Received: 17 May 2021 | Accepted: 8 December 2021

DOI: 10.1111/bjso.12520

the british psychological society
promoting excellence in psychology

ARTICLE

Is use of the general system justification scale across countries justified? Testing its measurement equivalence

Denise Vesper | **Cornelius J. König** | **Rudolf Siegel** | **Malte Friese**

Department of Psychology, Saarland University, Saarbrücken, Germany

Correspondence
Denise Vesper, Saarland University, Campus A1 3, 66123 Saarbrücken, Germany.
Email: denise.vesper@uni-saarland.de

Abstract

System justification is a widely researched topic in social and political psychology. One major measurement instrument in system justification research is the General System Justification Scale (G-SJS). This scale has been used, among others, for comparisons across social groups in different countries. Such comparisons rely on the assumption that the scale is measurement equivalent. However, this assumption has never been comprehensively tested. Thus, the present two studies assessed the measurement equivalence of the G-SJS following classic measurement equivalence guidelines (i.e., multigroup confirmatory factor analyses) in Study 1 and using a new method for comparing larger numbers of groups in Study 2 (i.e., alignment optimization). In Study 1, we analysed the measurement equivalence in Great Britain (n = 444), Germany (n = 454), and France (n = 463). In Study 2, we used a publicly available dataset consisting of 66 samples from 30 countries (N = 13,495) to again assess the measurement equivalence of the scale. Results indicated (partial) metric equivalence, but not scalar equivalence in both studies. Overall, the studies indicate that mean comparisons across the examined countries are not warranted with the current form of the G-SJS. The scale needs to be revised for valid cross-country comparisons of means.

KEYWORDS
measurement equivalence, measurement invariance, multi-country comparison, system justification

INTRODUCTION

System justification theory recently turned 25 years old (Jost, 2019). Over the years, numerous publications support the system justification construct, its antecedents, and predictions (e.g Jost et al., 2017; Osborne et al., 2019; van der Toorn et al., 2011). In fact, system justification theory is one of the most influential theories in recent social psychological research. The original article by Jost and Banaji (1994) has been cited 3,958 times, according to Google Scholar (as of October 8th, 2021), and the publication of articles dealing with system justification has increased exponentially since the publication of the original article (Osborne et al., 2019).

What many of these publications have in common is usage of the General System Justification Scale (G-SJS, Kay & Jost, 2003). This scale has been used in different contexts, age groups, and countries. Such comparisons depend on the scale's measurement equivalence (ME) across the assessed groups and countries. ME ensures that observed differences between groups or countries are based on differences in the construct itself and are not caused by improper translation or different understanding of items by different participant groups (Vandenberg & Lance, 2000). To the extent that the (often implicit) assumption of ME is violated, the validity of the conclusions drawn from comparisons may be compromised. However, although key for cross-group comparisons, ME of the G-SJS has rarely been examined, one exception being Vargas-Salfate et al. (2018) who tested the ME for a shortened scale. The aim of this article was to take a first step into filling this gap by assessing the measurement equivalence of cross-country comparisons building on samples from Great Britain, Germany, and France in Study 1 and 66 samples from 30 countries in Study 2.

Theoretical background

What is system justification?

System justification can be defined as the motivation to bolster and defend the societal status quo (Jost et al., 2017). According to system justification theory, people are motivated to defend and justify the existing economic, social, and political system. For instance, the social system can be justified by evaluating high-status groups as deserving of their positions rather than admitting that the current social order may be unjust. One important aspect of the theory is that it predicts system-justifying motivations not only among those who profit from the societal status quo, mainly high-status persons, but also among those who are disadvantaged by the system. Socially disadvantaged people can justify a system that provides them with little support, meaning that their electoral decisions, therefore, do not necessarily favor parties that would change the system to their advantage; they engage in such justifications to feel better about existing inequalities and power asymmetries (van der Toorn et al., 2011).

After initially focusing on stereotyping and prejudice, system justification theory research expanded to account for a wide range of outcomes (Jost, 2019). For example, system justification theory has been used to explain differences in political ideologies (Jost et al., 2017), justification of downsizing (Richter & König, 2017), and appraisals of pay entitlement (O'Brien et al., 2012). System justification has also been found to be an important contributor to resistance to change (Jost et al., 2017). Additionally, several factors have been identified that predispose individuals to justify the system, including a strong national identification (Carter et al., 2011) and perceived powerlessness (van der Toorn et al., 2015).

Cross-country comparisons and measurement equivalence of the G-SJS

System justification theory assumes that all people are generally motivated to justify the system to a certain extent, but the level of this motivation varies both between and within people (Jost, 2019). The G-SJS (Kay & Jost, 2003) is intended to measure this variability between people in system-justifying tendencies. Although this scale has been widely used as a measure of system justification motivation, it has not been without theoretical critique: Some researchers rather interpret it as a measure of the status quo perception instead of a measure of a motivation (Owuamalam et al., 2018). Furthermore, some items of the scale have been argued to measure other constructs than system justification. In particular, the item "[country name] is the best country to live in" might rather assess national attachment instead of system justification (Owuamalam et al., 2019). National attachment is, however, theoretically different from system justification and also varies between countries (Becker et al., 2017).

Despite these criticisms, the G-SJS has been used to examine system justification in different social groups and countries around the world (e.g., Brandt et al., 2020; Vargas-Salfate et al., 2018; Zmigrod, 2020). Although this cross-country research represents only a fraction of the research on system justification, such research is critical for system justification theory because the theory seeks to describe general processes that are broadly applicable to people in different societies with different political systems. For example, Cichocka and Jost (2014) compared system justification scores between capitalist and post-Communist countries. Another study found that system justification was positively related to life satisfaction and negatively related to depression across 18 countries (Vargas-Salfate et al., 2018). Furthermore, Brandt et al. (2020) assessed participants' social status and perceived legitimacy of the social system across 30 countries and 66 samples. They found that people with higher status assessed the social system as more legitimate than those with lower status. At the same time, they also reported considerable variation across people and countries, emphasising the relevance of ensuring that the measurement instruments used capture the same construct across individuals and countries.

A prerequisite for cross-country comparisons is that the measurement instruments used assess the same construct in the same way in all countries (i.e., the instruments are measurement equivalent or measurement invariant – both terms are used synonymously; Vandenberg & Lance, 2000). This implies that participants from different countries with the same level of the latent construct of interest should provide the same manifest response on the measurement instrument (Hirschfeld & von Brachel, 2014). If the ME has not been established, it is unclear whether and to what extent similar responses indicate similar levels of the latent construct. In this case, results from different countries cannot be easily compared.

The need to establish ME is often overlooked in cross-country and cross-group research (Flake et al., 2017), including research using the G-SJS: System justification motivation and its relations with other constructs might differ across countries or other groups because of true differences in this form of motivation or because respondents in different countries or from different groups other than countries understand the scale items differently. Such differences could be caused by incorrect translations or different interpretations in different political, economic, and social contexts.

A lack of ME can distort statistical conclusions and practical implications based on analyses of observed mean differences between groups (N. Schmitt & Ali, 2014). Hence, manifest differences between group samples might be misattributed to real-group differences, while in fact they merely reflect different interpretations or understandings of the scale items. Conversely, manifest non-differences between group samples might be falsely misattributed to a lack of real differences between groups. Hence, ensuring ME of the G-SJS across countries is important to provide cross-cultural validity to research using this scale (see Osborne et al., 2019, for a discussion of cross-cultural generalisability of the system justification theory).

