**DICHIARAZIONE SOSTITUTIVA DI CERTIFICAZIONE e/o SOSTITUTIVA DELL'ATTO DI NOTORIETÀ**
**(Art. 46 e 47 D.P.R. n. 445 del 28/12/2000)**

La sottoscritta

COGNOME Lomazzi NOME Vera

CODICE FISCALE LMZVRE78R54B300Y NATA IL 14/10/1978

A Busto Arsizio PROV. VA

RESIDENTE IN Cologne (Germany)

INDIRIZZO Weinsbergstr. 84, Cologne (Germany) CAP 50823

DOMICILIO

Consapevole delle sanzioni penali nel caso di dichiarazioni mendaci, di formazione o uso di atti falsi, richiamate dall'art. 76 del D.P.R. 445/2000

**DICHIARA CHE**

relativamente alle pubblicazioni in collaborazione in cui nel testo non sia esplicitamente indicato il contributo individuale degli autori, la progettazione e impostazione complessiva va considerata come frutto del lavoro congiunto dei coautori, mentre la stesura materiale delle singole parti è da attribuirsi come di seguito indicato.

In riferimento all'articolo:

Maineri A., **Lomazzi V.**, Luijkx R. (2021), Studying the priming effect of family norms on gender roles' attitudes: an experimental design. *Survey Research Methods*, 15 (1), pp. 43-64 doi:10.18148/srm/2021.v15i1.7656.

L'attribuzione della stesura materiale delle parti è la seguente:

- a cura di Vera Lomazzi i paragrafi: 1. Introduzione (43-44); 2. The context effect of gender norms on gender role attitudes (pp. 44-45).
- a cura di Angelica Maineri il paragrafo 5. Results (pp. 49-56).
- I tre autori hanno equamente contribuito alla stesura dei paragrafi: 3. The current study (p. 46); 4. Data and Methods (pp. 46-49); 6. Conclusion (56-58).

*Luogo, data* *La dichiarante*

Cologne, 09.04.2021

Vera Lomazzi

Survey Research Methods (2021)
Vol. 15, No. 1, pp. 43-64
doi:10.18148/srm/2021.v15i1.7656
European Survey Research Association



# Studying the context effect of family norms on gender role attitudes: an experimental design

Angelica M. Maineri
Department of Sociology
Tilburg University, The Netherlands

Vera Lomazzi
GESIS—Leibniz Institute for Social Sciences
Cologne, Germany

Ruud Luijkx
Department of Sociology
Tilburg University, The Netherlands and
Department of Sociology and Social Research
University of Trento, Italy

The measurement of gender role attitudes has been found to be problematic in previous studies, especially in comparative perspective. The present study adopts a novel approach and investigates the position of the gender role attitudes scale in the questionnaire as a potential source of bias. In particular, the present study aims at assessing the context effect of the family norms question on the measurement of gender role attitudes by adopting the theoretical perspective of the construal model of attitudes, according to which the adjacent questions constitute the context for interpreting and answering a stimulus. The study employs data from the CROss-National Online Survey panel, which was fielded in 2017 and contained an experiment where the order of the questions under investigation varied. The reliability, invariance and validity of the measurement of gender role attitudes across experimental settings and countries (Estonia, Great Britain and Slovenia) are explored adopting several analytical techniques within the Structural Equation Modelling (SEM) framework. Whereas the measurement of gender role attitudes resulted stable between experimental settings, some differences emerged in terms of criterion and, to a lesser extent, construct validity. Differences among the countries emerged, indicating that the cultural context may interact with the question context. Finally, we stress that the measurement is overall poor, urging survey infrastructures to investigate more in depth the formulation of the items measuring gender role attitudes.

*Keywords:* gender role attitudes; context effect; CRONOS; experiment; family norms; validity; measurement quality

## 1 Introduction

People have different opinions about what they consider to be the appropriate roles for women and men in society, in particular regarding the division of responsibilities in the public and private spheres. For example, one might support the gendered specialization of tasks and roles with a priori allocation of unpaid care activities to women and paid work to men. In contrast to the segregation of roles in specific domains because of gender, people can in fact express the preference for more progressive gender roles, supporting women's role in the public sphere as well as men's role in the private one. Authors generally refer to these beliefs as gender role attitudes (Alwin, 2005; Braun, 2008; K. S. Lee, Alwin, & Tufiş, 2007; Walter, 2018b).

Gender role attitudes (GRA) are considered a good proxy to measure gender equality values (Bergh, 2007). For this reason, GRA have been used by scholars interested in studying how these values change over time (Brewster & Padavic, 2000; Cotter, Hermsen, & Vanneman, 2011; Inglehart, 1983; K. S. Lee et al., 2007; Lomazzi, 2017a; Savelyev, 2014; Scott, Alwin, & Braun, 1996) and how they vary among countries (Albrecht, Edin, & Vroman, 2000; André, Gesthuizen, & Scheepers, 2013; Lomazzi, Israel, & Crespi, 2018; Panayotova & Brayfield, 1997; Sjöberg, 2004) in order to evaluate whether societies are developing more or less egalitarian cultures. As a matter of fact, most cross-sectional surveys, such as the International Social Survey Programme (ISSP), the World Values Survey (WVS), the Generations and Gender Programme (GGP), the Eurobarometer (EB), and the European Values Study (EVS) include items to investi-

Contact information: Angelica M. Maineri, Department of Sociology, Tilburg University, Prof. Cobbenhagenlaan 225, 5037 DB Tilburg (NL) (E-Mail: a.m.maineri@tilburguniversity.edu)

gate GRA, and provide information concerning several countries at many points in time.

Despite being largely used for substantive research, the measurement of GRA[1] appears problematic. When analyzing a significant number of countries, there is often a lack of measurement equivalence (Constantin & Voicu, 2015; Davidov, Muthen, & Schmidt, 2018; Lomazzi, 2018; van de Schoot et al., 2013; van Vlimmeren, Moors, & Gelissen, 2017). This is attributed not only to methodological discrepancies in data collection, sampling or translation, but also to cultural differences (Seddig & Lomazzi, 2019; van de Vijver & Tanzer, 2004). In particular, the operationalization of the concept itself[2], which may not suit the specific gender cultures that developed over time (Pfau-Effinger, 2004), and its sensitivity to cultural differences (Braun, 1998, 2009), enhance the risk of misleading results in cross-cultural comparisons. These cultural differences refer, for example, to the cross-cultural meaning of items, structural differences, prevailing cultural values, which differ across societies and over time (Braun, 1998, 2009; Seddig & Lomazzi, 2019; Walter, 2018b).

As reported by Walter (2018b), these critical aspects concern most of the GRA instruments nowadays available in cross-sectional surveys, including the one used by EVS2008. M. Voicu and Tufiş (2012) reported that the GRA indicator built on multiple items from EVS2008 was not tenable across the years and samples. Further investigation confirmed such instability (Lomazzi, 2017b) and revealed considerable variation in reliability across countries, and an inconsistent structure configuration of the measurement model among the 26 countries analyzed. Previous research (André et al., 2013; Baxter & Kane, 1995; Bolzendahl & Myers, 2004; Kroska & Elman, 2009; Sjöberg, 2004) provided empirical evidence of the effect of some socio-demographic variables on people's gender beliefs. For example, women and people who are more educated, less religious and with higher economic resources tend to express more egalitarian gender role attitudes. This empirical knowledge has been used to assess the construct validity of the GRA scale by contrasting the effect of these variables on the GRA scale surveyed in 2008 and in the previous wave of EVS 1999 (Lomazzi, 2017b). While the tests on the EVS1999 data were consistent with the evidence from previous research, the construct validity in 2008 was poor and lacked any systematic pattern. Looking for a source of such instability, the study by Lomazzi (2017b) wondered whether the new questions on family norms[3] introduced in the EVS2008 immediately before the GRA scale may have provided normative stimuli on family relations and, therefore, may have modified the context of response. Indeed, questions asked before the GRA scale define the contextual framework for the respondents' interpretation of the items, their judgment and the expression of related attitudes (Tourangeau, Rasinski, Bradburn, & D'Andrade, 1989).

So far, this hypothesis has not been empirically explored. The question of whether and how the normativity framed by the questionnaire context can affect the way the respondents interpret the GRA battery and express their opinion is the core issue of the current study.

Method biases affecting the measurement of GRA have not been thoroughly explored to date. In the present study, we aim at contributing to the literature on challenges to the measurement of gender role attitudes by questioning the position of the GRA scale in the questionnaire[4] as a potential source of bias (Lomazzi, 2017b; Tfaily, 2010; Walter, 2018a).

## 2 The context effect of family norms on gender role attitudes

Despite the fact that some early approaches tended to see attitudes as stable constructs that persist in the same

[1]Typical questions in the GRA scale ask the respondents to express their agreement with statements like: "A job is all right, but what most women really want is home and children"(EVS, ISSP, WVS), or "A working mother can establish just as warm and secure a relationship with her children as a mother who doesn't work" (EVS, ISSP, WVS). For extensive reviews of GRA items, see: Davis and Greenstein, 2009; Grunow, Begall, and Buchler, 2018; Walter, 2018b.

[2]For instance, the items often focus on women roles only, and are limited to the private sphere and to the consequences of mothers' participation in the labor market on family life (Braun, 1998; Constantin & Voicu, 2015; Walter, 2018a, 2018b). Also, the formulation of these items often assumes the male breadwinner pattern as the mainstream family model.

[3]In the 2008 wave, five items were added to the family norms items included in EVS1999 (1,2):

1. A man has to have children in order to be fulfilled (since 1999);
2. A marriage or a long-term stable relationship is necessary to be happy (since 1999);
3. Homosexual couples should be able to adopt children;
4. It is alright for two people to live together without getting married;
5. It is a duty towards society to have children;
6. People should decide for themselves whether to have children or not;
7. When a parent is seriously ill or fragile, it is mainly the adult child's duty to take care of him/her.

[4]While in the ISSP thematic module "Family and Changing Gender Roles" the GRA scale is the very first battery proposed to the respondents (Scholz, Jutz, Edlund, Öun, & Braun, 2014), in most of the other surveys either this measurement comes after questions concerning very different topics, as it generally happens in the Eurobarometer, or it follows items investigating issues that can be somehow associated with gender norms. For example, in the EVS2008, the GRA scale followed a battery of items containing strong statements about the way a family should or should not look like.

way over an individual's lifetime (Allport, 1935; Cook & Flay, 1978; Petty & Cacioppo, 1981), other perspectives assumed that attitudes are not solid constructs; on the contrary, they are unstable and subject to change (Schwarz, 1999; Tourangeau & Rasinski, 1988; Tourangeau, Rasinski, Bradburn, & D'Andrade, 1989; Wilson & Hodges, 1992). The change can be attributed to the individual's identity development, leading to value system transformations which, in turn, affect attitudes. Attitudes are considered the expression or the application of more complex value constructs and deeper beliefs which constitute an individual's values system (Halman & de Moor, 1993). This is a basic notion used by scholars to study value changes and their connection with social transformations (Ester, Halman, & de Moor, 1993; Halman & de Moor, 1993; Inglehart, 2003; Rokeach, 1968). Additionally, Wilson and Hodges (1992) referred to factors like mood, very recent experiences or people's behaviors and other social context elements to explain the possible causes of attitude instability. In the context of a questionnaire, also adjacent questions can constitute a stimulus influencing the expression of attitudes (Feldman & Lynch, 1988; Schwarz, 1999; Tourangeau & Rasinski, 1988). Since many survey programs collect information on people's attitudes, the issue of attitude stability in survey research is relevant for methodological researchers, as well as for scholars interested in unbiased substantive research.

