Examining the Structural Validity of Stereotype Content Scales – A Preregistered Re-Analysis of Published Data and Discussion of Possible Future Directions

INTERNATIONAL REVIEW OF SOCIAL PSYCHOLOGY

RESEARCH ARTICLE

MARIA-THERESE FRIEHS
PATRICK F. KOTZUR
JOHANNA BÖTTCHER
ANN-KRISTIN C. ZÖLLER
TABEA LÜTTMER
ULRICH WAGNER
FRANK ASBROCK
MAARTEN H. W. VAN ZALK

Author affiliations can be found in the back matter of this article

]u[ubiquity press

ABSTRACT

The Stereotype Content Model (SCM) plays a prominent role in social perception research when comparing the evaluation of different targets on warmth and competence dimensions. However, there is scarce information on the SCM's measurement properties. Thus, in this article, we provide a comprehensive test of the SCM's structural validity (i.e., reliability, dimensionality, cross-group comparability of measurement properties). We re-analysed published SCM data from English speaking participants (study 1: 78 datasets from 43 original publications, *N* = 20,819) and German participants (study 2: 29 datasets from 23 original publications, *N* = 10,854). We used confirmatory factor analyses to assess the scales' reliability and dimensionality as well as measurement invariance assessment to examine cross-group comparability as a precondition for meaningful and valid mean-value comparison. We found on average good reliabilities of the SCM scales. In contrast, about 35% of all 1093 examined SCM measurement models presented adequate scale dimensionality, and regarding the scales' cross-group comparability, we found (partial) scalar measurement invariance in about 11% of all cases. These findings indicate considerable validity concerns in published SCM research, as a meaningful and valid measurement of warmth and competence was not given in approximately two thirds of all cases, and mean-value comparisons were potentially biased due to lacking cross-group comparability for about eight out of nine cases. We propose future directions to improve the measurement quality and validity in SCM research and invite fellow researchers to constructively discuss these ideas.

CORRESPONDING AUTHOR:

Maria-Therese Friehs

Faculty of Psychology, FernUniversität in Hagen, Germany

maria-therese.friehs@fernuni-hagen.de

KEYWORDS:

Stereotype Content Model; Social Perception; Confirmatory Factor Analysis; Measurement Invariance; Re-Analysis

TO CITE THIS ARTICLE:

Friehs, M.-T., Kotzur, P. F., Böttcher, J., Zöller, A.-K. C., Lüttmer, T., Wagner, U., Asbrock, F., & van Zalk, M. H. W. (2022). Examining the Structural Validity of Stereotype Content Scales – A Preregistered Re-Analysis of Published Data and Discussion of Possible Future Directions. *International Review of Social Psychology,* 35(1): 1, 1–18. DOI: *https://doi.org/10.5334/irsp.613*

The Stereotype Content Model (SCM; Fiske et al., 2002) is one of the most prominent models of social perception (Abele et al., 2021). It proposes that social groups are evaluated using two fundamental dimensions: *warmth* (i.e., a group's perceived intention), which is predicted by low competition and threat; and *competence* (i.e., a group's capacity to act on their intentions), which is predicted by status (Fiske et al., 2007; Kervyn et al., 2015). These two dimensions align with past person and group perception research going back to the 1940s and are compatible with some of the dimensions found in other lines of social perception research (e.g., Abele & Wojciszke, 2007; Koch et al., 2016; Leach et al., 2007; Yzerbyt et al., 2005).

Conceptually, the SCM's main focus is the simultaneous assessment and comparison of multiple groups—hereafter called targets[1]—from a societal perspective (Abele et al., 2021). This fact notwithstanding, the SCM's record of research applications is very diverse (far exceeding the level of intergroup evaluation) and very extensive, impressively demonstrated by about 900 results in the Web of Science database on that topic as of November 2021. The SCM body of research is so comprehensive that a full review would exceed the scope of this article, but importantly, most applications so far have focused on the comparison of warmth and competence values between different targets or samples, and consequently, mean-value comparisons were the most frequently used statistical approach.

