401	11	2.00E-04	0.993	0.045
action	Mode desig	config	inv_data3_	43.762	5	0	0.989	0.085
action	Mode desig	metric	inv_data3_	41.957	8	0	0.99	0.063
action	Mode desig	scalar	inv_data3_	59.391	11	0	0.986	0.064
belong	Length	config	inv_data3_	291.184	6	0	0.923	0.204
belong	Length	metric	inv_data3_	242.26	9	0	0.937	0.151
belong	Length	scalar	inv_data3_	222.152	15	0	0.944	0.11
belong	Mode desig	config	inv_data3_	296.645	6	0	0.924	0.212
belong	Mode desig	metric	inv_data3_	253.58	9	0	0.936	0.159
belong	Mode desig	scalar	inv_data3_	252.397	15	0	0.938	0.121
childhood	Length	config	inv_data3_	345.689	20	0	0.912	0.121
childhood	Length	metric	inv_data3_	317.296	25	0	0.921	0.103
childhood	Length	scalar	inv_data3_	281.012	35	0	0.934	0.08
childhood	Mode desig	config	inv_data3_	374.765	20	0	0.91	0.13
childhood	Mode desig	metric	inv_data3_	345.488	25	0	0.919	0.11
childhood	Mode desig	scalar	inv_data3_	312.358	35	0	0.93	0.087
concern	Length	config	inv_data3_	11.794	6	0.0667	0.997	0.029
concern	Length	metric	inv_data3_	16.171	10	0.0949	0.997	0.023
concern	Length	scalar	inv_data3_	34.795	15	0.0026	0.989	0.034
concern	Mode desig	config	inv_data3_	3.339	6	0.7652	1	0
concern	Mode desig	metric	inv_data3_	7.437	10	0.6837	1	0
concern	Mode desig	scalar	inv_data3_	13.865	15	0.5358	1	0
concern_gr	Length	config	inv_data3_	0	0	0	1	0
concern_gr	Length	metric	inv_data3_	7.022	2	0.0299	0.996	0.047
concern_gr	Length	scalar	inv_data3_	13.057	5	0.0228	0.993	0.038
concern_gr	Mode desig	config	inv_data3_	0	0	0	1	0
concern_gr	Mode desig	metric	inv_data3_	7.727	2	0.021	0.996	0.052
concern_gr	Mode desig	scalar	inv_data3_	24.365	5	2.00E-04	0.985	0.06
democracy	Length	config	inv_data3_	58.745	10	0	0.933	0.067
democracy	Length	metric	inv_data3_	60.56	14	0	0.936	0.055
democracy	Length	scalar	inv_data3_	83.273	19	0	0.912	0.055
democracy	Mode desig	config	inv_data3_	57.88	10	0	0.921	0.067
democracy	Mode desig	metric	inv_data3_	60.251	14	0	0.924	0.056
democracy	Mode desig	scalar	inv_data3_	99.363	19	0	0.868	0.063
democracy	Length	config	inv_data3_	7.496	4	0.1119	0.989	0.028
democracy	Length	metric	inv_data3_	7.756	7	0.3546	0.998	0.01
democracy	Length	scalar	inv_data3_	14.352	11	0.2141	0.989	0.017
democracy	Mode desig	config	inv_data3_	11.516	4	0.0213	0.975	0.042
democracy	Mode desig	metric	inv_data3_	11.852	7	0.1055	0.984	0.026
democracy	Mode desig	scalar	inv_data3_	18.705	11	0.0666	0.974	0.026
elections	Length	config	inv_data3_	467.036	42	0	0.945	0.098