There are different types of ME: configural, metric, and scalar equivalence (see supplemental online material [SOM] S1 for further explanations; Vandenberg & Lance, 2000). Scalar equivalence is a prerequisite for comparing mean values across groups and for multilevel analyses because it implies that the scales have the same operational definition across groups (Cheung & Rensvold, 2002). Hence, scalar

equivalence would have been needed to compare the means of system justification across countries, as done by Vargas-Salfate et al. (2018) in their preliminary analyses, or conduct multilevel analyses, as in the study by Brandt et al. (2020). However, Vargas-Salfate et al. (2018) reported themselves that the short versions of the G-SJS they used did not show to be invariant; hence, their comparisons of mean values should be interpreted with caution.

Furthermore, the practical importance and magnitude of non-equivalence should also be considered. Statistical evidence for measurement non-equivalence does not necessarily imply severe distortions in the interpretation of measurement scores from different groups, particularly in large samples. Thus, researchers have been advised to calculate an effect size known as "d_{MACS}" (with MACS = mean and covariance structure) on the item level and, as an unstandardised effect size, *impact* on the scale level (Nye & Drasgow, 2011). The main aim of the present study was to examine the measurement equivalence of the G-SJS and to quantify the magnitude of non-equivalence, if any (RQ1).

In addition to its main aim, that is, testing ME, the present study also examined aspects of the G-SJS's convergent validity (Kay & Jost, 2003). Ensuring convergent validity is important to verify the functioning of the scale across countries. Not only should participants in different countries understand and interpret the scale items in the same way (ME) but the scale should also exhibit similar relationships with other relevant constructs across countries (convergent validity).

Two constructs well-suited to test for the G-SJS's convergent validity are political orientation and willingness to strike. System-justifying outcomes are often interpreted as indicators of conservatism (Jost et al., 2017) as the needs underlying system justification are also at the root of political conservatism (Jost et al., 2003) with conservatism defined as containing the two interrelated aspects of resisting social change and accepting equality (Jost et al., 2003, 2009). Jost et al. (2001) concluded that motivation to defend the existing social system is weaker among supporters of left-wing ideology than of right-wing ideology. Conservatism can, thus, be called a "prototypical system–justifying ideology" (Jost et al., 2003, p. 63) because the main components of conservative belief systems focus on acceptance of inequality and opposition to change (Huntington, 1957).

Jost (2019) inferred that system justification is almost always positively related to the endorsement of politically conservative ideologies. This relation has been shown for samples from Great Britain (Zmigrod et al., 2018) and Germany (Jost, 2019). France is the only country in which a negative correlation between system justification and conservatism has been found (Langer et al., 2020): System justification was associated with liberal-socialist attitudes and with liberal/leftist preferences regarding immigration and welfare, contradicting results from other countries and theoretical arguments. Langer et al. (2020) attributed the results from France to the Enlightenment ideals of 'liberté, égalité, and fraternité' stemming from the French revolution, which are still deeply entrenched in France, up to the point that they might represent the societal status quo. Although this might be a valid explanation for their results, varying correlation patterns with key variables across countries could be problematic for the international comparability of the scale. Thus, the question of whether the relation between system justification and political orientation is similar across countries is critical for future research. Next to its relation with political orientation, system justification can have many consequences for political behavior, including participation in collective action (Jost et al., 2017). For example, willingness to protest for change can be undermined by system justification as it fosters resistance to change (Jost et al., 2012). Hence, system-justifying beliefs should be negatively associated with support for protest and willingness to protest. One form of collective action is people's willingness to strike to collectively enforce jointly held economic or other goals against employers. We, therefore, used willingness to strike as a second criterion for the G-SJS's convergent validity in Study 1. Willingness to strike can be a generalisation of dissatisfaction with one part of one's work to other parts and a function of dissatisfaction in many areas (Stagner, 1956). In Study 1, we had the opportunity to examine the convergent validity of the G-SJS, in that we assessed whether the reported associations between system justification and political orientation by Jost (2019) and Langer et al. (2020) could be replicated in different samples (RQ2a) and whether system justification is negatively correlated with willingness to strike in Great Britain, Germany, and France (RQ2b).

In summary, the objective of this research was to examine the ME of the G-SJS. In Study 1, we drew on a sample of three countries: Great Britain, Germany, and France. In Study 2, we extended the scope of our analyses to 30 countries by using the dataset from the study Brandt et al. (2020). Additional analyses examined convergent validity of the G-SJS in Study 1.

STUDY 1 - METHODS

The data were collected in the context of a larger preregistered study pursuing a different research question (https://aspredicted.org/blind.php?x=tx4q7x). Corresponding results can be found in a different article (Vesper & König, in preparation). A list of all collected variables and the data can be found at https://osf.io/34p9j/?view_only=5e248dda2c18460289bf75a71d430c5d.

Sample

Participants were recruited through an online panel provider in Great Britain, Germany, and France and received 0.50 €/0.50 £ as an incentive. A total of 1652 persons completed the study. Since these data were primarily collected for another study concerning attitudes toward strikes, people who were not currently employed were screened out (n = 92). Participants who did not permit their data to be used for scientific purposes were also screened out (n = 33, following Meade & Craig, 2012). Several steps were taken to ensure data quality (Meade & Craig, 2012): First, to deal with overly swift completion, all participants were excluded who completed the items faster than a rate of two seconds per item (n = 78, Huang et al., 2012). Second, if participants chose the same response option for more than six successive items, their data were excluded (n = 88, Niessen et al., 2016). This exclusion procedure was pre-registered. After this procedure, N = 1361 people remained in the final sample.

In the final sample, the mean age was 46.33 (SD = 10.03) and 66.9% of the participants were female. In the British sample (n = 444), the mean age was 46.82 (SD = 10.68) and 65.8% were female. In the German sample (n = 454), the mean age was 44.80 (SD = 10.64) and 65.4% of the participants were female. In the sample from France (n = 463), the mean age was 47.36 (SD = 8.53) and 69.3% were female.

Materials

We used the original eight-item G-SJS (Kay & Jost, 2003) for the British sample, the German translation by Ullrich and Cohrs (2007), and the French translation used in Langer et al. (2020), which we received from P. Vasilopoulos (personal communication, January 17, 2020) to measure *system justification*. Items were rated on a nine-point scale ranging from 1 = "Do not agree at all" to 9 = "Agree completely" (see SOM, Table S1 for a list of all items).

To measure *political orientation*, we used a single item that had been used previously in all three languages, which read: "In politics, people sometimes talk of 'left' and 'right'. Where would you place yourself on this scale, where 1 means the left and 11 means the right?" (Breyer, 2015; European Social Survey Round 2019; Kroh, 2007).

To measure *willingness to strike*, we developed three items based on the study by Akkerman et al. (2013), which were rated on a five-point scale ranging from 1 = "Not at all" to 5 = "Very likely." The items were "I would strike for more money", "I would strike for better working conditions", and "I would strike for better working hours". Following recommended procedures (e.g., Schaffer & Riordan, 2003), we translated this scale from German into English and French via a translation-backtranslation process by two individuals per language who were fluent in both German and English or German and French, respectively. Any differences found between the original items and the back-translated versions were discussed, and agreement was reached on the most appropriate translation.