Various methodological and conceptual aspects of the SCM are currently debated, including the relative dominance and adequate number of the fundamental dimensions of social perception (e.g., Abele et al., 2021; Brambilla et al., 2021; Koch et al., 2016; Stanciu, 2015), the relevance of the target-rater-relation (Koch et al., 2020), or how to most effectively assess stereotype content (e.g., David et al., 2018; Halkias & Diamantopoulos, 2020; Kotzur et al., 2020). In the present paper, we will contribute to some of these debates by focusing on the SCM's measurement properties.

In psychological research, measurement instruments (i.e., scales) are typically developed following established procedures to make sure that, among other things, researchers measure what they intend to measure (i.e., construct validity) and that the scales present adequate measurement (i.e., psychometric) properties for the intended statistical analyses. These criteria are also referred to as *structural validity* (Flake et al., 2017; Flake & Fried, 2020), and aspects of interest include item performance (i.e., how strongly a single item relates to the overall scale), reliability (i.e., measurement consistency), dimensionality (i.e., the underlying scale structure), and cross-group comparability (i.e., whether the measurement properties perform similarly in different sub-samples or when assessing different targets). In the case of the SCM, some initial structural validity information was given in the scale development process (Fiske et al., 1999, 2002). Halkias and Diamantopoulos (2020) have reviewed this procedure and identified a number of critical issues, including the use of small homogenous samples, participants' potential fatigue due to the high number of items and targets, a lack of robust methodology for structural validity assessment, and a non-transparent item selection process. The authors concluded that the SCM's scale development process was 'highly problematic' (Halkias & Diamantopoulos, 2020: 719) because of its lack of convincing structural validity evidence for the used warmth and competence scales. Later research often built on these unsatisfactorily developed scales, while also flexibly amending new or excluding existing items, oftentimes without providing a rationale for these decisions (Halkias & Diamantopoulos, 2020). Additionally, reviewing the subsequent SCM literature reveals that the SCM scales' structural validity continued to lack methodologically rigorous examination and comprehensive reporting: while most studies have reported high reliabilities (e.g., Durante et al., 2013; Fiske et al., 2002), few publications have investigated the dimensionality or cross-group comparability of the SCM. To explore dimensionality, Fiske and colleagues (1999, 2002) conducted principal component analyses on the used item pools, which often revealed up to four more dimensions than theoretically expected. Subsequent publications frequently replicated this procedure (e.g., Asbrock, 2010). As an exploratory approach, principal component analysis empirically generates a certain item-scale pattern based on observed covariations. However, a stricter test of dimensionality would be achieved by confirmatory factor analysis (CFA), which evaluates theoretically pre-defined expectations about item-scale relationships against the empirical reality (Brown, 2015). We identified only nine published SCM studies that have reported CFA results (Grigoryan et al., 2020; Hackbart et al., 2020; Halkias & Diamantopoulos, 2020; Janssens et al., 2015; Kotzur et al., 2019, 2020; Stanciu, 2015; Stanciu et al., 2017; Vauclair et al., 2017). Even fewer have examined cross-group comparability, that is, ensuring that the measurement properties of the SCM scales do not differ between samples or evaluated targets. ***Table 1*** presents an overview of studies assessing the SCM's dimensionality or cross-group comparability.

In short, we conclude that the SCM scales have not been developed according to modern standards using widely available advanced statistical procedures (e.g., Bandalos, 2018; Brown, 2015). This resulted in a lack of knowledge about the SCM scales' structural validity (especially regarding dimensionality and cross-group comparability). Consequently, we cannot be certain whether the underlying constructs are indeed two-dimensional (one warmth factor, one competence factor; dimensionality) and whether the SCM scales measure warmth and competence comparably for all targets to validly compare mean-values (cross-