elections	Length	metric	inv_data3_	417.454	49	0	0.953	0.084
elections	Length	scalar	inv_data3_	384.077	63	0	0.959	0.069
elections	Mode desig	config	inv_data3_	498.808	42	0	0.921	0.102
elections	Mode desig	metric	inv_data3_	454.222	49	0	0.93	0.089
elections	Mode desig	scalar	inv_data3_	436.669	63	0	0.935	0.075
environme	Length	config	inv_data3_	41.247	10	0	0.97	0.053

1								
2	environme	Length	metric	inv_data3_	41.749	14	1.00E-04	0.973	0.042
3	environme	Length	scalar	inv_data3_	46.119	19	5.00E-04	0.974	0.036
4	environme	Mode desig	config	inv_data3_	52.992	10	0	0.955	0.063
5	environme	Mode desig	metric	inv_data3_	58.794	14	0	0.953	0.054
6	environme	Mode desig	scalar	inv_data3_	75.867	19	0	0.941	0.053
7	european	Length	config	inv_data3_	36.27	6	0	0.996	0.067
8	european	Length	metric	inv_data3_	35.541	9	0	0.997	0.051
9	european	Length	scalar	inv_data3_	35.736	15	0.0019	0.997	0.035
10	european	Mode desig	config	inv_data3_	70.36	6	0	0.993	0.1
11	european	Mode desig	metric	inv_data3_	82.307	9	0	0.992	0.087
12	european	Mode desig	scalar	inv_data3_	98.907	15	0	0.991	0.072
13	immigratio	Length	config	inv_data3_	6.964	4	0.1378	0.998	0.026
14	immigratio	Length	metric	inv_data3_	9.441	7	0.2225	0.998	0.018
15	immigratio	Length	scalar	inv_data3_	17.066	11	0.1059	0.996	0.022
16	immigratio	Mode desig	config	inv_data3_	7.175	4	0.1269	0.997	0.027
17	immigratio	Mode desig	metric	inv_data3_	8.803	7	0.2671	0.998	0.015
18	immigratio	Mode desig	scalar	inv_data3_	18.934	11	0.0623	0.993	0.026
19	marriage	Length	config	inv_data3_	146.964	19	0	0.948	0.077
20	marriage	Length	metric	inv_data3_	144.99	24	0	0.951	0.066
21	marriage	Length	scalar	inv_data3_	134.535	29	0	0.957	0.056
22	marriage	Mode desig	config	inv_data3_	90.417	19	0	0.979	0.059
23	marriage	Mode desig	metric	inv_data3_	86.087	24	0	0.981	0.049
24	marriage	Mode desig	scalar	inv_data3_	84.535	29	0	0.983	0.042
25	national	Length	config	inv_data3_	61.115	6	0	0.988	0.089
26	national	Length	metric	inv_data3_	62.711	9	0	0.988	0.072
27	national	Length	scalar	inv_data3_	63.128	15	0	0.99	0.053
28	national	Mode desig	config	inv_data3_	65.509	6	0	0.984	0.096
29	national	Mode desig	metric	inv_data3_	72.121	9	0	0.983	0.081
30	national	Mode desig	scalar	inv_data3_	72.454	15	0	0.985	0.06
31	norms1	Length	config	inv_data3_	44.647	10	0	0.927	0.055
32	norms1	Length	metric	inv_data3_	51.305	14	0	0.921	0.048
33	norms1	Length	scalar	inv_data3_	65.291	19	0	0.902	0.046