The current study is particularly focused on the possible measurement bias of gender role attitudes caused by earlier questions on normative beliefs concerning family structures. In order to understand more thoroughly how this may occur, the construal model of attitudes (Schwarz, 1999; Wilson & Hodges, 1992) is particularly helpful. According to this perspective, people use the information immediately available to them to construct and manifest their attitudes. This "available information" can be retrieved from memory and activated by stimuli that may affect the process of attitude expression. Tourangeau and Rasinski (1988) argued that answering to an attitude question requires a four-step process and each of these steps can be influenced by the questionnaire context.

The first step concerns the interpretation of the question. When the topic is largely common, this step is often automatic. However, prior items can provide the contextual framework for the interpretation of the question and influence the respondent's perspective. In the case of EVS2008, for example, normative items like "It is a duty towards society to have children" may draw attention either to the right to self-determination and individualistic views, or to social pressure towards collective approaches to society. The respondent could interpret the following item "A job is alright but what most women really want is a home and children" from the perspective activated by the duty/freedom to have children.

In the second step, the respondents retrieve their beliefs on the issue presented in the question. Tourangeau and Rasinski (1988) referred to this phase as a "memory search", since the memory reacts to a stimulus through associative networks and performs the retrieval process using mechanisms of recall and recognition. This activation should come from reading the item but it could be influenced by previous questions, which may have triggered specific issues and made them immediately available for the retrieval. In our example, respondents may retrieve their general value orientation towards the freedom/duty to have children, especially in the case of women.

During the third step, respondents use the information they retrieved in the second step to make a judgment on the topic of the item. This judgment can be affected by the retrieval process and by the beliefs that earlier items may have triggered or overstimulated. Particularly in this step, two processes may take place and result in either carryover effects, in case respondents adopt the same references and thus information used for earlier questions is also used for the following, or in backfire effects, in case respondents react against items perceived as strong and normative or when they use extreme standard of comparison (Tourangeau & Rasinski, 1988). In the case of GRA, earlier family norm questions may act in both ways, which makes results difficult to control.

In the fourth and final step, respondents give answers balancing two processes: the selection of the answer categories proposed by the survey and the consistency check, which may involve the search for consistency with previous answers and the response to social desirability. Carryover and backfire effects can also take place in this step when respondents try to ensure consistency with previous answers, in particular when they wish to offer a specific idea of themselves, for instance by avoiding extreme positions.

The questionnaire context may influence the cognitive processes of answering to attitude questions, even if its effect may be unstable and not always clear or easy to identify and control (Tourangeau & Rasinski, 1988; Tourangeau, Rasinski, Bradburn, & D'Andrade, 1989; Tourangeau, Rasinski, Bradburn, & D'andrade, 1989; Wilson & Hodges, 1992). In the EVS2008, seven items expressing a strong normative position on family relations, functions and structure were introduced. These items might activate association processes in the retrieval and judgment steps which could influence the subsequent measurement of gender role attitudes. In fact, respondents may retrieve information concerning their general view on relationships between partners and family members, as well as the relationship between individuals and society. Such beliefs are also related to ideas concerning gendered functions, tasks and social roles, and therefore they activate interpretative and judgmental frameworks which can influence the answers to GRA items.

## 3 The current study

Built on the controversial and untenable results of the assessment of the GRA scale of the EVS2008, the current study aims at investigating the occurrence of a context effect in relation with earlier items tapping into family norms. In particular, this study explores whether the position in the questionnaire could affect the measurement of GRA. In order to test the presence of context effects, scholars generally employ experimental settings and compare the answers given by respondents under different conditions applying several techniques. Basic investigations mainly involve distribution checks (S. Lee, McClain, Webster, & Han, 2016; Tourangeau, Rasinski, Bradburn, & D'Andrade, 1989; Tourangeau, Singer, & Presser, 2003), mean comparisons (DeMoranville & Bienstock, 2003; Tormos, 2018), item correlations (S. Lee et al., 2016; McFarland, 1981), chi-square tests (McFarland, 1981; Stark et al., 2018; Tourangeau, Rasinski, Bradburn, & D'Andrade, 1989; Tourangeau et al., 2003), and Anova (DeMoranville & Bienstock, 2003; Tourangeau, Rasinski, Bradburn, & D'Andrade, 1989). While some authors investigate the context effects by employing the IRT approach (Rivers, Meade, & Lou Fuller, 2009), other scholars used different techniques to identify the consequences of such biases on construct validity, for example using regression analyses (Stark et al., 2018; Tormos, 2018), and on the model fit, by applying structural equation modeling (DeMoranville & Bienstock, 2003; B. Voicu, 2015).

Since the original EVS2008 data do not allow isolating possible effects of earlier questions, the study follows this vast literature and adopts an experimental design. The position of the GRA scale and its adjacent question is manipulated in order to compare the performance of the GRA scale in two different questionnaire settings. This allows for a detailed examination of the GRA scale when surveyed after the family norms question, contrasted with the situation when the order of the two questions is reversed. Based on previous research (Lomazzi, 2017b) and consolidate survey practice (Scholz et al., 2014), we expect that the GRA scale will perform better when it is not preceded by the family norms items. In order to evaluate the quality of the GRA measurement across experimental conditions and countries, the study will use several analytical tools, which are described in the next section.

## 4 Data and Methods

This study uses data from the CROss-National Online Survey panel (2018).[5] Unlike EVS2008, which relied on interviewer-administered interviews, CRONOS collected self-administered web-based questionnaires. The CRONOS panel consists of 6 waves and a welcome survey, administered online between December 2016 and December 2017 in Great Britain, Estonia and Slovenia. Respondents (~3000 invited) were recruited after participating in Round 8 of the European Social Survey (ESS). In Estonia and Slovenia, respondents were randomly selected from the population registers according to strata. Great Britain implemented a multistage sampling design and relied on an address-based sampling frame (for a more detailed description of the sampling strategies, see Survey, 2018). The panel allowed for several survey experiments. The methodology of CRONOS is described in greater details in Villar and Sommer (2017) and Villar et al. (2018).

Among the seven CRONOS surveys, the current study examines data from Wave 5 (November–December 2017), which contains the two batteries on GRA and family norms (FN), and retains the same item wordings of EVS2008. By varying the order of the batteries, two experimental settings are obtained. On the one hand, setting A retains the order FN-GRA, reproducing the order of the questions in the EVS2008. On the other hand, in setting B the order of the questions is reversed, i.e. GRA-FN. Respondents were randomly assigned to one of the two settings. The proper functioning of the randomization procedure was checked by assessing whether the distribution of sociodemographic characteristics (age, gender, educational level and area of residence) was the same between the two experimental groups (Mutz & Pemantle, 2015; Tormos, 2018). No significant differences emerged in the distribution of these variables across experimental groups, confirming that the randomization procedure had worked properly (see Table B1 in the Appendix). Taking into account only the sample units invited to take place in CRONOS, the participation rate in Wave 5 was 55% in Great Britain, 77% in Estonia and 85% in Slovenia, reaching 1833 respondents in total. As a reference, in ESS round 8, the response rates in these countries were: 68.4% in Estonia, 42.8% in United Kingdom and 55.9% in Slovenia[6].

### 4.1 Analytical strategy

In order to answer the research question, we gradually built up our analytical strategy. The overall goal is to evaluate whether the measurement of GRA performed well when it was not preceded by questions on family norms. Thus, the study compared the performance of the GRA measurement in the different experimental conditions by looking at different aspects, such as: reliability; measurement equivalence; construct and criterion validity. All the analyses were performed both on the full sample—in order to highlight the

[5]The CRONOS panel was implemented under the SERISS project, which received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654221.

[6]Compared to the whole gross sample of ESS round 8, response rates in CRONOS waves ranged between 20–25% in Estonia , 12–16% in Great Britain and 21–27% in Slovenia.

general trends—and on the three countries separately[7]—to rule out possible cross-national differences.

**Reliability.** When working with multidimensional concepts, reliability indicates the consistency of a measurement. In particular, internal consistency represents the correlations among items belonging to the same scale. The most widely used measure of reliability is Cronbach's alpha. Cronbach's alpha is sensitive to the number of items included and to the sample size, thus the results should be interpreted with caution. Usually, internal consistency is deemed good if Cronbach's alpha is larger than 0.8, and sufficient if it is above 0.6. In this study, Cronbach's alpha was computed, and bootstrapping (400 repetitions) applied to estimate the 95% confidence intervals. We tested the internal consistency of the GRA scale within each subsample of country/experimental setting, so as to evaluate whether reliability is higher when GRA precedes FN.

**Measurement model and model fit.** Before building more complex models, it was necessary to identify a measurement model able to adequately fit the data. Although the same GRA scale has already been formally surveyed in EVS2008, the measurement model has been proven to be unclear and unstable in previous research (Lomazzi, 2017b; M. Voicu & Tufiş, 2012). Due to the instability found in the measurement model in EVS2008, it was therefore necessary to identify a tenable measurement model of GRA using the new data collected. For this purpose, an exploratory factor analysis (EFA) has been carried out, in order to identify a potential configural model.

The quality of the measurement model emerged from the EFA has been evaluated with a Confirmatory Factor Analysis (CFA) model, in a structural equation modeling (SEM) framework. The evaluation of the model fit is carried out considering several fit statistics. Since chi-square tests are known to be sensitive to sample size (Saris, Satorra, & Sörbom, 1987), other chi-square based goodness-of-fit measures are generally considered preferable as, for example, the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the standardized root mean residual (SRMR) (Hu & Bentler, 1999; West, Taylor, & Wu, 2012). In order to be deemed acceptable, the model fit should meet the following criteria: CFI value higher than 0.90, and RMSEA and SRMR values lower than 0.08. The CFA model has been fitted in each country and experimental setting separately, with the aim of identifying a stable model, which allowed to further investigate the context effect.

**Measurement equivalence.** Measurement equivalence (or invariance) is considered a prerequisite to carry out meaningful comparisons of correlations, regression coefficients, or means between different groups (Davidov et al., 2018). The assessment of equivalence (i.e., measurement invariance) consists in evaluating "whether or not, under different condition of observing and studying phenomena, measurement operations yield measures of the same attribute" (Horn & McArdle, 1992, pg. 117). Multi-Group Confirmatory Factor Analysis (MGCFA) is one of the most used techniques to evaluate to what extent the same measurement model fits the data of different groups (Brown, 2015). MGCFA builds on a strict approach to invariance, which implies "exact equivalence" between parameters across the groups and an increasing level of restrictive conditions that results in different levels of measurement equivalence. The least restrictive is configural invariance, which requires the same latent variables across groups. Metric invariance imposes factors to have the same loadings. Achieving this level is considered necessary to be able to compare unstandardized regression coefficients, factor variances and covariances. Scalar invariance, which also imposes equivalence between indicator intercepts, is the most restrictive and is required to meaningfully compare factor means. Chen (2007) suggests the following criteria in order to evaluate the differences between levels of measurement invariance, depending on the sample size involved in the assessment: for sample sizes larger than 300 units, as in the current study, differences between the constrained and unconstrained models are too sizeable, thus leading to reject the more constrained model, when the change in CFI is larger than $-0.010$, complemented by a change in RMSEA larger than 0.015, or a change in SRMR larger than 0.030 (or 0.010 when moving from the metric to the scalar model).