REFERENCE	MODELLED FACTORS	METHOD OF ANALYSIS	RESULTS CFA (ACCEPTED/TESTED)	RESULTS MI
Grigoryan et al. (2020)	warmth, competence, status, competition	MGCFA[1]	/	full configural and metric, partial scalar MI
Hackbart et al. (2020)	warmth, competence	CFA	1/1 target	/
Halkias & Diamantopoulos (2020)	warmth, competence	CFA and MGCFA	Study 6: 1/1 target Study 7: 1/1 target	Study 7: full scalar MI
Janssens et al. (2015)	warmth, competence	CFA[1]	Study 1: reasonable Study 2: acceptable after deleting an item	Study 1: no acceptable fit Study 2: partial MI[2]
Kotzur et al. (2019)	warmth, competence	CFA and alignment optimization[1]	10/16 targets	full metric and partial scalar MI
Kotzur et al. (2020)	warmth, competence	CFA and MGCFA across two conditions	Study 1: 1/6 targets Study 2: 5/18 targets after item exclusion Study 3: 4/13 before + 4/13 after item exclusion	Study 1: scalar MI Study 2: (partial) scalar MI for 4 targets Study 3: (partial) scalar MI for 6 targets
Stanciu (2015)	warmth and competence (two factor model) vs. trustworthiness, friendliness, efficacy, conscientiousness (four factor model)	CFA and MGCFA[1]	Two factor model: 1/25 Four factor model: 13/25	MI for two targets applying the four factor-model.
Stanciu et al. (2017)	warmth, competence	CFA[1]	Study 1: 2/2 Study 2: 22/22	/
Vauclair et al. (2016)	warmth, competence, four BIAS map behaviours	CFA[1]	1/1 targets in both samples	partial scalar MI

Table 1 An Overview of Confirmatory Factor Analysis and Measurement Invariance Examination in the Current SCM Literature.

Note. MGCFA = Multiple-group confirmatory factor analysis; MI = Measurement Invariance [1] The applied methods and/or model fit criteria deviated from the ones chosen in this paper. [2] MI level not specified.

group comparability). This lack of knowledge about the properties and qualities of SCM scales is highly problematic, because 'if a construct of interest is studied with poor measurement, the ability to make any claims about the phenomenon is severely curtailed because what exactly is being measured is unknown and that uncertainty trickles down to the primary results' (Flake et al., 2017: 370).

In the present manuscript, we address this research gap around the structural validity of SCM scales by systematically examining the measurement properties of existing SCM scales in two data re-analyses. In study 1, we focused on studies published in the English language and based on English-speaking samples, because English is the original language in which the SCM was proposed and because most published studies are conducted in English. However, recent meta-scientific discussions have pointed out both a general need for replication of (social) psychological results (e.g., Earp & Trafimow, 2015; Schimmack, 2020) and a need to consider context as an important influence on research findings (e.g., Pettigrew, 2018, 2021). Consequently, in study 2, we replicated our approach in another research context, namely the one we as authors know best: German SCM scales with German samples. In both studies, our analyses addressed the following research questions (RQ):

1.
    a. How reliable are published SCM scales?[2]
    b. To what extent do published SCM scales support the theoretically proposed dimensionality by assessing warmth and competence as two separate factors across different targets?
2. To what extent are SCM scales comparable in their measurement properties across different targets or samples, thus allowing for meaningful and valid mean-value comparisons of warmth and competence?[3]

Contingent on our results, we will deduce concrete implications for SCM researchers as well as potential steps to increase structural validity in future SCM research.

METHODS

The two studies relied on the same methodological approach, which we outline below. The few insignificant procedural differences between the studies are explained in OSM-1. Our analyses were preregistered on OSF (study 1: *https://osf.io/gqmvz/?view_only=9a8fc0053b634ace8ea8941b6c9423b7*; study 2: *https://osf.io/486h7/?view_only=e1b25da1084f4e248a621be36b31a153*).[4] The analysis code and detailed results for all analyses are stored on the Open Science Framework (see *https://osf.io/jqzet/*).

We re-analysed publications published until January 2020 (study 1)/Mid-August 2020 (study 2) that

(I) compared SCM dimensions between different targets or samples;
(II) directly referred to the SCM and not to a related model of social perception;
(III) measured warmth and competence by at least two items each, and
(IV) used English-language scales and/or samples (study 1)/German-language scales and German samples (study 2).