TABLE 1 Descriptive statistics and internal consistencies of the general system justification scale and willingness to strike scale for the three samples (N_{UK} = 444, N_{DE} = 454, and N_{FR} = 463)

Sample	Scale	*M*	*SD*	Cronbach's α	McDonald's ω
UK	General system justification	4.28	1.43	.81	.82 [.78-.85]
DE	General system justification	4.47	1.61	.88	.88 [.86-.90]
FR	General system justification	3.93	1.54	.85	.85 [.83-.88]
UK	Willingness to strike	3.54	1.18	.91	.91 [.89-.93]
DE	Willingness to strike	3.93	1.09	.90	.90 [.87-.92]
FR	Willingness to strike	3.38	1.21	.87	.87 [.84-.89]
UK	Political orientation[a]	5.72	1.99	--	--
DE	Political orientation[a]	5.53	1.91	--	--
FR	Political orientation[a]	6.10	2.31	--	--

Note: UK = Great Britain, DE = Germany, FR = France. Numbers in brackets represent the 95% confidence interval.
[a]One-item scale.

Table 1 shows descriptive statistics and reliability scores in terms of Cronbach's α and McDonald's ω (McDonald, 1999) for the G-SJS, willingness to strike, and political orientation for all three samples.

Procedure

At the beginning, participants chose their preferred language. After that, a welcome page explained the purpose of the study. This page was followed by socio-demographic questions. Here, participants who indicated that they were not currently employed were screened out. Next, all other participants filled out various scales including the G-SJS (Kay & Jost, 2003) and several others not of interest for present purposes. At the end, the participants were asked whether their data could be used for scientific purposes, were thanked for their participation, and returned to the online panel provider site to receive their compensation.

Statistical analyses: tests of measurement equivalence

Statistical analyses were carried out with R 3.6.1 (R Core Team, 2019) and additional R packages: careless (Yentes & Wilhelm, 2018), MBESS (Kelley, 2019), lavaan (Rosseel, 2012), dmacs (Dueber, 2019), sem (Fox et al., 2017), and semTools (Jorgensen et al., 2019). To assess the model fit in the confirmatory factor analyses (CFAs), we used the comparative fit index (CFI), Tucker–Lewis index (TLI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR). To evaluate model fit, we followed recommendations by Hu and Bentler (1999) who consider cutoff values for CFI and TLI ≥ .95, SRMR ≥ .08, and RMSEA ≥ .06 as indicating a good model fit.

To evaluate ME, the standard three-step process using multigroup CFAs was followed (see also the SOM, S1 for further explanations on statistical analyses of ME; Hirschfeld & von Brachel, 2014; Vandenberg & Lance, 2000). As $\Delta\chi^2$ highly depends on sample size, ΔCFI was used to compare models (Cheung & Rensvold, 2002): The equivalence hypothesis was rejected if changes in CFI of −.01 or more between the tested model and the less constrained model were observed.

Statistical analyses: measurement equivalence effect sizes

ME analyses can be supplemented by effect sizes in order to assess the magnitude and practical importance of non-equivalence (Nye & Drasgow, 2011). We calculated so-called *impact* on the scale level and d_{MACS} on the item level, as suggested by Nye and Drasgow (2011). In this study, we focused on the scale-level effect size of *impact*; all analyses regarding d_{MACS} can be found in the SOM (S3) as well as a more detailed description on how to calculate *impact* (S1). On the scale level, *impact* can be calculated as an unstandardised effect size to assess the magnitude of non-equivalence. *Impact* reflects the construct-relevant differences in the measure (Ock et al., 2020). In order to estimate *impact*, one must first calculate Δ*mean*. Δ*mean* reflects the number of observed differences in mean composite scores between the assessed groups that can be attributed to non-equivalence. To estimate Δ*mean*, a researcher sums up the differences in item means that can be attributed to non-equivalence between the referent and focal groups (Nye & Drasgow, 2011). Negative Δ*mean* values indicate that differential item functioning (DIF, i.e., whether the items work in the same way in the groups, Janssen, 2011) results in higher means for the focal group than the referent group. *Impact* is then estimated as the difference between the observed differences and Δ*mean* between the focal group and the referent group.

STUDY 1 - RESULTS

Preliminary analyses can be found in the SOM (S2). Based on these, the reverse-coded items "British/French/German society needs to be radically restructured" (Item 3) and "Our society is getting worse every year" (Item 7) were excluded from the following analyses because the CFAs of all three samples showed a barely acceptable fit when these items were included.

Results RQ 1: measurement equivalence of the G-SJS

To assess the ME of the G-SJS, we conducted three analytic steps. First, we tested for configural equivalence. Configural equivalence means that the number of latent variables and the latent variables' pattern of loadings on the indicators is the same in all groups (Chen, 2008). The configural model showed just acceptable model fit according to the suggestions of Hu and Bentler (1999), $\chi^2(27) = 193.26$, $p < .001$, CFI = .948, TLI = .91, RMSEA = .12, 90% CI [.10,.13], SRMR = .03; hence, full configural equivalence was established.

Second, we tested for metric equivalence, that is, the assumption that the factor loadings have similar magnitude across groups. This form of equivalence is required to compare the relationships between latent variables across different groups (Chen, 2008). The model for metric equivalence showed also just acceptable fit, $\chi^2(37) = 275.69$, $p < .001$, CFI = .926, TLI = .91, RMSEA = .12, 90% CI [.11, .13], SRMR = .07. In contrast to the configural model, the change in the CFI was ΔCFI = −.023 and, therefore, larger than the threshold of ΔCFI = −.01 (Cheung & Rensvold, 2002).

Before concluding that this should be interpreted as evidence against metric equivalence, we inspected the modification indices of the metric equivalence model and tested for partial metric equivalence by sequentially releasing individual loading constraints and retesting the model (Vandenberg & Lance, 2000). With respect to metric equivalence, releasing the constraints for items means that all items are still required to load onto the same factors in each sample, but the requirement that the loadings be of the same magnitude across samples can be dropped for some items (Vandenberg & Lance, 2000). The item "In general, the British/German/French political system operates as it should" (Item 2) had the highest modification indices. Thus, we relaxed the constraints on this item and tested again for metric equivalence. This model showed a little improvement (ΔCFI = −.011). As the change in CFI was still above the cutoff of −.01, we additionally released the constraints on the item "Most policies serve the greater good" (Item 5). Releasing the constraints for Item 2 and Item 5 significantly improved the model

fit (ΔCFI = −.004). Following Steenkamp and Baumgartner (1998) and Vandenberg and Lance (2000), a factor can be assumed to be partially equivalent if more than half of the items loading onto the factor are equivalent. Hence, as the constraints for two of the six items were released, partial metric equivalence for the one-factor model was found.

Finally, we tested for scalar equivalence. Scalar equivalence implies that the item intercepts are also similar across groups. Hence, there should be no systematic response biases. This form of ME is necessary in order to meaningfully compare the means of the latent variables across different groups (Chen, 2008). The scalar equivalence model showed just acceptable fit, $\chi^2(39) = 342.29$, $p < .001$, CFI = .906, TLI = .89, RMSEA = .13, 90% CI [.12, .14], SRMR = .07. Compared with the partial metric model, the change in CFI equalled ΔCFI = −.039, greater than the threshold of ΔCFI = −.01, indicating that the scalar model had worse fit than the partial metric model (Cheung & Rensvold, 2002). A further release of item constraints to test for partial scalar equivalence was not possible, as two of the six remaining items had already been released of their constraints, and releasing a third item would have led to half of the items being freed, which Steenkamp and Baumgartner (1998) and Vandenberg and Lance (2000) do not recommend. Thus, scalar measurement equivalence could not be established.

Taken together, we found partial metric equivalence, but no scalar equivalence for the scale in the three countries. Hence, we can meaningfully compare difference scores on the items across these three countries (Steenkamp & Baumgartner, 1998), but comparing means across the three countries is not warranted (but see the section below on effect sizes of non-equivalence). Remember that these results pertain to the already reduced scale after excluding two of eight items that led to poor model fit in preliminary analyses.