Even if partially invariant solutions are generally considered acceptable (Byrne, Shavelson, & Muthén, 1989), the strict requirements implied by this analytical strategy often result in the preclusion of mean comparisons, especially when dealing with many groups (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2012). This stimulated a lively debate among scholars who have different perspectives on this matter. Whereas scholars like Welzel and Inglehart (Welzel, Brunkert, Inglehart, & Kruse, 2019; Welzel & Inglehart, 2016), challenge the accepted practice of testing for measurement invariance, others (Alemán & Woods, 2016; Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014; Sokolov, 2018) support the importance of measurement equivalence in comparative studies, but challenge the concept of exact equivalence implied by techniques rooted in the frequentist approach, like MGCFA. To include cultural variability in the assessment, the concept of "approximate equivalence" has been introduced (Asparouhov & Muthén, 2014; Cieciuch, Davidov, Schmidt, Algesheimer, & Schwartz, 2014; Davidov et al., 2018; Muthén & Asparouhov, 2012; van de Schoot et al., 2013) and new techniques have been developed, in particular in the Bayesian framework, hinting towards their promising application in particular when the assessment involves a large number of groups.

Concerns about measurement equivalence are particularly

[7] As long as it was allowed by the sample size.

relevant when dealing with concepts that have been demonstrated to be sensitive to cultural biases, as it is the case with GRA (Braun, 2009; Constantin & Voicu, 2015; Seddig & Lomazzi, 2019). Scholars interested in the comparability of the existing GRA scales hardly found equivalence in cross-sectional settings. In their study, Constantin and Voicu (2015) examined the GRA scale from ISSP2002 (32 countries) and WVS2005 (46 countries) using MGCFA. In both the cases, metric invariance was achieved but mean comparisons would not be reliable. Lomazzi (2018) performed similar analyses on data from WVS2010 and contrasted the results obtained adopting exact and approximate approaches. By using MGCFA, scalar invariance was achieved for a subgroup of 27 countries of the 59 included in the analysis. The alignment optimization allowed, instead, to meaningfully compare 35 out of the 59 countries. Similar results have been obtained by studies investigating the GRA scale surveyed in 36 countries by ISSP2012: whereas MGCFA results show the achievement of partial metric invariance, the alignment optimization allowed for a reliable factor means comparison of 36 countries (Lomazzi & Seddig, 2020). In the case of ISSP2012, the lack of invariance has been demonstrated to derive from cultural bias (Seddig & Lomazzi, 2019), but also method biases are well-known sources of non-equivalence (van de Vijver & Tanzer, 2004). The current study aims at investigating whether the the underlying structure of GRA scale is comparable between experimental groups and hence evaluate whether the order of the questions affected the measurement structure. The study involves a small number of groups and, therefore, MGCFA was employed to assess measurement equivalence across countries and between the two experimental settings.

**Criterion and construct validity.** Validity assessments are relevant to evaluate whether the instrument measures what it is assumed to measure. In order to establish whether the question order affected the measurement validity of GRA, in the current study criterion and construct validity tests are carried out.

A measurement instrument has criterion validity if it results associated with a separate measure, which is considered a "golden standard" (Bannigan & Watson, 2009; DeVellis, 2016). The dataset in use hardly contains any variables that could properly fit in the definition of "golden standard" for the GRA measurement. The assessment is therefore conducted by a) considering the association between the GRA indicator and an external variable correlated with the concept but not belonging to the GRA scale; b) assessing the correlations with the external variable by known socio-demographic groups. In the first case (a), the study examined the correlation with an item which is sometimes used to investigate attitudes towards gender roles, and does not belong to the GRA scale: this is the item "When jobs are scarce, men have more right to a job than women". The chosen external variable is part of the main ESS survey and was thus administered months before the GRA items, and in a face-to-face setting. The variable is meant to capture attitudes towards gender egalitarianism in the labor market, and its positive correlation with the GRA dimension would indicate criterion validity. The purpose of this test is to assess whether criterion validity differs in the two experimental settings, as we expect criterion validity to be stronger in setting B (GRA-FN) compared to setting A (FN-GRA). In the second case (b), the study assessed the correlation between the GRA dimension and "When job are scarce…" by age groups; by interacting the measure of gender egalitarianism and age groups, we thus provide a more refined test of criterion validity. Criterion validity was assessed in the SEM framework, by including the external criterion as an exogenous variable in the unconstrained CFA model and adding a covariance term between this and the GRA scale latent variable(s). The standardized covariance indicates the correlation. The resulting model was fitted not only on the full sample, but also by country and experimental setting, in order to establish under which question order the GRA measurement performed better. The model was, moreover, tested separately of each age group.

Furthermore, construct validity is assessed. Since previous research reported the lack of construct validity of the GRA measurements surveyed in EVS2008 contrasted to the EVS1999 ones (Lomazzi, 2017b), the comparison of construct validity between the different experimental settings and countries is particularly relevant for the purpose of this study. A measurement has construct validity when the observed relationships with other variables are similar to the expectations derived from theoretical and empirical studies. Studies investigating gender role attitudes consistently identify the main predictors of gender beliefs (André et al., 2013; Baxter & Kane, 1995; Bolzendahl & Myers, 2004; Kroska & Elman, 2009; Sjöberg, 2004). Accordingly, Table 1 summarizes the expected effect of the most relevant sociodemographic predictors. The operationalization of these variables is described more extensively in section 4.2. The compliance of the measurement with theoretical expectations was assessed in the SEM environment by adding an explanatory part to the previously selected CFA model. The linear relationships between the predictors and the latent GRA constructs have been compared across countries and experimental settings. Model fit was also taken into account using the aforementioned indices (CFI, RMSEA, SRMR); additionally, the explained variance is represented by the Coefficient of Determination (CD): the closer the value is to 1, the better the fit.

To sum up, the current study proceeds as follows: after assessing the measurement's reliability, we selected a tenable measurement model, where observed variables load on latent variable(s), and we checked whether the measurement model was equivalent across countries/experimental conditions. We

Table 1
*List and brief description of predictors used to test construct validity*

| Predictor | Known effect |
|---|---|
| Gender | Women tend to be more egalitarian than men |
| Age groups | Older generations tend to be more traditional than the younger ones |
| Educational level | More educated people tend to be more egalitarian |
| Current marital status | People who are experiencing or had experienced an institutionalized form of relationship tend to be more traditionalist than those who are single or living with a partner |
| Church attendance | People who do not experience an institutionalized form of religiosity (hence, those who do not attend religious services) tend to be more egalitarian |

then proceeded by adding the explanatory part, namely by adding exogenous variables that correlate with, or explain, the latent variable(s): in this way, we were able to assess criterion and construct validity. All analyses are conducted with the software StataMP, version 16. The syntax is available as Supplementary Material.

### 4.2 Measurements

The GRA items and their descriptive statistics are listed in Table 2 (see also Fig. A1 in the Appendix for the mean of each GRA item by country and experimental setting). They were presented to the respondents one by one, and each had four answer categories (1 Agree strongly; 2 Agree; 3 Disagree; 4 Disagree strongly). Due to the instability of the measurement in EVS2008 (cf. Lomazzi, 2017b; M. Voicu & Tufiş, 2012), and to the differences in data collection between EVS2008 and CRONOS, the factorial structure of the GRA scale has been preliminarily examined with an EFA. Table 3 reports the results of the EFA (principal factors extraction with Varimax rotation) performed on the full sample ($N = 1,791$ after list-wise deletion of missing on the items). Consistently with findings from EVS2008 (Lomazzi, 2017b), a three-factor structure emerged (see Table 3). Each of the three factors captures respectively 18%, 7% and 6% of the variance of the items. The first factor, "Caring responsibilities", comprises items related to child and household care and the role of the parents. The second factor, "Division of labor", taps opinions on women's work-family balance (cf. Valentova, 2016). The third and last factor, "Economic role of women", concerns employment and responsibilities concerning income. The three-factor structured emerged in the EFA is later assessed in a confirmatory setting.

The previous measurement potentially affecting the measurement of GRA is a battery investigating Family Norms. This scale comprises two dimensions (identified through factor analysis, see Table B2 in the Appendix). The first deals with normative beliefs concerning traditional family form ("A man has to have children in order to be fulfilled"; "A marriage or a long-term stable relationship is necessary to be happy"; "It is a duty towards society to have children"; "When a parent is seriously ill or fragile, it is mainly the adult child's duty to take care of him/her"). The second concerns new family forms ("Homosexual couples should be able to adopt children"; "It is alright for two people to live together without getting married"; "People should decide for themselves whether to have children or not"). The answer options span from 1 (Agree strongly) to 5 (Disagree strongly).

In order to test criterion and construct validity, additional variables were employed. As for criterion validity, the agreement with the sentence "When jobs are scarce, men have more right to a job than women", ranging from 1 (Agree strongly) to 5 (Disagree strongly), was used. Regarding construct validity, several variables were considered: women were identified from the variable measuring gender; age groups were aggregated from the continuous age variable; the educational level was captured by an ordinal variable in 7 categories representing an adapted version of ISCED main levels; a reduced version of marital status was created by identifying those who are married (legal marriage or civil union), previously married (widowed/divorced), living with a partner or single; finally, a dichotomous variable indicated those who never attend religious services. A list-wise deletion of missing values on these variables has led to the exclusion of 18 cases (resulting in $N = 1,815$, without considering, for the time being, the missing values in the GRA items[8]). The descriptive statistics of all the variables used can be found in Table 4.

## 5 Results

### 5.1 Reliability

Overall, the GRA scale displays low internal consistency (Cronbach's alpha=0.61—see Table 5), and the results indicate that there are no significant differences between experi-

[8]In contrast to what usually occurs with the GRA scale in a face-to-face setting, there is a low amount of don't-knows and no-answers overall (mostly below 1%). After checking that such distribution did not differ between experimental settings and that it did not exceed 1% for each variable, a list-wise deletion of missing values is performed in each analysis depending on the GRA variables employed.