34	norms1	Mode desig	config	inv_data3_	35.454	10	1.00E-04	0.921	0.049
35	norms1	Mode desig	metric	inv_data3_	42.619	14	1.00E-04	0.911	0.043
36	norms1	Mode desig	scalar	inv_data3_	50.814	19	1.00E-04	0.902	0.039
37	norms2	Length	config	inv_data3_	533.466	70	0	0.869	0.063
38	norms2	Length	metric	inv_data3_	552.465	79	0	0.866	0.06
39	norms2	Length	scalar	inv_data3_	577.417	89	0	0.861	0.057
40	norms2	Mode desig	config	inv_data3_	541.654	70	0	0.889	0.079
41	norms2	Mode desig	metric	inv_data3_	556.068	79	0	0.888	0.075
42	norms2	Mode desig	scalar	inv_data3_	606.518	89	0	0.878	0.073
43	pol_system	Length	config	inv_data3_	7.008	6	0.3201	0.999	0.012
44	pol_system	Length	metric	inv_data3_	11.256	9	0.2585	0.997	0.015
45	pol_system	Length	scalar	inv_data3_	24.543	15	0.0564	0.988	0.024
46	pol_system	Mode desig	config	inv_data3_	16.144	6	0.013	0.988	0.04
47	pol_system	Mode desig	metric	inv_data3_	15.089	9	0.0885	0.993	0.025
48	pol_system	Mode desig	scalar	inv_data3_	28.128	15	0.0208	0.984	0.029
49	pol_watch	Length	config	inv_data3_	27.512	4	0	0.967	0.073
50	pol_watch	Length	metric	inv_data3_	34.902	7	0	0.961	0.06
51	pol_watch	Length	scalar	inv_data3_	50.892	11	0	0.944	0.057

pol_watch	Mode desig	config	inv_data3_	16.541	4	0.0024	0.98	0.054
pol_watch	Mode desig	metric	inv_data3_	32.978	7	0	0.958	0.059
pol_watch	Mode desig	scalar	inv_data3_	111.109	11	0	0.839	0.092
policy	Length	config	inv_data3_	24.309	10	0.0068	0.949	0.036
policy	Length	metric	inv_data3_	34.407	14	0.0018	0.928	0.036
policy	Length	scalar	inv_data3_	43.518	19	0.0011	0.913	0.034
policy	Mode desig	config	inv_data3_	10.674	10	0.3835	0.997	0.008
policy	Mode desig	metric	inv_data3_	15.992	14	0.3139	0.991	0.011
policy	Mode desig	scalar	inv_data3_	41.636	19	0.002	0.895	0.033
society	Length	config	inv_data3_	29.147	6	1.00E-04	0.958	0.059
society	Length	metric	inv_data3_	27.689	9	0.0011	0.966	0.043
society	Length	scalar	inv_data3_	33.049	15	0.0046	0.968	0.033
society	Mode desig	config	inv_data3_	35.406	6	0	0.964	0.067
society	Mode desig	metric	inv_data3_	40.106	9	0	0.962	0.057
society	Mode desig	scalar	inv_data3_	46.677	15	0	0.962	0.044
surveillance	Length	config	inv_data3_	0.379	2	0.8275	1	0
surveillance	Length	metric	inv_data3_	3.661	4	0.4538	1	0
surveillance	Length	scalar	inv_data3_	11.499	8	0.175	0.999	0.02
surveillance	Mode desig	config	inv_data3_	2.581	2	0.2751	1	0.016
surveillance	Mode desig	metric	inv_data3_	4.503	4	0.3422	1	0.011