Bilateral country comparisons

Next, we tested ME for two out of three countries. For Great Britain and France, we found partial scalar equivalence after releasing the constraints of Item 1 ("In general, I find society to be fair") and Item 5, $\chi^2(26) = 133.44$, $p < .001$, CFI = .942, TLI = .93, RMSEA = .10, 90% CI [.08, .11], SRMR = .06, ΔCFI = −.009, respectively. For Great Britain and Germany, we also found partial scalar equivalence when the constraints for Items 2 and 5 were released, $\chi^2(24) = 179.35$, $p < .001$, CFI = .928, TLI = .91, RMSEA = .12, 90% CI [.10, .14], SRMR = .05, ΔCFI = −.008, respectively. Finally, for Germany and France, we found partial metric equivalence, but no scalar equivalence. Partial metric equivalence was achieved by releasing the constraints of Item 1, $\chi^2(22) = 161.82$, $p < .001$, CFI = .942, TLI = .92, RMSEA = .12, 90% CI [.10, .14], SRMR = .05, ΔCFI = −.007. To achieve partial scalar equivalence, the constraints of half of the items would have needed to be released.

Effect sizes of measurement non-equivalence

To calculate the effect sizes of measurement non-equivalence, we first chose Great Britain as the referent group for comparing the samples from Great Britain and Germany and Great Britain and France because the original English version of the scale was used in Great Britain (following Nye & Drasgow, 2011). Next, we chose Germany as the referent group for the German–French comparison. The item "Society is set up so that people usually get what they deserve" (Item 8) was chosen as the referent item, following suggestions on how to choose a referent item by Nye and Drasgow (2011). These criteria state that *n*-1 tests of equivalence are conducted for each of the items in the scale. The only thing that varies across these tests is the item serving as the referent in each test. An item is chosen as referent item if this item is equivalent across each of the *n*-1 tests of equivalence. This resulted in three comparisons of five items; hence, fifteen item comparisons in total.

On the scale level, the *impact* for the British–German comparison was −0.83; this refers to the true construct-relevant difference between these two samples. This indicates that the German sample has

TABLE 2 Effects of non-equivalence on scale-level properties

Scale	*Impact*	Δ*mean*	Observed difference	% of observed difference due to Δ*mean*
UK[a] -DE	−0.83	0.28	−0.55	51%
UK[a]-FR	2.09	−1.10	0.99	111%
DE[a]-FR	2.91	−1.38	1.53	90%

Note: Items were z-standardised. *Impact* refers to the true differences in the construct. Negative values in Δ*mean* indicate that DIF (differential item functioning) results in higher means for the focal group than the referent group. UK = Great Britain, DE = Germany, FR = France.
[a]Referent group.

higher system justification tendencies than the British sample. The *impact* was larger than the observed difference between these two groups (−0.55), indicating that non-equivalence inflated the British scores. The amount of observed difference that can be attributed to DIF was, hence, Δ*mean* = 0.28 for this comparison (see Table 2). Δ*mean* was in the opposite direction of the *impact* (*impact* = −0.83), indicating that non-equivalence reduced the observed differences in system justification relative to their true mean differences, thus resulting in an observed difference that was smaller than the actual construct-relevant difference (i.e., *impact*). The results can be found in Table 2. Results regarding the item level effect sizes of measurement non-equivalence can be found in the SOM (S3 and Table S2).

For the British–French comparison, the *impact* was 2.09. Hence, the British sample had higher system justification tendencies than the French sample. The observed difference between these two groups was 0.99, suggesting that the true difference between these two samples was actually larger than the observed difference indicated. The amount of the observed difference attributable to DIF was Δ*mean* = −1.10. The negative value indicates that DIF resulted in a higher mean in the French sample than the British sample. Moreover, the Δ*mean* for this comparison was negative, whereas the observed difference was positive (0.99). Hence, non-equivalence reduced the differences between the two samples by inflating the French scores.

For the German–French comparison, the *impact* was 2.91, indicating that the German sample had higher system justification tendencies than the French sample. The observed difference was 1.53, and the amount of the observed difference that could be attributed to DIF was Δ*mean* = −1.39. This indicates that the DIF resulted in higher item means in the French sample than the German sample. As in the British–French comparison, the Δ*mean* was negative, while the observed mean difference was positive. Thus, the true difference between these two countries was actually larger than the observed difference indicated, which is shown by *impact* = 2.91. Hence, the non-equivalence reduced the differences between the two samples by inflating the scores in the French sample.

Results RQ 2: convergent validity of the G-SJS

To answer RQ2a, that is, whether the associations between system justification and political orientation (i.e., conservatism) in the three countries could be replicated, correlations between political orientation and system justification scores were calculated for each sample. The system justification score was calculated without Items 3 and 7 as these two items had been excluded after the CFAs. Table 3 shows the results. System justification was significantly positively correlated to a right-wing political orientation in the British sample (replicating previous results). The same relation between these two constructs was also found in our French sample (contrary to previous results), but was less than half as strong, and failed to pass the significance threshold of $p < .05$. In the German sample, the correlation between system justification and political orientation was descriptively negative (contrary to previous results) and not significant, too. Thus, the RQ2a results were inconclusive.

RQ2b concerned the relationship between system justification and willingness to strike. For the French sample, a significant negative correlation was found (Table 3). The British and the German samples

TABLE 3 Correlations between system justification, political orientation, and willingness to strike in the three samples

	1 UK/ DE/ FR	2 UK/ DE/ FR	3 UK/ DE/ FR
1. System justification	-		
2. Political orientation	.25***/ −.07.^{117}/.09$^{.058}$	-	
3. Willingness to strike	−.06.$^{.247}$/ −.09.$^{.051}$/ −.21***	−.27***/−.06.$^{.183}$/−.18***	-
M	4.28/ 4.47/ 3.93	5.72/ 5.53/ 6.10	3.54/ 3.93/ 3.38
SD	1.43/ 1.61/ 1.54	1.99/ 1.91/ 2.31	1.18/ 1.09/ 1.21

Note: N_{UK} = 444, $N_{Germany}$ = 454, N_{France} = 463. UK = Great Britain, DE = Germany, FR = France. Higher values in political orientation correspond to a more conservative political ideology. The superscripted numbers represent the significance level of the corresponding correlations.

*p < .05, **p < .01, and ***p < .001.

exhibited considerably smaller, non-significant negative relationships, although the correlation in the German sample was close to significance. Hence, RQ2b can be answered by stating that system justification was negatively related with willingness to strike in the French, but not in the British and German samples.

STUDY 1 – DISCUSSION

When comparing all three groups, partial metric equivalence was reached but not full scalar equivalence. When considering two-country comparisons, we found partial scalar equivalence for the comparisons between Great Britain and France and Great Britain and Germany, respectively. Hence, in these pairings, comparisons of mean values were allowed; the comparison of mean values between Germany and France was not justified. Note, however, that we needed to exclude two reverse-coded items from the scale that negatively impacted baseline model fit and to additionally free several items from their constraints to achieve this partial scalar equivalence. In addition, the *impact* (i.e., the effect of non-equivalence on the scale level) of the comparisons between the French sample and the other two samples indicated that the observed scores were influenced by non-equivalence. The percentage of observed mean difference attributable to DIF ranged from 51% (UK–German comparison) to 111% (UK–French comparison). A percentage of 111% suggests that the effects of DIF are larger than the observed mean differences in this case. For the British–French comparison, Δ*mean* was negative while the observed mean difference was positive. This shows that the referent group (the British sample) had a higher observed mean, but the effects of DIF lowered this effect by increasing the mean of the focal group (the French sample). Thus, the true difference was larger than the observed differences indicated, and conclusions based on the observed descriptive statistics might be affected. This also applies to the other two comparisons as 50% and 90% of observed mean difference attributable to DIF hinted at substantial influences of DIF in the scale values. On the item level, several items were considered problematic (see Table S3 in the SOM for further information).