Table 2
*Descriptive statistics of GRA items*

| Variable | Mean | Std. Dev. | Min | Max | % non-substantive answers[a] | N (total) |
|---|---|---|---|---|---|---|
| A working mother can establish just as warm and secure a relationship with her children as a mother who does not work* (*workmother*) | 3.18 | 0.68 | 1 | 4 | 0.44 | 1,833 |
| A pre-school child is likely to suffer if his or her mother works (*childsuffer*) | 2.83 | 0.68 | 1 | 4 | 0.49 | 1,833 |
| A job is alright but what most women really want is a home and children (*jobalright*) | 2.72 | 0.76 | 1 | 4 | 0.87 | 1,833 |
| Being a housewife is just as fulfilling as working for pay (*housewife*) | 2.46 | 0.75 | 1 | 4 | 0.87 | 1,833 |
| Having a job is the best way for a woman to be an independent person* (*independent*) | 2.93 | 0.67 | 1 | 4 | 0.55 | 1,833 |
| Both the husband and wife should contribute to household income* (*hhincome*) | 3.08 | 0.69 | 1 | 4 | 0.87 | 1,833 |
| In general, fathers are as well suited to look after their children as mothers* (*fathersuit*) | 3.03 | 0.63 | 1 | 4 | 0.82 | 1,833 |
| Men should take as much responsibility as women for the home and children* (*menrespons*) | 3.28 | 0.54 | 1 | 4 | 0.76 | 1,833 |

Items marked with * have been turned so that high values represent more egalitarian views.
[a] Don't know/I prefer not to answer

Table 3
*Factor loadings after varimax rotation ($N = 1,791$)*

| Items on Gender role attitudes | Caring responsibilities | Division of labor | Economic role of women |
|---|---|---|---|
| workmother[a] | 0.45[b] | 0.33 | 0.10 |
| childsuffer | 0.35 | 0.52[b] | 0.04 |
| jobalright | 0.12 | 0.55[b] | −0.02 |
| housewife | −0.15 | 0.42[b] | 0.27 |
| independent[a] | 0.04 | 0.10 | 0.49[b] |
| hhincome[a] | 0.24 | −0.02 | 0.50[b] |
| fathersuit[a] | 0.56[b] | 0.07 | 0.08 |
| menrespons[a] | 0.55[b] | 0.04 | 0.11 |
| Variance explained | 18% | 7% | 6% |

Extraction method: principal factors
[a] Items have been turned so that high values represent more egalitarian views. [b] Factor loadings > 0.4

mental conditions nor across countries. Although reliability seems higher in setting A (FN-GRA), compared to setting B (GRA-FN), the overlapping confidence intervals do not allow to draw conclusions. Reliability appears to be particularly low (below the widely used threshold of 0.6) in Estonia. Overall, these findings suggest that the correlations among items belonging to the GRA scale are moderately weak.

### 5.2 Measurement model and model fit

This section reports the results of the Confirmatory Factor Analysis (CFA) and related model fit, assessed adopting the most commonly used criteria (Chen, 2007; Hu & Bentler, 1999).

The factor structure emerged in the EFA (see Table 3) was tested in a CFA framework, with three latent variables

Table 4
*Descriptive statistics of explanatory variables*

| Variable | Mean/% | Std. Dev. | Min | Max | N |
|---|---|---|---|---|---|
| When jobs are scarce | 4.26 | 0.86 | 1 | 5 | 1,815 |
| Gender: Female | 0.57 | - | 0 | 1 | 1,815 |
| Age (continuous) | 48.12 | 16.37 | 18 | 94 | 1,815 |
| 18–34 | 0.25 | - | 0 | 1 | 1,815 |
| 35–59 | 0.48 | - | 0 | 1 | 1,815 |
| 60+ | 0.28 | - | 0 | 1 | 1,815 |
| Educational attainment | 4.66 | 1.65 | 1[a] | 7[b] | 1,815 |
| Marital status | | | | | |
| Married | 0.49 | - | 0 | 1 | 1,815 |
| Previously married | 0.12 | - | 0 | 1 | 1,815 |
| Living with partner | 0.20 | - | 0 | 1 | 1,815 |
| Single | 0.19 | - | 0 | 1 | 1,815 |
| No church attendance | 0.39 | - | 0 | 1 | 1,815 |
| Country | | | | | |
| EE—Estonia | 0.33 | - | 0 | 1 | 1,815 |
| GB-GBN—Great Britain | 0.33 | - | 0 | 1 | 1,815 |
| SI—Slovenia | 0.33 | - | 0 | 1 | 1,815 |

[a] Lower secondary [b] Higher tertiary

Table 5
*Cronbach's alpha by country and experimental setting ($N = 1,791$)*

| Experimental setting | EE | GB | SI | Overall |
|---|---|---|---|---|
| A. FN-GRA | 0.57 (0.48–0.65) | 0.65 (0.58–0.72) | 0.65 (0.57–0.73) | 0.63 (0.59–0.67) |
| B. GRA-FN | 0.58 (0.51–0.66) | 0.62 (0.54–0.71) | 0.61 (0.52–0.70) | 0.60 (0.55–0.64) |
| Overall | 0.58 (0.52–0.64) | 0.64 (0.59–0.68) | 0.63 (0.58–0.68) | 0.61 (0.58–0.64) |

95% Confidence intervals computed via bootstrapping (400 replications) in parentheses

loading on 8 observed items (see Fig. 1). The model, estimated on the full sample, displayed a poor model fit (see Table 6), with a significant chi-square ($\chi^2 = 452,991$, df=17, $p < .001$), a high value of the RMSEA (0.120) and a low CFI (0.770). The low values of the standardized factor loadings, in particular those of housewife and independent, indicate poor convergent validity, and the strong correlation between the latent dimensions Gra1 ("Caring responsibilities") and Gra2 ("Division of labor"), which equals 0.67 ($p < .001$), indicate poor discriminant validity. If tested by country and/or by experimental condition, the model does not converge in some conditions, particularly in setting B in Great Britain and Slovenia (see Table 6).

Despite several attempts of modifying the model (see Supplementary materials), e.g. by adding error covariances among the observed variables, the measurement model retaining the three-factors structure always performed very poorly and/or did not converge in some conditions. These difficulties in reaching convergence and, generally speaking, sufficient model fit could be explained by the weakness of this measurement and the problematic conceptualization of these items already pointed out in the literature (Braun, 1998, 2009; Grunow et al., 2018; Walter, 2018b).

A reduced model, limited to items related to the dimension tapping Caring responsibilities, appeared to be stable across subsamples. The reduced model comprises only one latent dimension and three observed variables (see Fig. 2); the model is just-identified, thus model fit is not reported. The model converged in all countries, experimental conditions, and combinations thereof. All standardized factor loadings were above 0.4, and residual variances of the observed items ranged between 0.59 and 0.83; there is more unexplained variance of the indicators (especially, *workmoth*) than explained by the latent variable, indicating that the quality of the measurement is quite poor. Nevertheless, further analyses focus on this reduced and more stable model.

Table 6
*model fit measures for the 3-factors model by country and experimental setting, and for the full sample*

| experimental setting | country | n | $\chi^2$(df = 17) | p-value | rmsea | cfi | srmr |
|---|---|---|---|---|---|---|---|
| a (fn–gra) | ee | 295 | 81.374 | 0.000 | 0.113 | 0.782 | 0.078 |
| b (gra–fn) | ee | 288 | 90.501 | 0.000 | 0.123 | 0.725 | 0.083 |
| a (fn–gra) | gb | 299 | 134.202 | 0.000 | 0.152 | 0.692 | 0.095 |
| b (gra–fn) | gb | 296 | -[a] | -[a] | -[a] | -[a] | -[a] |
| a (fn–gra) | si | 308 | 99.241 | 0.000 | 0.125 | 0.795 | 0.092 |
| b (gra–fn) | si | 287 | -[a] | -[a] | -[a] | -[a] | -[a] |
| a (fn–gra) + b (gra–fn) | ee + gb + si | 1773 | 452.991 | 0.000 | 0.120 | 0.770 | 0.073 |

df= degrees of freedom; rmsea= root mean square error of approximation; cfi= comparative fit index; srmr= standardized root mean residual.
[a] No convergence

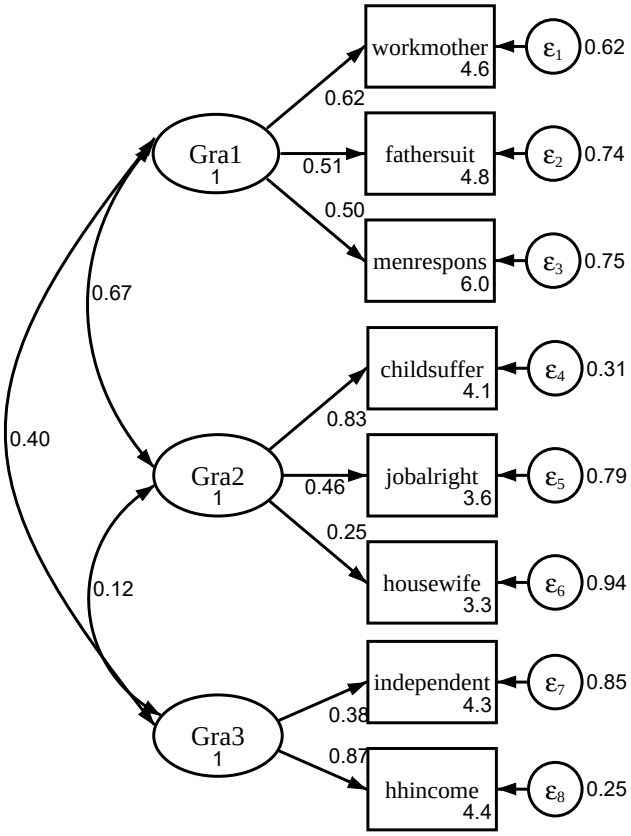


*Figure 1*. CFA 3-factors model; standardized coefficients ($N = 1,773$)

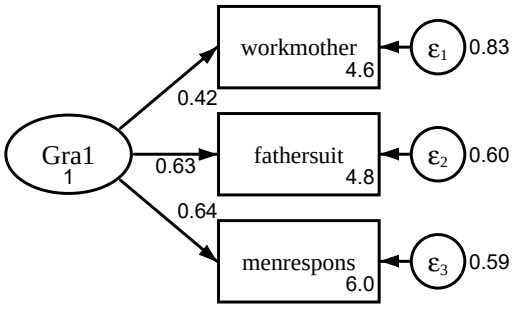


*Figure 2*. CFA reduced model; standardized coefficients ($N = 1,794$)

### 5.3 Measurement equivalence

This section reports the results of the MGCFA, employed to assess the equivalence of the measurement model of the GRA subdimension on Caring Responsibilities (GRA-CR) across experimental settings and countries. The goal of these tests was to assess whether the context of the questionnaire had affected the relationship between the GRA-CR latent dimension and the observed indicators, hence making the underlying measurement unstable and hindering comparisons.

Scalar invariance was achieved in both the Estonian and the Slovenian sample (see Table 7): in Estonia, the model fit overall did not deteriorate excessively when adding constraints; in Slovenia, although the change in RMSEA from the configural to the metric model was larger than 0.015, the change in SRMR was only slightly above 0.030, combined with a decrement in CFI smaller than -0.010; all parameter changes from the metric to the scalar model were acceptable. In the British sample however, due to the large change in the CFI from the metric to the scalar model, only the metric model could be accepted (despite the change in SRMR equal to 0.033. thus slightly above the set threshold).