surveillance	Mode desig	scalar	inv_data3_	23.29	8	0.003	0.993	0.042
traditional_	Length	config	inv_data3_	0	0	0	1	0
traditional_	Length	metric	inv_data3_	6.908	2	0.0316	0.979	0.046
traditional_	Length	scalar	inv_data3_	8.737	5	0.12	0.984	0.026
traditional_	Mode desig	config	inv_data3_	0	0	0	1	0
traditional_	Mode desig	metric	inv_data3_	4.656	2	0.0975	0.985	0.035
traditional_	Mode desig	scalar	inv_data3_	17.36	5	0.0039	0.932	0.048
trust	Length	config	inv_data3_	1818.979	272	0	0.934	0.054
trust	Length	metric	inv_data3_	1673.096	289	0	0.941	0.05
trust	Length	scalar	inv_data3_	1526.458	323	0	0.949	0.044
trust	Mode desig	config	inv_data3_	2430.497	272	0	0.909	0.086
trust	Mode desig	metric	inv_data3_	2187.848	289	0	0.92	0.078
trust	Mode desig	scalar	inv_data3_	2015.204	323	0	0.929	0.07
trust_pl	Length	config	inv_data3_	446.216	20	0	0.985	0.137
trust_pl	Length	metric	inv_data3_	383.351	25	0	0.987	0.112
trust_pl	Length	scalar	inv_data3_	333.235	35	0	0.99	0.087
trust_pl	Mode desig	config	inv_data3_	534.704	20	0	0.985	0.154
trust_pl	Mode desig	metric	inv_data3_	457.098	25	0	0.987	0.126
trust_pl	Mode desig	scalar	inv_data3_	527.103	35	0	0.985	0.114
work	Length	config	inv_data3_	34.621	10	1.00E-04	0.97	0.038
work	Length	metric	inv_data3_	36.928	14	8.00E-04	0.972	0.031
work	Length	scalar	inv_data3_	54.375	19	0	0.956	0.033
work	Mode desig	config	inv_data3_	29.635	10	0.001	0.98	0.043
work	Mode desig	metric	inv_data3_	34.312	14	0.0019	0.979	0.037
work	Mode desig	scalar	inv_data3_	53.793	19	0	0.964	0.041

var	scale	CFI_dif	sig_CFI	sig_grp
v98	3	NA	NA	FALSE
v98	3	0.004	FALSE	FALSE
v98	3	0.002	FALSE	FALSE
v98	3	NA	NA	FALSE
v98	3	0.001	FALSE	FALSE
v98	3	-0.004	FALSE	FALSE
v164	4	NA	NA	FALSE
v164	4	0.014	FALSE	FALSE
v164	4	0.007	FALSE	FALSE
v164	4	NA	NA	FALSE
v164	4	0.012	FALSE	FALSE
v164	4	0.002	FALSE	FALSE
v267	4	NA	NA	FALSE
v267	4	0.009	FALSE	FALSE
v267	4	0.013	FALSE	FALSE
v267	4	NA	NA	FALSE
v267	4	0.009	FALSE	FALSE
v267	4	0.011	FALSE	FALSE
v212	5	NA	NA	FALSE
v212	5	0	FALSE	FALSE
v212	5	-0.008	FALSE	FALSE
v212	5	NA	NA	FALSE
v212	5	0	FALSE	FALSE
v212	5	0	FALSE	FALSE
v217	5	NA	NA	FALSE
v217	5	-0.004	FALSE	FALSE
v217	5	-0.003	FALSE	FALSE
v217	5	NA	NA	FALSE
v217	5	-0.004	FALSE	TRUE
v217	5	-0.011	TRUE	FALSE
v133	10	NA	NA	FALSE
v133	10	0.003	FALSE	TRUE
v133	10	-0.024	TRUE	FALSE
v133	10	NA	NA	FALSE
v133	10	0.003	FALSE	TRUE
v133	10	-0.056	TRUE	FALSE
v134	10	NA	NA	FALSE
v134	10	0.009	FALSE	FALSE