The relation between system justification and political orientation emerged only partly as expected (with higher political orientation scores indicating greater conservatism). Previous research found positive correlations in Great Britain (Zmigrod et al., 2018) and Germany (Jost, 2019) and a negative correlation in France (Langer et al., 2020). In the present study, system justification was significantly positively correlated with political orientation in Great Britain, non-significantly negatively correlated in Germany, and non-significantly positively correlated in France. Thus, these results differ from previous findings for two of the three countries and are inconsistent with the assumption that conservatism is a "prototypical system-justifying ideology" (Jost et al., 2003, p. 63).

Perhaps, the German participants perceived the status quo in their country as already rather left/liberal based on policies such as the rather pro-refugee political agenda of the German federal government

at the time. Thus, German conservatives might have reported relatively low system justification motivation and been motivated to challenge the status quo, for example, by changing over a more restrictive refugee policy. Then again, Germany had been led by a conservative chancellor and party for 15 years at the time of data collection, speaking against the idea that Germans perceived their country as particularly leftist. Thus, there is no easy explanation for the descriptively negative correlation between system justification and political ideology in Germany.

The reasoning that the current status quo is perceived as rather left/liberal aligns with Langer et al.'s (2020) arguments with respect to France as these researchers also found a negative correlation between system justification and conservatism in their French sample. However, the French participants in our study did not follow this pattern. Additionally, some research has found a negative quadratic relationship between system justification and political conservatism in a European setting (Caricati, 2019). Hence, further research is needed to investigate the relationship between political conservatism and system justification as the association might not be as straightforward as previously assumed.

The relation between system justification and willingness to strike as a proxy for participation in collective action was negative in all three countries, as expected. However, these correlations were small and not significant in Great Britain and Germany. Thus, these findings only partially align with the assumption that people who are less system-justifying are less willing to accept inequality and to participate in collective action (Jost et al., 2017).

STUDY 2 – METHODS

In Study 2, we broadened the scope of our analyses to ensure that our results were valid and not due to peculiarities of Study 1. To this end, we used the dataset from Brandt et al. (2020), which contains 66 samples from 30 countries. The dataset is available at https://osf.io/qw47m/. Note that in Study 2, we used a different method to assess ME – the alignment optimisation method (Asparouhov & Muthén, 2014; Magraw-Mickelson et al., 2020; Marsh et al., 2018). This method is better suited for comparisons of more than three samples. On the downside, it does not allow for calculating ME effect sizes.[1]

Sample

A total of 14,469 persons were included in the dataset, and of these, only those who had no missing values in the G-SJS scale (N = 13,494) were included in our analyses. The mean age was 25.05 (SD = 10.48), and 66.6% of the participants were female. See Brandt et al. (2020) for a further description of their studies (but note that some figures might differ as we did not exclude participants who had missing values in other constructs than system justification).

Materials

Brandt et al. (2020) used the original eight-item G-SJS (Kay & Jost, 2003) and own translations to measure *system justification*. Items were rated on a self-developed seven-point response scale ranging from −3 = "Disagree strongly" to +3 = "Agree strongly". Note that this response scale differed from the one used in the original scale (and in Study 1) that ranged from 1 = "Do not agree at all" to 9 = "Agree completely".

[1]We also used the alignment method for the data from Study 1 to ensure that the results were not confounded by the used statistical method. We obtained metric, but no scalar equivalence for the 8-item version as well as the 6-item version. The degree of non-equivalence across parameters was 45.8% of intercepts (8-item version) and 44.4% of intercepts (6-item version), respectively. Thus, the results corroborated those reported in Study 1 using the MG-CFA procedure.

Statistical analyses: tests of measurement equivalence

Statistical analyses were carried out with R 4.0.5 (R Core Team, 2019) and additional R packages: dplyr (Wickham et al., 2021), MBESS (Kelley, 2019), lavaan (Rosseel, 2012), psych (Revelle, 2020), and sirt (Robitzsch, 2020). To assess the model fit in the CFAs, we used the comparative fit index (CFI), Tucker–Lewis index (TLI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR). Again, we followed recommendations by Hu and Bentler (1999) who consider cutoff values for CFI and TLI ≥ .95, SRMR ≥ .08, and RMSEA ≥ .06 as indicating a good model fit.

We used the alignment optimisation method (Asparouhov & Muthén, 2014; Magraw-Mickelson et al., 2020; Marsh et al., 2018) to assess that ME that is well-suited for comparisons of more than three samples. We used the sirt-package (Robitzsch, 2020) to conduct the alignment optimisation method in R and followed recommendations of Fischer and Karl (2019) regarding settings of tolerance and alignment power. This method assumes rather approximate than exact invariance. Alignment optimisation starts with a common configural model that contains all groups (Magraw-Mickelson et al., 2020). In this configural model, the intercepts and loadings are unconstrained instead of a separate baseline model for each group as with the MG-CFA. Starting from this configural model, the process uses maximum likelihood (ML) estimation to fit an optimal set of measurement parameters and then computes approximations based on that. Based on the optimal model, means and variances of the latent variable can be computed.

The alignment method is executed in two steps: First, a configural model representing the best fitting model among all groups is established while fixing the factor means to 0 and fixing variances to 1, without constraining the loadings or intercepts (Magraw-Mickelson et al., 2020). Thus, alignment optimisation works with the assumption that there is a degree of non-invariance and its goal is to keep this non-invariance to a minimum. Second, the factor means and variances are freely estimated and undergo an optimisation process for every group factor mean and item parameter (Asparouhov & Muthén, 2014). When the minimisation point has been reached, a researcher can compare factor means and factor variances across groups using a "post-estimation algorithm" and identifies each model parameter such as loadings and intercepts that is significantly different from the average of that parameter across all groups. Those estimates that are significantly different are flagged as non-invariant. The output then provides the latent means based on this model plus the parameters flagged as non-invariant (Asparouhov & Muthén, 2014; Magraw-Mickelson et al., 2020). Thus, the alignment process allows for the estimation of reliable means despite the presence of some measurement non-invariance. As a threshold, Asparouhov and Muthén (2014) recommended 20% non-invariant parameters as acceptable.

STUDY 2 – RESULTS

Preliminary analyses regarding the CFAs for every sample of the dataset can be found in the SOM (Table S4). We found metric equivalence for the overall sample, but no scalar equivalence. The degree of non-equivalence across parameters was 55.4% of intercepts (that are crucial to achieve scalar equivalence) and hence above the recommended threshold of 20% (Asparouhov & Muthén, 2014). Based on the results of Study 1, we also conducted the analyses without the recoded items 3 and 7. Again, metric equivalence was established, but no scalar equivalence, and the percentage of non-equivalent item parameters was 52.8% of intercepts. Further details and the calculations are available in the SOM (S1 and Table S4).

STUDY 2 – DISCUSSION

Using the alignment optimization method, metric, but not scalar equivalence was established for the 30 countries in the dataset from Brandt et al. (2020) where a modified response scale was used. These

findings are consistent with those from Study 1. They indicate that the validity of comparisons of G-SJS means across the assessed countries is questionable with the current version of the scale.