Measurement equivalence was also assessed across countries within each experimental setting separately (see Table 8), in order to test whether the measurement of the GRA-CR dimension was more stable under one of the conditions. In both cases, only metric invariance was achieved, as the wors-

Table 7
*Measurement invariance assessment across experimental settings by country, model fit measures*

| | | $\chi^2$-Test | | | | | |
|---|---|---|---|---|---|---|---|
| Country | Model | $\chi^2$ | df | p-value | RMSEA | CFI | SRMR |
| EE ($N$ = 586) | configural | 0.000 | 0 | - | 0.000 | 1.000 | 0.000 |
| | metric | 0.646 | 2 | 0.724 | 0.000 | 1.000 | 0.014 |
| | scalar | 1.225 | 4 | 0.874 | 0.000 | 1.000 | 0.014 |
| GB ($N$ = 603) | configural | 0.000 | 0 | - | 0.000 | 1.000 | 0.000 |
| | metric | 3.166 | 2 | 0.205 | 0.044 | 0.992 | 0.033 |
| | scalar | 7.876 | 4 | 0.096 | 0.057 | 0.973 | 0.033 |
| SI ($N$ = 605) | configural | 0.000 | 0 | - | 0.000 | 1.000 | 0.000 |
| | metric | 4.038 | 2 | 0.133 | 0.058 | 0.991 | 0.031 |
| | scalar | 5.168 | 4 | 0.270 | 0.031 | 0.995 | 0.031 |

df= Degrees of Freedom; RMSEA= Root Mean Square Error of Approximation; CFI= Comparative Fit Index; SRMR= Standardized Root Mean Residual.

ening of the model fit from the metric to the scalar model is too large, especially as concerns the change in CFI (larger than 0.010) and RMSEA (larger than 0.015).

The tests of measurement invariance suggest that the measurement is mostly comparable across experimental settings, although to a lesser extent in Great Britain, where only metric invariance is achieved, thus not allowing mean comparisons of the latent GRA-CR items. In terms of cross-national comparison, the measurement is equally stable in the two experimental settings, as metric invariance is achieved in both settings. Evidence so far thus suggests that no context effect occurred in the measurement of GRA-CR.

### 5.4 Criterion validity

In the tests for criterion validity, stronger correlations in the expected directions indicate that the measurement under investigation – here, GRA egalitarianism in caring responsibilities– is better able to measure the concept that it is supposed to measure. Correlations were computed by adding an exogenous variable to the unconstrained SEM model, and adding a covariance between mentioned variables and the latent GRA-CR dimension. The standardized covariance can be read as the correlation between the two constructs.

Firstly, the correlation of the GRA-CR dimension with an external variable measuring a construct in the same domain of GRA was checked (see Table 9). Correlation with "Men have more right to jobs ..." was, as expected, positive and moderate. The correlation was stronger in setting B (0.33, $p < .001$) than in setting A (0.21, $p < .001$); this was especially visible in Estonia, where the correlation in setting A was actually not significant (0.13, $p > .05$), and Great Britain. The correlations were substantially the same across experimental conditions in Slovenia.

In order to provide a further test of criterion validity, the correlation between GRA egalitarianism in Caring Responsibilities and "Men have more right" was compared across age groups. The correlation appeared stronger among the older age group, indicating that for elderly people there is a stronger link between gender egalitarianism in the labor market and in the care sphere (see Table 10). However, the correlations were weaker in Setting A compared to setting B for the youngest group (where the correlation was not significant, being equal to 0.13, $p > .05$) and for the elderly group (where the correlation was almost half of the correlation in setting B). In setting B, moreover, the correlation was stronger in the 18-34 age group than in the 35-59 age group, but it remained the strongest in the 60+ age group.

Overall, criterion validity seems to be stronger in setting B, namely when GRA precedes FN. Particularly in Estonia, when FN precedes GRA, the measurement of GRA egalitarianism in Caring Responsibilities seems not to be valid, as it does not correlate with the external indicator of gender role egalitarianism.

### 5.5 Construct validity

The compliance of the GRA-CR measurement with theoretically-driven expectations on the influences of sociodemographic characteristics on gender role egalitarianism was assessed by adding explanatory variables to the measurement model (see Fig. 3 for visual representation of the model). The regression coefficients, estimated on the whole sample, are represented in Fig. 4, thus displaying the main effects of the predictors on the GRA dimension, leaving aside for the time being cross-national differences and question-order effects so as to have a baseline reference. Please remember that high values in the dependent variables represent

Table 8
*Measurement invariance assessment across countries by experimental setting, model fit measures*

| | | $\chi^2$-Test | | | | | |
|---|---|---|---|---|---|---|---|
| Experimental Setting | Model | $\chi^2$ | df | p-value | RMSEA | CFI | SRMR |
| A (FN–GRA) ($N$ = 910) | configural | 0.000 | 0 | - | 0.000 | 1.000 | 0.000 |
| | metric | 2.891 | 4 | 0.576 | 0.000 | 1.000 | 0.025 |
| | scalar | 60.689 | 8 | 0.000 | 0.147 | 0.808 | 0.029 |
| B (GRA–FN) ($N$ = 884) | configural | 0.000 | 0 | - | 0.000 | 1.000 | 0.000 |
| | metric | 1.586 | 4 | 0.811 | 0.000 | 1.000 | 0.018 |
| | scalar | 24.493 | 8 | 0.002 | 0.084 | 0.936 | 0.021 |

df= Degrees of Freedom; RMSEA= Root Mean Square Error of Approximation; CFI= Comparative Fit Index; SRMR= Standardized Root Mean Residual.

Table 9
*Standardized covariance (correlations) of "Caring responsibilities" with "Men have more right" from SEM model*

| | | | | 95% C.I. | |
|---|---|---|---|---|---|
| Country | Experimental setting | $N$ | $r$ | Lower | Upper |
| EE + GB + SI | A (FN-GRA) + B (GRA-FN) | 1794 | $0.27^{***}$ | 0.22 | 0.33 |
| EE + GB + SI | A (FN-GRA) | 910 | $0.21^{***}$ | 0.13 | 0.29 |
| | B (GRA-FN) | 884 | $0.33^{***}$ | 0.26 | 0.41 |
| EE | A (FN-GRA) | 296 | 0.13 | −0.01 | 0.27 |
| | B (GRA-FN) | 290 | $0.37^{***}$ | 0.23 | 0.50 |
| GB | A (FN-GRA) | 301 | $0.33^{***}$ | 0.18 | 0.46 |
| | B (GRA-FN) | 302 | $0.47^{***}$ | 0.34 | 0.58 |
| SI | A (FN-GRA) | 313 | $0.20^{**}$ | 0.07 | 0.33 |
| | B (GRA-FN) | 292 | $0.22^{**}$ | 0.08 | 0.36 |

$^{*}$ $p < 0.05$ $^{**}$ $p < 0.01$ $^{***}$ $p < 0.001$

more egalitarian gender role views in the sphere of Caring responsibilities.

Following expectations, it can be noticed in Fig. 4 that overall women tended to be more egalitarian than men, elderly groups were less egalitarian than the youngest group, single people were more egalitarian than married people, and those who did not attend church are more egalitarian that those who did. Yet, the educational level did not have a significant impact on GRA-CR egalitarianism, and those who were previously married and those who cohabit with a partner did not significantly differ from married people. Coefficients were small size, yet the explained variance was 8.6% (CD=0.086). Overall, the model fit was acceptable: although the CFI was below the threshold of 0.9, the RMSEA and the SRMR were both smaller than 0.08 (see Table 11).

We further split the analyses by experimental group and country, in order to test whether construct validity varies depending on the order of the questions and on the country (see Table 11 for model fit and Fig. 5 for unstandardized coefficients of explanatory part of the model). Most of the independent variables taken into account did not have a significant correlation with the GRA-CR dimension in any of the country/experimental setting, and the few significant coefficients were small in size and sparse. In Estonia, for instance, we observed significant gender differences only in setting A (FN-GRA), and age differences only in setting B (GRA-FN). In Great Britain, construct validity resulted extremely low, as hardly any of the predictors had a significant relationship with the GRA-CR dimension in any of the two settings. In Slovenia, instead, in setting B, significant coefficients were found in association with being a woman, being 60+, being single and not attending church; yet, educational levels had a small correlation with GR-CR egalitarianism only in setting A. Additionally, the explained variance represented by the coefficient of determination (CD) was higher in setting B (13.3%) compared to setting A (7.2%). Model fit of the full model was better in setting B compared to setting A in Estonia: whereas the RMSEA and SRMR were smaller than 0.08

Table 10
*Standardized covariance (correlations) of "Caring responsibilities" with "Men have more right" by age group from SEM model*

| | | 18–34 | | | 35–59 | | | 60 and above | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 95% C.I. | | | 95% C.I. | | | 95% C.I. | |
| Country | Experimental setting | *r* | Lower | Upper | *r* | Lower | Upper | *r* | Lower | Upper |
| EE + GB + SI | A (FN–GRA) + B (GRA–FN) | 0.21*** | 0.09 | 0.31 | 0.24*** | 0.16 | 0.32 | 0.37*** | 0.25 | 0.49 |
| EE + GB + SI | A (FN–GRA) | 0.13 | −0.01 | 0.27 | 0.23*** | 0.12 | 0.34 | 0.24* | 0.03 | 0.45 |
| | B (GRA–FN) | 0.31*** | 0.14 | 0.47 | 0.23*** | 0.12 | 0.35 | 0.47*** | 0.32 | 0.62 |
| *N* | | | 441 | | | 856 | | | 497 | |

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Table 11
*Model fit of full SEM model*

| Experimental setting | Country | N | $\chi^2$(df = 16) | p-value | RMSEA | CFI | SRMR | CD |
|---|---|---|---|---|---|---|---|---|
| A (FN-GRA) + B (GRA-FN) | EE + GB + SI | 1794 | 93.142 | 0.000 | 0.052 | 0.883 | 0.024 | 0.086 |
| A (FN-GRA) | EE | 296 | 37.055 | 0.002 | 0.067 | 0.828 | 0.038 | 0.050 |
| B (GRA-FN) | EE | 290 | 17.661 | 0.344 | 0.019 | 0.978 | 0.028 | 0.156 |
| A (FN-GRA) | GB | 301 | 38.507 | 0.001 | 0.068 | 0.812 | 0.039 | 0.167 |
| B (GRA-FN) | GB | 302 | 31.137 | 0.013 | 0.056 | 0.856 | 0.031 | 0.118 |
| A (FN-GRA) | SI | 313 | 25.605 | 0.060 | 0.044 | 0.921 | 0.028 | 0.072 |
| B (GRA-FN) | SI | 292 | 32.781 | 0.008 | 0.060 | 0.896 | 0.031 | 0.133 |

df= Degrees of Freedom; RMSEA= Root Mean Square Error of Approximation; CFI= Comparative Fit Index; SRMR= Standardized Root Mean Residual; CD = Coefficient of Determination.