v134	10	-0.009	FALSE	FALSE
v134	10	NA	NA	FALSE
v134	10	0.009	FALSE	TRUE
v134	10	-0.01	TRUE	FALSE
v176	4	NA	NA	FALSE
v176	4	0.008	FALSE	FALSE
v176	4	0.006	FALSE	FALSE
v176	4	NA	NA	FALSE
v176	4	0.009	FALSE	FALSE
v176	4	0.005	FALSE	FALSE
v199	5	NA	NA	FALSE

v199	5	0.003	FALSE	FALSE
v199	5	0.001	FALSE	FALSE
v199	5 NA	NA		FALSE
v199	5	-0.002	FALSE	TRUE
v199	5	-0.012	TRUE	FALSE
v194	4 NA	NA		FALSE
v194	4	0.001	FALSE	FALSE
v194	4	0	FALSE	FALSE
v194	4 NA	NA		FALSE
v194	4	-0.001	FALSE	FALSE
v194	4	-0.001	FALSE	FALSE
v185	10 NA	NA		FALSE
v185	10	0	FALSE	FALSE
v185	10	-0.002	FALSE	FALSE
v185	10 NA	NA		FALSE
v185	10	0.001	FALSE	FALSE
v185	10	-0.005	FALSE	FALSE
v65	3 NA	NA		FALSE
v65	3	0.003	FALSE	FALSE
v65	3	0.006	FALSE	FALSE
v65	3 NA	NA		FALSE
v65	3	0.002	FALSE	FALSE
v65	3	0.002	FALSE	FALSE
v189	4 NA	NA		FALSE
v189	4	0	FALSE	FALSE
v189	4	0.002	FALSE	FALSE
v189	4 NA	NA		FALSE
v189	4	-0.001	FALSE	FALSE
v189	4	0.002	FALSE	FALSE
v149	10 NA	NA		FALSE
v149	10	-0.006	FALSE	TRUE
v149	10	-0.019	TRUE	FALSE
v149	10 NA	NA		TRUE
v149	10	-0.01	TRUE	FALSE
v149	10	-0.009	FALSE	FALSE
v151	10 NA	NA		FALSE
v151	10	-0.003	FALSE	FALSE
v151	10	-0.005	FALSE	FALSE
v151	10 NA	NA		FALSE
v151	10	-0.001	FALSE	TRUE
v151	10	-0.01	TRUE	FALSE
v145	4 NA	NA		FALSE
v145	4	-0.002	FALSE	FALSE
v145	4	-0.009	FALSE	FALSE
v145	4 NA	NA		FALSE
v145	4	0.005	FALSE	FALSE
v145	4	-0.009	FALSE	FALSE
v208	5 NA	NA		FALSE
v208	5	-0.006	FALSE	TRUE
v208	5	-0.017	TRUE	FALSE

v208	5	NA	NA	TRUE
v208	5	-0.022	TRUE	TRUE
v208	5	-0.119	TRUE	FALSE
v103	10	NA	NA	TRUE
v103	10	-0.021	TRUE	TRUE
v103	10	-0.015	TRUE	FALSE
v103	10	NA	NA	FALSE
v103	10	-0.006	FALSE	TRUE
v103	10	-0.096	TRUE	FALSE
v221	4	NA	NA	FALSE
v221	4	0.008	FALSE	FALSE
v221	4	0.002	FALSE	FALSE
v221	4	NA	NA	FALSE
v221	4	-0.002	FALSE	FALSE
v221	4	0	FALSE	FALSE
v205	4	NA	NA	FALSE
v205	4	0	FALSE	FALSE
v205	4	-0.001	FALSE	FALSE
v205	4	NA	NA	FALSE
v205	4	0	FALSE	FALSE
v205	4	-0.007	FALSE	FALSE
v82	5	NA	NA	TRUE
v82	5	-0.021	TRUE	FALSE
v82	5	0.005	FALSE	FALSE
v82	5	NA	NA	TRUE
v82	5	-0.015	TRUE	TRUE
v82	5	-0.053	TRUE	FALSE
v115	4	NA	NA	FALSE
v115	4	0.007	FALSE	FALSE
v115	4	0.008	FALSE	FALSE
v115	4	NA	NA	FALSE
v115	4	0.011	FALSE	FALSE
v115	4	0.009	FALSE	FALSE
v32	4	NA	NA	FALSE
v32	4	0.002	FALSE	FALSE
v32	4	0.003	FALSE	FALSE
v32	4	NA	NA	FALSE
v32	4	0.002	FALSE	FALSE
v32	4	-0.002	FALSE	FALSE
v46	5	NA	NA	FALSE
v46	5	0.002	FALSE	TRUE
v46	5	-0.016	TRUE	FALSE
v46	5	NA	NA	FALSE
v46	5	-0.001	FALSE	TRUE
v46	5	-0.015	TRUE	FALSE