GENERAL DISCUSSION

The current article investigated the applicability of the G-SJS (Kay & Jost, 2003) in different countries by assessing measurement equivalence in samples from Great Britain, Germany, and France in Study 1 and across 30 countries in Study 2, as well as its convergent validity in Study 1. In Study 1, the one-factor model exhibited good fit in all three samples after excluding the two reverse-coded items, and the G-SJS was partially metric equivalent across the three countries. Partial metric equivalence allows for meaningful comparisons of difference scores on the items across the three countries (for an overview of practical questions regarding ME, see Table 4; Steenkamp & Baumgartner, 1998). We did not find scalar equivalence across the three countries, indicating that the construct mean values in the three countries cannot be readily compared. Focusing on bilateral comparisons allowed us to calculate effect sizes of measurement non-equivalence. For some comparisons, these were substantial.

In Study 2, we also found metric, but not scalar equivalence across the 30 countries using an analysis method well-suited for comparisons of larger numbers of samples (Asparouhov & Muthén, 2014). Note that this method did not allow to estimate effect sizes of non-equivalence. The convergent validity analyses in Study 1 were rather inconclusive as the relation between system justification and political orientation showed only the expected direction in the British sample. Furthermore, the relationship between system justification and willingness to strike was only as expected in the French sample.

What are the substantial implications of the present research for researchers not specifically interested in measurement issues, but in using the scale in their research? Based on the results of our two studies, we do not recommend calculating an ANOVA to compare scale means of several countries. Due to lack of scalar equivalence in the present samples, these comparisons might be distorted by construct-irrelevant variance in other samples as well. Hence, comparisons such as those made by van der Toorn et al. (2010) between a U.S. and a Hungarian sample should be interpreted with caution. For such comparisons, the ME should first be established. The same applies to comparisons that are made

TABLE 4 Questions and answers regarding further use of the general system justification scale

Question	Answer
May I calculate an ANOVA for the scale means of several countries?	The lack of scalar equivalence does not permit an ANOVA to be conducted. Comparisons would be distorted by this construct-irrelevant variance.
May I compare the scale's correlations with external criteria across countries?	As scalar equivalence is also a prerequisite for interpreting differences in variances, the answer is no. With differences in variances, we mean the variance in mean G-SJS values within a given sample, which in turn enters the correlation.
May I assume the same factor structure of the scale in different countries?	The one-factor model exhibited good fit in all three countries in Study 1. Insofar, the answer is yes. Nevertheless, the items used to assess this factor do not seem well-suited for every country (which can be seen in Table S4). Future research should always start with a CFA for each sample, followed by a test of measurement equivalence via multigroup confirmatory analysis or alignment optimisation.
May I assume that all items of the scale are similarly important in different countries?	First, the negatively worded Items 3 and 7 should be improved as they impaired the model fit in all three samples in Study 1. However, in Study 2, they did not have a huge impact on the non-equivalence. Second, the other items showed different effect sizes of non-equivalence in the three comparisons in Study 1; therefore, which items are similarly interpreted and which are not seems to depend on the specific samples in question. This was also found in Study 2 where Item 4 had most unique item parameters per item, followed by Items 6 and 3.

between capitalist and post-Communist societies (Cichocka & Jost, 2014) and the multilevel analyses comparing the relations of subjective status and perceived legitimacy across countries as in Brandt et al. (2020). If measurement invariance is not established before interpreting the results, one cannot be sure whether the observed differences might not (at least partly) be based on improper translations or different interpretations of items across groups. The same applies to the comparison of the scale's correlations with external criteria across countries. This is why we did not compare the correlations between the system justification scores of the three countries' political orientation and willingness to strike across the three countries in Study 1. Instead, we focused on interpreting each correlation in isolation. To make these implications easily accessible, Table 4 summarises pertinent questions for further use of the scale along with answers based on the results of the present studies.

Our statistical results also align with previous criticism of the scale from a theoretical standpoint: Owuamalam et al. (2019) criticised that some items measure rather national attachment, which is a theoretically different construct than system justification. National attachment has also been shown to vary between countries (Becker et al., 2017). Hence, differences in these items might not be based on differences in system justification, but rather in the concept of national attachment to the respective group. In accordance with this, we found that the item "The United Kingdom/Germany/France is the best country in the world to live in" exhibited the most unique item parameters in Study 2 and should, thus, be considered for further improvement. Hence, the earlier theoretical critique of this type of item and our empirical results independently led to similar conclusions. These should be considered for the further development of the scale.

Another point regarding the items is that we had to exclude the two negatively worded items before conducting the MG-CFAs in Study 1 to achieve a satisfactory model fit in all three countries. This highlights the common problem in ME testing that negatively worded items are often interpreted differently across countries, which might distort the intended unidimensionality of a scale by creating a separate factor (e.g., Herche & Engelland, 1996; Lindwall et al., 2012; D. P. Schmitt & Allik, 2005). Based on Study 1, the two negatively worded items should be considered for further improvement or discarded from the scale altogether (both alternatives would, admittedly, require new validation work, Flake et al., 2017). Note, however, that these negatively worded items did not pose particular problems in Study 2.

Not only the statistical but also the practical impact of ME findings (i.e., effect sizes) should be considered (N. Schmitt & Ali, 2014), which was possible in Study 1. The effect sizes showed that the non-equivalence on the scale level between France and the other two countries was particularly substantial. For example, the non-equivalence resulted in item means that were 1.39 standard deviations higher in the French sample than the German sample. Hence, changes to the scale items might be required before the scale can be used for valid cross-country comparisons.

Finally, the aim of the present research was to elucidate that considering ME is an important issue when working with the G-SJS (Kay & Jost, 2003). To this end, we conducted cross-country comparisons of system justification as measured with the G-SJS. However, the implications of our findings go well-beyond cross-country comparisons. To be sure, such comparisons comprise only a small part of research on system justification. Importantly, the same logic applies to all kinds of comparisons of system justification tendencies across groups as measured with the G-SJS, within the same country or across countries (e.g., different ethnic groups, high- and low-status groups, Jost et al., 2003). Thus, future research should also assess whether the G-SJS is measurement equivalent in such comparisons and not only across countries.

Limitations and future research

Several limitations need to be mentioned. First, although the samples in Study 1 were comparatively large (Ns > 440), they had other limitations. This is evident from the fact that a considerable number of participants had to be excluded – for reasons such as too swift completion or selecting identical response options across a large number of consecutive items. By excluding these participants, we tried to enhance the quality of our samples and ensure that our analyses are valid. Nevertheless, it might be

useful to further validate our results in other, ideally representative, samples for each country. Second, Brandt et al. (2020) used a different response scale than the original G-SJS and the one we used in Study 1. Despite the converging evidence, in terms of ME, the results might be affected by this difference between studies and also by the difference between the used analysis methods. To address this issue, we conducted the analyses from Study 1 with both the MG-CFA and the alignment method and obtained comparable results, hinting that this should not be a major problem (see SOM, S4).

Future research on the G-SJS (Kay & Jost, 2003) and system justification motivation, in general, could address several areas. In particular, the results from the German sample (i.e., the negative relation between system justification and conservatism) should be replicated. If found again, researchers should investigate the reasons for this unexpected relation. It would also be interesting to assess whether this negative relationship emerges in additional countries, particularly countries that are known to have a rather left/liberal status quo. Finally, further research could also investigate the differences we found between Germany and France in more depth to disentangle whether these differences are due to lack of ME (i.e., different item formulations or item understandings) or real differences in the conceptualisation of system justification.