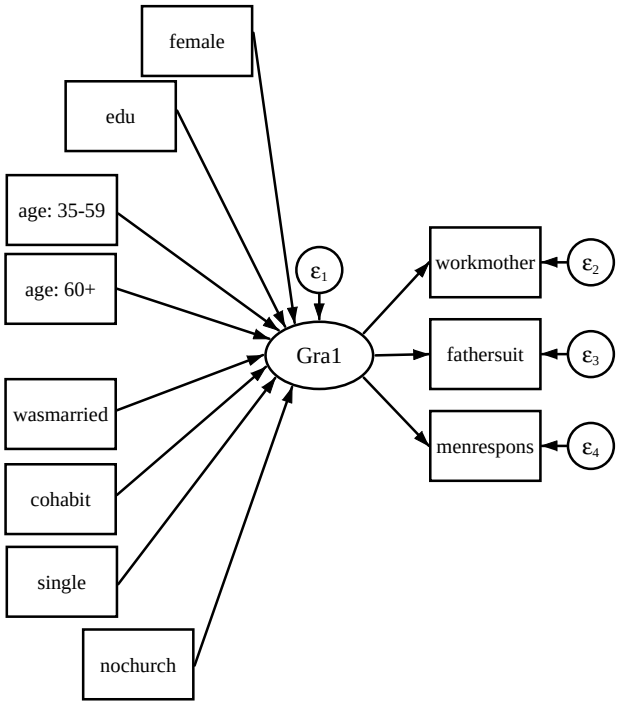


*Figure 3*. Full SEM model

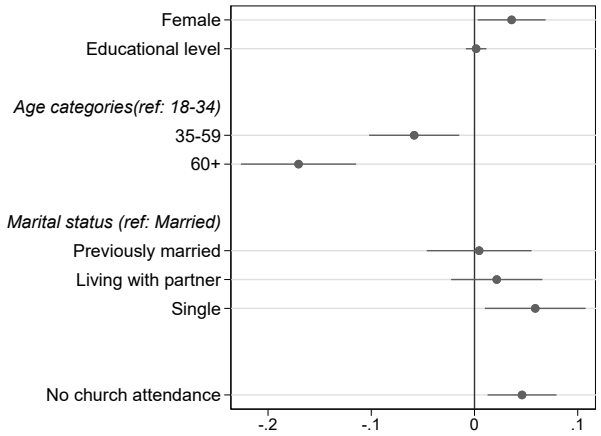


*Figure 4*. Baseline model estimated via SEM. Each dot represents the point estimate of the unstandardized coefficient, the lines the confidence intervals ($N$ = 1794)

in both settings, CFI was larger than 0.9 only in setting B. In Slovenia, the chi-square, RMSEA, CFI and SRMR indicators seemed to yield better outcomes in setting A; especially the CFI was larger than 0.9 only in setting A. The pattern was mixed in Great Britain: albeit the RMSEA and SRMR were

smaller than the threshold of 0.08, the CFI remained below 0.9 in both conditions.

All in all, construct validity appeared low and there were only few differences between experimental settings. The evidence is slightly leaning towards the setting in which GRA precedes FN, as construct validity seems to be higher in setting B, at least in terms of model fit in Estonia, and in terms of explanations aligning with theoretical expectations in Slovenia. By looking at the correlations of the predictors with the GRA dimension, construct validity appeared to be especially low in Great Britain compared to the other countries.

## 6 Conclusion and discussion

In the present study, a survey experiment fielded in a probability-based online panel was used to detect if the order of questions has an impact on the measurement of gender role attitudes. Several analytical techniques are used to address the question. The study presents a novel approach in the study of the GRA, as previous studies questioning the quality and the comparability of GRA measurements did not take into account potential biases due to the position of the scale in the questionnaire. Moreover, whereas previous studies on question order effects on single questions/items, the present study dealt with a multidimensional measurement. Despite large variability in the results, summarized in Table 12, some general conclusions can be drawn, and future lines of research can be suggested.

The results of the experiment analyzed in this study are not clear-cut. The measurement model including all eight GRA items and three dimensions appeared poor in quality and highly unstable, a condition which hindered further comparison across experimental groups. This made it necessary to scale down to a simplified model, by focusing on one dimension of GRA only. The tests of measurement invariance indicated that the evidence of a context effect in the measurement structure of GRA-CR is weak. The measurement appeared to be substantially stable across countries and experimental settings; surely, no experimental setting outperformed the other. Yet, in terms of measurement quality, the validity tests seem to lean slightly towards the condition in which GRA preceded FN: in setting B, both criterion validity and, to a lesser extent, construct validity appeared higher compared to setting A. The difference is however small, and construct validity in particular is generally low; moreover, evidence is limited to one dimension of GRA. Hence, we cannot ultimately conclude that the GRA scale performed better when it was not preceded by the question on family norms. While we could expect low validity in setting A (FN-GRA) following previous studies (e.g. Lomazzi, 2017b), the situation did not appear to sensibly improve when the FN question is moved after GRA.

What clearly emerged from our study is that differences among countries matter. Although the measurement structure of the GRA-CR dimension appeared stable in cross-national perspective, the validity tests yielded different results in each country. Criterion and construct validity appeared particularly low in Estonia when FN preceded GRA (setting A), whereas construct validity was almost absent in both settings in Great Britain. In Slovenia, the alignment with theoretical explanations at the basis of construct validity was visibly stronger when GRA came first (setting B). One potential explanation for these cross-national differences is suggested by Braun (2009). When analyzing potential problems in cross-country comparison of GRA, the author pinpointed how in former socialist countries it is likely to find more egalitarian attitudes towards female employment, channeling gender role differentiation into the family care sphere. This may perhaps explain why for Estonian respondents (and, to a lesser extent, the Slovenian too) the FN question appeared more visibly problematic in terms of quality for the measurement of GRA, compared to British respondents. Moreover, the experiment the present study is based on the assumption that the strong normative statements in the FN question could affect the interpretation and assessment of gender role attitudes, e.g. the item "it is a duty towards society to have children" could trigger either individualistic views or social pressure to adopt collective approaches. It may be, however, that respondents from the different countries were not equally sensitive to these stimuli. Stark et al. (2018) questioned the generalizability of results of question-order experiments, often restricted to English-speaking countries: our study fielded experiments in three distinct European countries, whose respondents showed different reactions to the question order. As the authors suggested, contextual characteristics may influence the occurrence of the question-order effect, depending on the conditions under which that effect should emerge (Stark et al., 2018). For instance, they mentioned the cultural salience of the norm of evenhandedness (i.e. making two subsequent judgements consistent to avoid bias), the average level of education (as a proxy for cognitive abilities) and the different perception of the contrast among the issues involved in the question-order experiment (see Stark et al., 2018).

The performance of the GRA scale was overall poor. Setting aside differences between experimental conditions, quality indicators like reliability, validity and goodness-of-fit tests showed a problematic situation. This may also be because the items do not allow to properly differentiate respondents, since there is a large tendency to express agreement with the egalitarian-loaded items (and disagreement with the others), avoiding extreme answer categories. Future research should investigate whether and why GRA items are not able anymore to grasp differences in gender egalitarianism, not even within the same country. In light of these results, and following other authors (e.g. Walter, 2018a), we urge survey programs to investigate in depth the current wording of

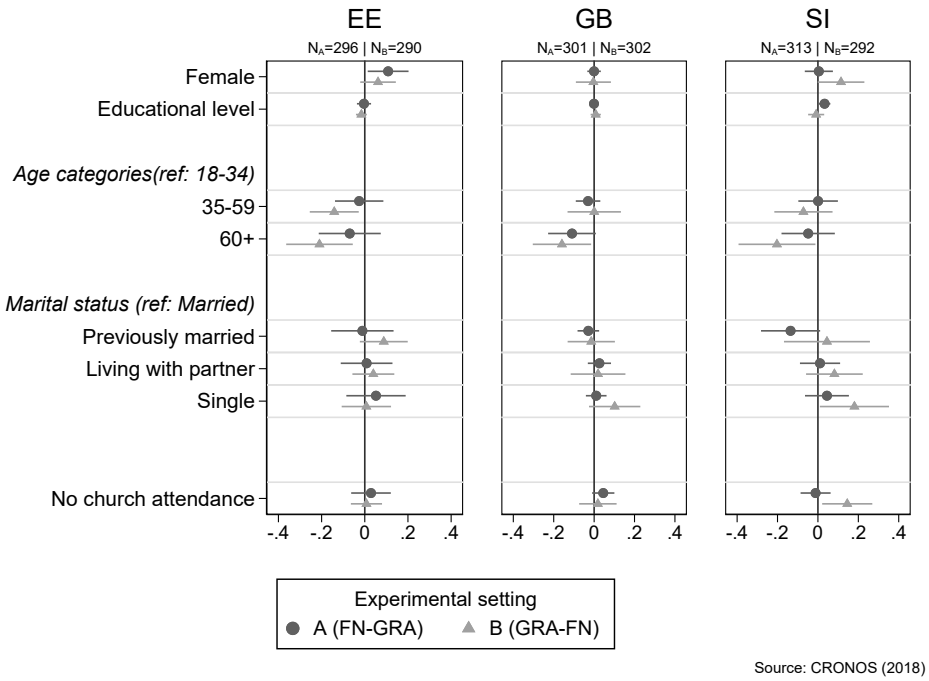


*Figure 5*. Unstandardized coefficients estimated via SEM. Each dot represents the point estimate of the unstandardized coefficient, the lines the confidence intervals

Table 12
*Summary of results*

| Type of test | GRA-CR measurement performance |
|---|---|
| Reliability (GRA scale) | No sizeable differences between experimental settings |
| Measurement invariance | Across experimental settings within countries: Metric invariance in GB, scalar invariance in EE and SI<br>Across countries within experimental settings: Metric invariance across countries in both settings |
| Criterion validity | Stronger in setting B (GRA-FN) |
| Construct validity | Slightly stronger in setting B (GRA-FN), but overall low (especially in GB) |

the items, so as to consider whether better ways to measure gender role attitudes in contemporary societies should be devised, while retaining the possibility of analyzing trends over time. Future research may expand this study and investigate FN and GRA items simultaneously, for instance along the lines of the study by van Vlimmeren et al. (2017): despite tackling a concept which is broader than gender role attitudes, such a study could shed light on the context effect by looking at it from a different perspective, e.g. not only GRA being influenced by FN, but also, vice versa, FN being influenced by GRA. Additionally, methods devised to control for priming effects, such as the one proposed by B. Voicu (2015), could be tested in the framework of FN and GRA to evaluate whether measurement quality increases.

One limitation of the experiment evaluated in this study concerns the fact that the GRA items were the first of the questionnaire, in setting B. Although this was done on purpose to avoid other potential priming elements, it is also possible that a complete lack of context for such complex items made the interpretation difficult for respondents. In CRONOS Wave 1 a similar experiment—with the same design but different wording of items—was fielded. However, the experiment was preceded by other items, perhaps creating a more natural flow of questions for the respondent, and thus facilitating the answering process. Further research should look into this aspect and, more generally, investi-

gate the best position of the GRA question in the questionnaire. Moreover, since two similar experiments were fielded in CRONOS among the same respondents, it would be possible to test whether the same respondent reacted differently to the same GRA item when asked under two different experimental conditions.