CONCLUSION

The main aim of the present research was to assess the measurement equivalence of the G-SJS (Kay & Jost, 2003). Based on two studies (Study 1: three countries, Study 2: 30 countries), our results indicate that it is not justified to compare mean values across countries – in fact, the scale-level effects of non-equivalence were quite large. Thus, caution is required when comparing samples from different countries using the G-SJS in its current form.

ACKNOWLEDGMENT

Open access funding enabled and organized by ProjektDEAL.

CONFLICT OF INTEREST

We have no conflict of interests to disclose. We would like to thank Keri Hartman for proof-reading this document.

AUTHOR CONTRIBUTION

Denise Vesper: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Project administration (equal); Software (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Cornelius J. König:** Conceptualization (equal); Funding acquisition (equal); Resources (equal); Supervision (equal); Writing – review & editing (equal). **Rudolf Siegel:** Formal analysis (equal); Software (equal). **Malte Friese:** Conceptualization (equal); Supervision (equal); Writing – review & editing (equal).

OPEN RESEARCH BADGES

This article has earned an Open Data, for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://osf.io/34p9j/?view_only=82053bc8f56e443bbab03bc92c14077f.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Open Science Framework at https://osf.io/34p9j/?view_only=5e248dda2c18460289bf75a71d430c5d (Study 1) and https://osf.io/qw47m (Study 2).

ORCID

Denise Vesper https://orcid.org/0000-0002-1585-9243
Cornelius J. König https://orcid.org/0000-0003-0477-8293
Rudolf Siegel https://orcid.org/0000-0002-6021-804X
Malte Friese https://orcid.org/0000-0003-0055-513X

REFERENCES

Akkerman, A., Born, M. J., & Torenvlied, R. (2013). Solidarity, strikes, and scabs: How participation norms affect union members' willingness to strike. *Work and Occupations*, *40*(3), 250–280. https://doi.org/10.1177/0730888413481481

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

Becker, J. C., Butz, D. A., Sibley, C. G., Barlow, F. K., Bitacola, L. M., Christ, O., Khan, S. S., Leong, C.-H., Pehrson, S., Srinivasan, N., Sulz, A., Tausch, N., Urbanska, K., & Wright, S. C. (2017). What do national flags stand for? An exploration of associations across 11 countries. *Journal of Cross-Cultural Psychology*, *48*(3), 335–352. https://doi.org/10.1177/0022022116687851

Brandt, M. J., Kuppens, T., Spears, R., Andrighetto, L., Autin, F., Babincak, P., Badea, C., Bae, J., Batruch, A., Becker, J. C., Bocian, K., Bodroža, B., Bourguignon, D., Bukowski, M., Butera, F., Butler, S. E., Chryssochoou, X., Conway, P., Crawford, J. T., … Zimmerman, J. L. (2020). Subjective status and perceived legitimacy across countries. *European Journal of Social Psychology*, *50*(5), 921–942. https://doi.org/10.1002/ejsp.2694

Breyer, B. (2015). Left-Right self-placement (ALLBUS). *Zusammenstellung sozialwissenschaftlicher Items und Skalen [Compilation of Social Science Items and Scales]*. https://doi.org/10.6102/zis83

Caricati, L. (2019). Evidence of decreased system justification among extreme conservatives in non-American samples. *The Journal of Social Psychology*, *159*(6), 725–745. https://doi.org/10.1080/00224545.2019.1567455

Carter, T. J., Ferguson, M. J., & Hassin, R. R. (2011). Implicit nationalism as system justification: The case of the United States of America. *Social Cognition*, *29*(3), 341–359. https://doi.org/10.1521/soco.2011.29.3.341

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*(5), 1005–1018. https://doi.org/10.1037/a0013193

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Cichocka, A., & Jost, J. T. (2014). Stripped of illusions? Exploring system justification processes in capitalist and post-Communist societies. *International Journal of Psychology*, *49*(1), 6–29. https://doi.org/10.1002/ijop.12011

Dueber, D. (2019). *dmacs: Measurement nonequivalence effect size calculator (0.1.0) [Computer software]*. Retrieved from https://github.com/ddueber/dmacs

European Social Survey Round 9 (2019). *ESS-9 2018 documentation report*. Edition 1.2. European Social Survey Data Archive, NSD - Norwegian Centre for Research Data for ESS ERIC. https://doi.org/10.21338/NSD-ESS9-2018

Fischer, R., & Karl, J. A. (2019). A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Frontiers in Psychology*, *10*, 1507. https://doi.org/10.3389/fpsyg.2019.01507

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Fox, J., Zhenghua, N., & Byrnes, J. (2017). *sem: Structural equation models (v3.1-9) [R package version 3.1-9]*. Retrieved from https://CRAN.R-project.org/package=sem

Herche, J., & Engelland, B. (1996). Reversed-polarity items and scale unidimensionality. *Journal of the Academy of Marketing Science*, *24*(4), 366–374. https://doi.org/10.1177/0092070396244007

Hirschfeld, G., & von Brachel, R. (2014). Improving multiple-group confirmatory factor analysis in R: A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research, and Evaluation*, *19*(1), https://doi.org/10.7275/qazy-2946

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Huntington, S. P. (1957). Conservatism as an ideology. *The American Political Science Review*, *51*(2), 454–473. https://doi.org/10.2307/1952202

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2019). *semTools: Useful tools for structural equation modeling (R package version 0.5-2) [Computer software]*. Retrieved from https://CRAN.R-project.org/package=semTools

Jost, J. T. (2019). A quarter century of system justification theory: Questions, answers, criticisms, and societal applications. *British Journal of Social Psychology*, *58*(2), 263–314. https://doi.org/10.1111/bjso.12297

Jost, J. T., & Banaji, M. R. (1994). The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology*, *33*(1), 1–27. https://doi.org/10.1111/j.2044-8309.1994.tb01008.x

Jost, J. T., Becker, J., Osborne, D., & Badaan, V. (2017). Missing in (collective) action: Ideology, system justification, and the motivational antecedents of two types of protest behavior. *Current Directions in Psychological Science*, *26*(2), 99–108. https://doi.org/10.1177/0963721417690633

Jost, J. T., Blount, S., Pfeffer, J., & Hunyady, G. (2003). Fair market ideology: Its cognitive-motivational underpinnings. *Research in Organizational Behavior*, *25*, 53–91. https://doi.org/10.1016/S0191-3085(03)25002-4

Jost, J. T., Burgess, D., & Mosso, C. O. (2001). Conflicts of legitimation among self, group, and the system: The integrative potential of system justification theory. In J. T. Jost, & B. Major (Eds.), *The psychology of legitimacy: Emerging perspectives on ideology, justice, and intergroup relations* (pp. 363–388). Cambridge University Press.

Jost, J. T., Chaikalis-Petritsis, V., Abrams, D., Sidanius, J., van der Toorn, J., & Bratt, C. (2012). Why men (and women) do and don't rebel: Effects of system justification on willingness to protest. *Personality and Social Psychology Bulletin*, *38*(2), 197–208. https://doi.org/10.1177/0146167211422544

Jost, J. T., Federico, C. M., & Napier, J. L. (2009). Political ideology: Its structure, functions, and elective affinities. *Annual Review of Psychology*, *60*(1), 307–337. https://doi.org/10.1146/annurev.psych.60.110707.163600

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin*, *129*(3), 339–375. https://doi.org/10.1037/0033-2909.129.3.339

Jost, J. T., Pelham, B. W., Sheldon, O., & Sullivan, B. N. (2003). Social inequality and the reduction of ideological dissonance on behalf of the system: Evidence of enhanced system justification among the disadvantaged. *European Journal of Social Psychology*, *33*(1), 13–36. https://doi.org/10.1002/ejsp.127

Kay, A. C., & Jost, J. T. (2003). Complementary justice: Effects of "poor but happy" and "poor but honest" stereotype exemplars on system justification and implicit activation of the justice motive. *Journal of Personality and Social Psychology*, *85*(5), 823–837. https://doi.org/10.1037/0022-3514.85.5.823

Kelley, K. (2019). *MBESS: The MBESS R Package (4.6.0) [Computer software]*. Retrieved from https://CRAN.R-project.org/package=MBESS

Kroh, M. (2007). Measuring left–right political orientation: The choice of response format. *Public Opinion Quarterly*, *71*(2), 204–220. https://doi.org/10.1093/poq/nfm009

Langer, M., Vasilopoulos, P., McAvay, H., & Jost, J. T. (2020). System justification in France: Liberté, égalité, fraternité. *Current Opinion in Behavioral Sciences*, *34*, 185–191. https://doi.org/10.1016/j.cobeha.2020.04.004

Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., Raudsepp, L., Liukkonen, J., & Thøgersen-Ntoumani, C. (2012). Method effects: The problem with negatively versus positively keyed items. *Journal of Personality Assessment*, *94*(2), 196–204. https://doi.org/10.1080/00223891.2011.645936

Magraw-Mickelson, Z., Carrillo, A. H., Weerabangsa, M. M., Owuamalam, C. & Gollwitzer, M. (2020). Comparing classic and novel approaches to measurement invariance. Preprint. https://doi.org/10.31234/osf.io/pz8u9

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2018). What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods*, *23*(3), 524–545. https://doi.org/10.1037/met0000113

McDonald, R. P. (1999). Test homogeneity, reliability, and generalizability. In R. P. McDonald (Ed.), *Test theory: A unified treatment* (pp. 76–120). Erlbaum.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, *63*, 1–11. https://doi.org/10.1016/j.jrp.2016.04.010

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, *96*(5), 966–980. https://doi.org/10.1037/a0022955

O'Brien, L. T., Major, B. N., & Gilbert, P. N. (2012). Gender differences in entitlement: The role of system-justifying beliefs. *Basic and Applied Social Psychology*, *34*(2), 136–145. https://doi.org/10.1080/01973533.2012.655630

Ock, J., McAbee, S. T., Mulfinger, E., & Oswald, F. L. (2020). The practical effects of measurement invariance: Gender invariance in two Big Five personality measures. *Assessment*, *27*(4), 657–674. https://doi.org/10.1177/1073191119885018

Osborne, D., Sengupta, N. K., & Sibley, C. G. (2019). System justification theory at 25: Evaluating a paradigm shift in psychology and looking towards the future. *British Journal of Social Psychology*, *58*(2), 340–361. https://doi.org/10.1111/bjso.12302

Owuamalam, C. K., Rubin, M., & Spears, R. (2018). A critical review of the (un)conscious basis for system-supporting attitudes of the disadvantaged. *Social and Personality Psychology Compass*, *12*(11), e12419. https://doi.org/10.1111/spc3.12419

Owuamalam, C. K., Rubin, M., & Spears, R. (2019). Revisiting 25 years of system motivation explanation for system justification from the perspective of social identity model of system attitudes. *British Journal of Social Psychology*, *58*(2), 362–381. https://doi.org/10.1111/bjso.12285

R Core Team (2019). *R: A language and environment for statistical computing (3.6.1) [Computer software]*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Revelle, W. (2020). *psych: Procedures for Personality and Psychological Research (2.1.3) [R package]*. Northwestern University Evanston. Retrieved from https://CRAN.R-project.org/package=psych

Richter, M., & König, C. J. (2017). Explaining individuals' justification of layoffs. *Journal of Applied Social Psychology*, *47*(6), 331–346. https://doi.org/10.1111/jasp.12442

Robitzsch, A. (2020). *sirt: Supplementary item response theory models (R package version 3.9-4) [Computer software]*. Retrieved from https://CRAN.R-project.org/package=sirt

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software*, *48*, 1–36.

Schaffer, B. S., & Riordan, C. M. (2003). A review of cross-cultural methodologies for organizational research: A best-practices approach. *Organizational Research Methods*, *6*(2), 169–215. https://doi.org/10.1177/1094428103251542

Schmitt, D. P., & Allik, J. (2005). Simultaneous administration of the Rosenberg Self-Esteem Scale in 53 nations: Exploring the universal and culture-specific features of global self-esteem. *Journal of Personality and Social Psychology*, *89*(4), 623–642. https://doi.org/10.1037/0022-3514.89.4.623

Schmitt, N., & Ali, A. A. (2014). The practical importance of measurement invariance. In C. E. Lance, & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 327–346). Routledge.

Stagner, R. (1956). *Psychology of industrial conflict*. Wiley.

Steenkamp, J.-B.-E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–90. https://doi.org/10.1086/209528

Ullrich, J., & Cohrs, J. C. (2007). Terrorism salience increases system justification: Experimental evidence. *Social Justice Research*, *20*(2), 117–139. https://doi.org/10.1007/s11211-007-0035-y

van der Toorn, J., Berkics, M., & Jost, J. T. (2010). System justification, satisfaction, and perceptions of fairness and typicality at work: A cross-system comparison involving the US and Hungary. *Social Justice Research*, *23*(2), 189–210. https://doi.org/10.1007/s11211-010-0116-1

van der Toorn, J., Feinberg, M., Jost, J. T., Kay, A. C., Tyler, T. R., Willer, R., & Wilmuth, C. (2015). A sense of powerlessness fosters system justification: Implications for the legitimation of authority, hierarchy, and government. *Political Psychology*, *36*(1), 93–110. https://doi.org/10.1111/pops.12183

van der Toorn, J., Tyler, T. R., & Jost, J. T. (2011). More than fair: Outcome dependence, system justification, and the perceived legitimacy of authority figures. *Journal of Experimental Social Psychology*, *47*(1), 127–138. https://doi.org/10.1016/j.jesp.2010.09.003

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Vargas-Salfate, S., Paez, D., Khan, S. S., Liu, J. H., & Gil de Zúñiga, H. (2018). System justification enhances well-being: A longitudinal analysis of the palliative function of system justification in 18 countries. *British Journal of Social Psychology*, *57*(3), 567–590. https://doi.org/10.1111/bjso.12254

Vesper, D., & König, C.J. (in preparation). Measurement equivalence of the English, German, and French version of the strike attitude scale.

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *dplyr: A grammar of data manipulation (R package version 1.0.5) [Computer software]*. Retrieved from https://CRAN.R-project.org/package=dplyr

Yentes, R. D. & Wilhelm, F. *careless: Procedures for computing indices of careless responding (R package version 1.1.3) [Computer software]*. Available at https://github.com/ryentes/careless

Zmigrod, L. (2020). The role of cognitive rigidity in political ideologies: Theory, evidence, and future directions. *Current Opinion in Behavioral Sciences*, *34*, 34–39. https://doi.org/10.1016/j.cobeha.2019.10.016

Zmigrod, L., Rentfrow, P. J., & Robbins, T. W. (2018). Cognitive underpinnings of nationalistic ideology in the context of Brexit. *Proceedings of the National Academy of Sciences*, *115*(19), E4532–E4540. https://doi.org/10.1073/pnas.1708960115

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Vesper, D., König, C. J., Siegel, R., & Friese, M. (2022). Is use of the general system justification scale across countries justified? Testing its measurement equivalence. *British Journal of Social Psychology*, 61, 1032–1049. https://doi.org/10.1111/bjso.12520