Another potentially relevant difference between our starting point, EVS2008, and the experiment reported in this study, concerns the different mode of data collection, as EVS2008 relied on face-to-face interviews, and CRONOS data were collected via web surveys. Whereas the measurement of GRA may be sensitive to social desirability or acquiescence bias in interviewer-administered settings, e.g. depending on the gender of the interviewer, the situation seemingly did not sensibly improve in the web self-administered context of CRONOS, where the GRA measurement still proved unstable and of low validity. Additionally, presenting the items one by one, as it was done in the CRONOS survey, may have weakened the cognitive associations between the FN and GRA items, thus making it less likely for the context effect to occur. Nevertheless, it should be pointed out that a recent study comparing CRONOS to the face-to-face ESS interviews found that measurement quality was only slightly lower in the former, and that metric measurement invariance holds for most of the topics investigated (Cernat & Revilla, 2020).

It may be interesting to evaluate the GRA measurement in other surveys and modes. For instance, in the latest EVS wave (EVS2017), some of the GRA items were replaced, and the few items that were retained from the family norms scale of EVS2008 were placed after the GRA question. Although direct comparison is not allowed due to the different item wording, and the lack of an experimental design, it would be interesting to evaluate the performance of the GRA measurement in EVS2017 in light of the results of the present study. Moreover, due to its mixed-mode design, EVS2017 would offer the opportunity to evaluate the performance of the GRA measurement in different modes.

The initial question, regarding the causes of the poor measurement of GRA in the EVS2008, remains partially open. This study showed that question order is not likely to be responsible for the unexpected results yielded at the time, hence leaving the issue open to further investigation. Qualitative approaches, as cognitive interviews or online probing (Behr, Braun, Kaczmirek, & Bandilla, 2013; Braun, 2008), may provide useful insights for evaluating the cross-cultural comparability of GRA. Nevertheless, this study showed that the measurement of GRA performed poorly in a self-administered survey setting, and that differences between countries are relevant when investigating question order effects.

**Acknowledgements**

This work was made possible by the Internship grant awarded by the European Consortium for Sociological Research (ECSR) in 2018. We are grateful to the anonymous reviewers for the constructive comments on the manuscript. We are also thankful to Giulia Brandolini for contributing to the readability of the paper, and to Malina Voicu and Ana Villar for their support in designing the experiment.

**References**

Albrecht, J. W., Edin, P.-A., & Vroman, S. B. (2000). A cross-country comparison of attitudes towards mothers working and their actual labor market experience. *Labour*, *14*(4), 591–607. doi:10.1111/1467-9914.00147

Alemán, J., & Woods, D. (2016). Value orientations from the world values survey: How comparable are they cross-nationally? *Comparative Political Studies*, *49*(8), 1039–1067. doi:10.1177/0010414015600458

Allport, G. W. (1935). Attitudes. In *A handbook of social psychology.* (pp. 798–844). Worcester, MA, US: Clark University Press.

Alwin, D. F. (2005). Attitudes, beliefs, and childbearing. In A. Booth & A. Crouter (Eds.), *The new population problem: Why families in developed countries are shrinking and what it means* (Chap. 8, pp. 115–126). Mahwah, NJ: Lawrence Erlbaum Associates.

André, S., Gesthuizen, M., & Scheepers, P. (2013). Support for traditional female roles across 32 countries: Female labour market participation, policy models and gender differences. *Comparative Sociology*, *12*(4), 447–476. doi:10.1163/15691330-12341270

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508. doi:10.1080/10705511.2014.919210

Bannigan, K., & Watson, R. (2009). Reliability and validity in a nutshell. *Journal of clinical nursing*, *18*(23), 3237–3243.

Baxter, J., & Kane, E. (1995). Dependence and independence. a cross-national analysis of gender inequality and gender attitudes. *Gender & Society*, *9*(2), 193–215. doi:10.1177/089124395009002004

Behr, D., Braun, M., Kaczmirek, L., & Bandilla, W. (2013). Testing the validity of gender ideology items by implementing probing questions in web surveys. *Field Methods*, *25*(2), 124–141. doi:10.1177/1525822X12462525

Bergh, J. (2007). Gender attitudes and modernization processes. *International Journal of Public Opinion Research*, *19*(1), 5–23. doi:10.1093/ijpor/edl004

Bolzendahl, C. I., & Myers, D. J. (2004). Feminist attitudes and support for gender equality: Opinion change in women and men, 1974–1988. *Social Forces*, *83*(2), 759–789. doi:10.1353/sof.2005.0005

Braun, M. (1998). Gender roles. In J. Van Deth (Ed.), *Comparative politics: The problem of equivalence* (pp. 111–134). London, United Kingdom: Routledge.

Braun, M. (2008). Using egalitarian items to measure men's and women's family roles. *Sex Roles*, *59*(9-10), 644–656. doi:10.1007/s11199-008-9468-5

Braun, M. (2009). The role of cultural contexts in item interpretation. In M. Haller, R. Jowell, & T. Smith (Eds.), *The International Social Survey Programme, 1984–2009: Charting the globe* (pp. 395–408). London/New York: Routledge.

Brewster, K. L., & Padavic, I. (2000). Change in gender-ideology, 1977-1996: The contributions of intracohort change and population turnover. *Journal of Marriage and Family*, *62*(2), 477–487. doi:10.1111/j.1741-3737.2000.00477.x

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.

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.

Cernat, A., & Revilla, M. (2020). Moving from face-to-face to a web panel: Impacts on measurement quality. *Journal of Survey Statistics and Methodology*. doi: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. doi:10.1080/10705510701301834

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., & Schwartz, S. H. (2014). Comparing results of an exact versus an approximate (bayesian) measurement invariance test: A cross-country illustration with a scale to measure 19 human values. *Frontiers in Psychology*, *5*(September), 1–10. doi:10.3389/fpsyg.2014.00982

Constantin, A., & Voicu, M. (2015). Attitudes towards gender roles in cross-cultural surveys: Content validity and cross-cultural measurement invariance. *Social Indicators Research*, 733–751. doi:10.1007/s11205-014-0758-8

Cook, T. D., & Flay, B. R. (1978). The persistence of experimentally induced attitude change. *Advances in Experimental Social Psychology*, *11*(100), 1–57. doi:10.1016/S0065-2601(08)60004-0

Cotter, D., Hermsen, J. M., & Vanneman, R. (2011). The end of the gender revolution? Gender role attitudes from 1977 to 2008. *American Journal of Sociology*, *117*(1), 259–289. doi:10.1086/658853

CROss-National Online Survey panel. (2018). CRONOS 0-6 ESS8 integrated data file, edition 1.2 [CRONOS_ESS8_e01_2.dta]. NSD—Norwegian Centre for Research Data, Norway—Data Archive and distributor of CRONOS data for ESS ERIC. Retrieved from http://www.europeansocialsurvey.org/methodology/methodological_research/modes_of_data_collection/cronos.html

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*, 55–75. doi:10.1146/annurev-soc-071913-043137

Davidov, E., Muthen, B., & Schmidt, P. (2018). Measurement invariance in cross-national studies: Challenging traditional approaches and evaluating new ones. *Sociological Methods and Research*, *47*(4), 631–636. doi:10.1177/0049124118789708

Davis, S. N., & Greenstein, T. N. (2009). Gender ideology: Components, predictors, and consequences. *Annual Review of Sociology*, (35), 87–105. doi:10.1146/annurev-soc-070308-115920

DeMoranville, C. W., & Bienstock, C. C. (2003). Question order effects in measuring service quality. *International Journal of Research in Marketing*, *20*(3), 217–231. doi:10.1016/S0167-8116(03)00034-X

DeVellis, R. F. (2016). *Scale development: Theory and applications*. Sage publications.

Ester, P., Halman, L., & de Moor, R. (1993). Values shift in western societies. In P. Ester, L. Halman, & R. de Moor (Eds.), *The individualizing society. value change in europe and north america* (pp. 1–20). Tilburg: Tilburg University Press.

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

Grunow, D., Begall, K., & Buchler, S. (2018). Gender ideologies in europe: A multidimensional framework. *Journal of Marriage and Family*, *80*(1), 42–60. doi:10.1111/jomf.12453

Halman, L., & de Moor, R. (1993). Comparative research on values. In P. Ester, L. Halman, & R. de Moor (Eds.), *The individualizing society. value change in europe and north america* (pp. 21–36). Tilburg: Tilburg University Press.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117–144. doi:10.1080/03610739208253916

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), 1–55. doi:10.1080/10705519909540118

Inglehart, R. F. (1983). The persistence of materialist and post-materialist value orientations: Comments on van deth's analysis. *European Journal of Political Research*, *11*, 81–91. doi:10.1111/j.1475-6765.1983.tb00044.x

Inglehart, R. F. (2003). *Human values and social change: Findings from the values surveys*. Brill.

Kroska, A., & Elman, C. (2009). Change in attitudes about employed mothers: Exposure, interests, and gender ideology discrepancies. *Social Science Research*, *38*(2), 366–382. doi:10.1016/j.ssresearch.2008.12.004

Lee, K. S., Alwin, D. F., & Tufiş, P. A. (2007). Beliefs about women's labour in the reunified germany, 1991-2004. *European Sociological Review*, *23*(4), 487–503. doi:10.1093/esr/jcm015

Lee, S., McClain, C., Webster, N., & Han, S. (2016). Question order sensitivity of subjective well-being measures: Focus on life satisfaction, self-rated health, and subjective life expectancy in survey instruments. *Quality of Life Research*, *25*(10), 2497–2510. doi:10.1007/s11136-016-1304-8

Lomazzi, V. (2017a). Gender role attitudes in italy: 1988–2008. a path-dependency story of traditionalism. *European Societies*, *19*(4), 370–395. doi:10.1080/14616696.2017.1318330

Lomazzi, V. (2017b). Testing the goodness of the evs gender role attitudes scale. *BMS Bulletin of Sociological Methodology/ Bulletin de Methodologie Sociologique*, *135*(1), 90–100. doi:10.1177/0759106317710859

Lomazzi, V. (2018). Using alignment optimization to test the measurement invariance of gender role attitudes in 59 countries. *Methods, Data, Analyses*, *12*(1), 77–104. doi:10.12758/mda.2017.09

Lomazzi, V., Israel, S., & Crespi, I. (2018). Gender equality in europe and the effect of work-family balance policies on gender-role attitudes. *Social Sciences*, *8*(1), 5. doi:10.3390/socsci8010005

Lomazzi, V., & Seddig, D. (2020). Gender role attitudes in the international social survey programme: Cross-national comparability and relationships to cultural values. *Cross-Cultural Research*, *54*(4), 398–431. doi:10.1177/1069397120915454

McFarland, S. G. (1981). Effects of question order on survey responses. *Public Opinion Quarterly*, *45*(2), 208. doi:10.1086/268651

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological methods*, *17*(3), 313–335.

Mutz, D. C., & Pemantle, R. (2015). Standards for experimental research: Encouraging a better understanding of experimental methods. *Journal of Experimental Political Science*, *2*, 192–215. doi:10.1017/XPS.2015.4

Panayotova, E., & Brayfield, A. (1997). National context and gender ideology: Attitudes toward women's employment in hungary and the united states. *Gender & Society*, *11*(5), 627–655.

Petty, R. E., & Cacioppo, J. T. (1981). *Communication and persuasion: Central and peripheral routes to attitude change*. Brown, Dubuque, IA.

Pfau-Effinger, B. (2004). Socio-historical paths of the male breadwinner model - an explanation of cross-national differences. *British Journal of Sociology*, *55*(3), 377–399. doi:10.1111/j.1468-4446.2004.00025.x

Rivers, D. C., Meade, A. W., & Lou Fuller, W. (2009). Examining question and context effects in organization survey data using item response theory. *Organizational Research Methods*, *12*(3), 529–553. doi:10.1177/1094428108315864

Rokeach, M. (1968). *Beliefs, attitudes, and values: A theory of organization and change*. San Francisco: Jossey-Bass.

Saris, W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. *Sociological Methodology*, *17*, 105–129. Retrieved from http://www.jstor.org/stable/271030

Savelyev, Y. (2014). *Modernization and variations in emancipative values in european societies in 1995-2008: Test of inglehart's socialization hypothesis*. Retrieved from https://econpapers.repec.org/RePEc:hig:wpaper:48/soc/2014

Scholz, E., Jutz, R., Edlund, J., Öun, I., & Braun, M. (2014). *Issp 2012 family and changing gender roles iv: Questionnaire development*. GESIS-Technical Reports.

Schwarz, N. (1999). Self-reports. how the questions shape the answers. *American Psychologist*, *54*(2), 93–105. doi:10.1037/0003-066X.54.2.93

Scott, J., Alwin, D. F., & Braun, M. (1996). Generational changes in gender-role attitudes: Britain in a cross-national perspective. *Sociology*, *30*(3), 471–492. doi:10.1177/0038038596030003004

Seddig, D., & Lomazzi, V. (2019). Using cultural and structural indicators to explain measurement noninvariance in gender role attitudes with multilevel structural equation modeling. *Social Science Research*, *84*. doi:https://doi.org/10.1016/j.ssresearch.2019.102328

Sjöberg, O. (2004). The role of family policy institutions in explaining gender-role attitudes: A comparative multilevel analysis of thirteen industrialized countries. *Journal of European Social Policy*, *14*(2), 107–123. doi:10.1177/0958928704042003

Sokolov, B. (2018). The index of emancipative values: Measurement model misspecifications. *American Political*

*Science Review*, *112*(2), 395–408. doi:10.1017/S0003055417000624

Stark, T. H., Silber, H., Krosnick, J. A., Blom, A. G., Aoyagi, M., Belchior, A., … Yu, R.-R. (2018). Generalization of classic question order effects across cultures. *Sociological Methods & Research*, 1–36. doi:10.1177/0049124117747304

Survey, E. S. (2018). *Ess 2016 documentation report ed 2.1*.

Tfaily, R. (2010). Cross-community comparability of attitude questions: An application of item response theory. *International Journal of Social Research Methodology*, *13*(2), 95–110. doi:10.1080/13645570902920145

Tormos, R. (2018). Question-order effects in the evaluation of political institutions in decentralized polities. *International Journal of Public Opinion Research*, 1–26. doi:10.1093/ijpor/edy013

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *103*(3), 299–314.

Tourangeau, R., Rasinski, K. A., Bradburn, N., & D'Andrade, R. (1989). Belief accessibility and context effects in attitude measurement. *Journal of Experimental Social Psychology*, *25*, 401–421. doi:10.1016/0022-1031(89)90030-9

Tourangeau, R., Rasinski, K. A., Bradburn, N., & D'andrade, R. (1989). Carryover effects in attitude surveys. *Public Opinion Quarterly*, *53*, 495–524. Retrieved from http://poq.oxfordjournals.org/

Tourangeau, R., Singer, E., & Presser, S. (2003). Context effects in attitude surveys: Effects on remote items and impact on predictive validity. *Sociological Methods and Research*, *31*(4), 486–513. doi:10.1177/0049124103251950

Valentova, M. (2016). How do traditional gender roles relate to social cohesion? focus on differences between women and men. *Social Indicators Research*, *127*(1), 153–178. doi:10.1007/s11205-015-0961-2

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing offwith scylla and charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, *4*(OCT), 1–15. doi:10.3389/fpsyg.2013.00770

van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue europeenne de psychologie appliquee*, *54*(2), 119–135. doi:10.1016/j.erap.2003.12.004

van Vlimmeren, E., Moors, G. B. D., & Gelissen, J. P. T. M. (2017). Clusters of cultures : Diversity in meaning of family value and gender role items across europe. *Quality & Quantity*, *51*(6), 2737–2760. doi:10.1007/s11135-016-0422-2

Villar, A., & Sommer, E. (2017). *Web recruitment design plans and experimental testing. deliverable 7.3 of the seriss project funded under the european union's horizon 2020 research and innovation programme ga no: 654221*. Retrieved from www.seriss.eu/resources/deliverables

Villar, A., Sommer, E., Finnoy, D., Gaia, A., Berzelak, N., & Bottoni, G. (2018). *Cross-national online survey (cronos) panel data and documentation user guide*. ESS ERIC. London.

Voicu, B. (2015). Priming effects in measuring life satisfaction. *Social Indicators Research*, *124*(3), 993–1013. doi:10.1007/s11205-014-0818-0

Voicu, M., & Tufiş, P. A. (2012). Trends in gender beliefs in romania: 1993-2008. *Current Sociology*, *60*(1), 61–80. doi:10.1177/0011392111426648

Walter, J. G. (2018a). Measures of gender role attitudes under revision: The example of the german general social survey. *Social Science Research*, *72*(February), 170–182. doi:10.1016/j.ssresearch.2018.02.009

Walter, J. G. (2018b). The adequacy of measures of gender roles attitudes: A review of current measures in omnibus surveys. *Quality and Quantity*, *52*(2), 829–848. doi:10.1007/s11135-017-0491-x

Welzel, C., Brunkert, L., Inglehart, R. F., & Kruse, S. (2019). Measurement equivalence? a tale of false obsessions and a cure. *World Values Research*, *11*(3), 54–84. doi:10.2139/ssrn.2390636

Welzel, C., & Inglehart, R. F. (2016). Misconceptions of measurement equivalence: Time for a paradigm shift. *Comparative Political Studies*, *49*(8), 1068–1094. doi:10.1177/0010414016628275

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling*. (pp. 209–231). New York, NY, US: The Guilford Press.

Wilson, T. D., & Hodges, S. D. (1992). Attitudes as temporary constructions. In L. L. Martin & A. Tesser (Eds.), *The construction of social judjments* (pp. 37–65). Hillsdale, NJ: Erlbaum.

Appendix A
Figures

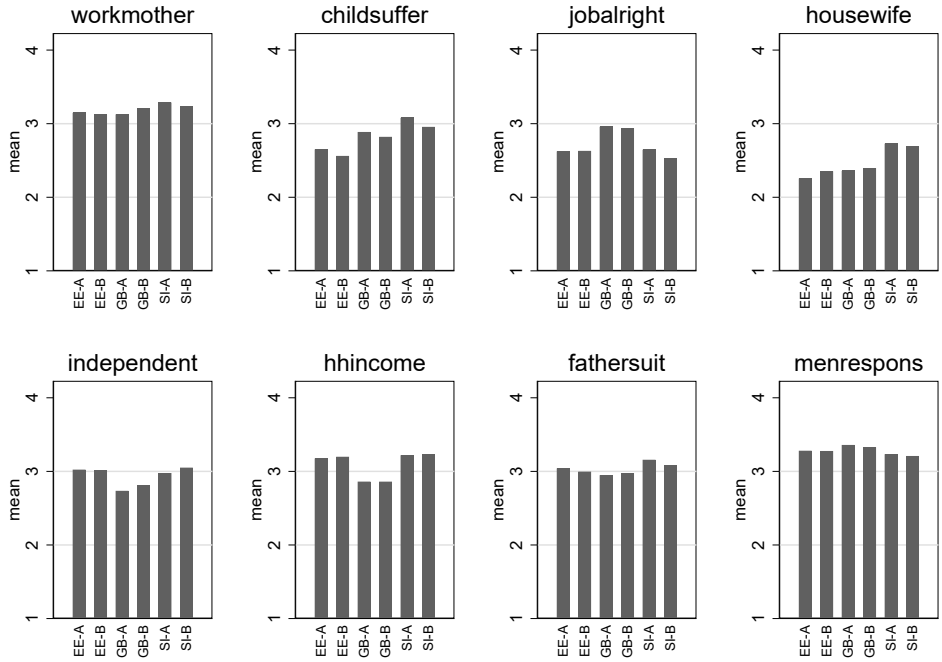


*Figure A1*. Mean of GRA items by country and experimental setting

## Appendix B
## Tables

Table B1
*Distribution of gender, age categories, education and area of residence by question order*

| | A. FN-GRA | B. GRA-FN | Comparison | | |
|---|---|---|---|---|---|
| Variable | % | % | $\chi^2$ | df | p-value |
| Gender | $N$ = 929 | $N$ = 904 | 1.01 | 1 | 0.315 |
| Female | 56.0 | 58.3 | - | - | - |
| Age | $N$ = 926 | $N$ = 902 | 5.25 | 2 | 0.073 |
| 18-34 | 26.8 | 22.2 | - | - | - |
| 35-59 | 46.2 | 49.2 | - | - | - |
| 60+ | 27.0 | 28.6 | - | - | - |
| Education | $N$ = 924 | $N$ = 899 | 0.78 | 2 | 0.677 |
| Lower (Isced 0-2) | 13.1 | 11.8 | - | - | - |
| Middle (Isced 3-4) | 36.1 | 37.3 | - | - | - |
| Higher (Isced 5-8) | 50.8 | 50.9 | - | - | - |
| Area of residence | $N$ = 917 | $N$ = 879 | 4.21 | 4 | 0.378 |
| Big city | 13.6 | 14.8 | - | - | - |
| Suburbs | 12.8 | 14.4 | - | - | - |
| Large town | 21.9 | 18.8 | - | - | - |
| Small town | 21.2 | 22.7 | - | - | - |
| Rural area | 30.4 | 29.2 | - | - | - |

df= Degrees of Freedom

Table B2
*Factor loadings after varimax rotation (N = 1,813)*

| Item on Family norms | Traditional family | New family |
|---|---|---|
| q1 A man has to have children in order to be fulfilled | 0.67[a] | −0.30 |
| q2 A marriage or a long-term stable relationship is necessary to be happy | 0.55[a] | −0.35 |
| q3 Homosexual couples should be able to adopt children | −0.28 | 0.48[a] |
| q4 It is alright for two people to live together without getting married | −0.16 | 0.57[a] |
| q5 It is a duty towards society to have children | 0.57[a] | −0.37 |
| q6 People should decide for themselves whether to have children or not | −0.17 | 0.47[a] |
| q7 When a parent is seriously ill or fragile, it is mainly the adult child's duty to take care of him/her | 0.33 | −0.06 |
| Variance explained by factors | 31% | 4% |

Extraction method: principal factors
[a] Factor loadings > 0.4