We scanned academic search engines (Web of Science, PsycInfo, PSYNDEX, Google Scholar, Fiske-Lab) for eligible datasets and sent out calls for data via professional mailing-lists (SPSSI, EASP, DGPs). We included all data in our re-analyses to which we gained access until the end of December 2020. The study identification process is summarised in the figures OSM-3 (study 1) and OSM-4 (study 2). Our level of analysis was datasets, which resulted in more re-analysed datasets than publications included in the re-analysis. The included publications are listed in OSM-6 (study 1) and OSM-9 (study 2). We emphatically thank all authors for sharing their data.

The modelling process is explained in more detail in OSM-5. Due to the high number of analysis steps per dataset, we used a four-eyes-principle according to which all analyses were carried out by one of the authors and independently examined by another.

First, to assess the warmth and competence scales' dimensionality and reliability (RQ1), we conducted CFA for each target using the SCM scale information described in the original publications. Thus, we specified one warmth factor predicting the mentioned warmth indicators and one correlated competence factor predicting the competence indicators. From the CFA models, we computed the reliability coefficient McDonald's ω_{total}[5] for warmth and competence for all successfully estimated CFA models (Hayes & Coutts, 2020). We accepted CFA models that completely fulfilled the model-fit criteria by Schermelleh-Engel and colleagues (2003). CFA models with non-acceptable model-fit did not support the claim that the used items form valid SCM scales and were therefore discarded from further analyses to address RQ2.

Second, we investigated the comparability of measurement properties across targets within each dataset using measurement invariance (MI) assessment. MI examines whether a construct measured across multiple targets is indeed comparable by introducing equality of model parameters in multiple-group confirmatory factor analysis (MGCFA; Byrne et al., 1989). For each dataset separately and in line with the research questions of the original publications, we applied MI assessment to all acceptable CFA models using increasingly restrictive, nested models (Vandenberg & Lance, 2000). We first tested all accepted CFA models for equal form (i.e., configural MI), that is, whether the number of factors and the factor-loading patterns are comparable across CFA models. If this model was acceptable (as indicated by the model-fit criteria mentioned above), we introduced equality restrictions for factor-loadings of identical indicators across CFA models (i.e., metric MI). Metric MI implies equal warmth and competence measurement units across targets and is a precondition for correlational/regression-based analysis, such as analysing warmth-competence correlations between different targets or predicting emotions and behavioural tendencies for different targets from warmth and competence. Metric MI was assumed if overall model-fit was acceptable, the χ^2-value did not increase significantly, and model-fit changes adhered to Chen's (2007) criteria. For acceptable metric MI models, we added equality restrictions to indicator-intercepts of identical indicators across CFA models (i.e., scalar MI). Scalar MI implies equal points-of-zero (i.e., equal item difficulty; Boer et al., 2018) of similar SCM indicators across targets. It thus forms the precondition for valid warmth and competence mean-value comparisons across targets. Scalar MI was assumed if overall model-fit was acceptable, the χ^2-value did not increase significantly, and model-fit changes again adhered to Chen's (2007) criteria.

If full metric or scalar MI was not achieved, we improved model-fit by introducing partial measurement invariance (Byrne et al., 1989). Partial MI allows exceptions from the equality constraints of measurement properties for some targets; thus, the equality-assumption is somewhat limited but still generally accepted (Davidov et al., 2014). We identified eligible equality constraints using modification indices and introduced partial MI by releasing constraints on the precondition that, for at least two indicators per factor, all parameters remained equal across all targets (Davidov et al., 2014). If introducing metric or scalar partial MI still resulted in unacceptable model-fit, we deleted the target with the highest χ^2-value contribution in the fully constrained MI model from analysis and repeated the complete process, always aiming to reach (partial) scalar MI as a requirement for RQ2.

RESULTS

STUDY 1: ENGLISH CONTEXT

Our final set of data consisted of 78 datasets from 43 publications (N = 20,819 participants; see OSM-3). Detailed information about each dataset (e.g., targets, scale wording, sample information) can be found in OSM-6. We first conducted CFA on each of the $K = 586$ targets to assess the reliability and dimensionality for the SCM scales (RQ1). Omega statistics across $K = 452$ models (excluding 134 models with implausible parameter estimates or which did not converge) revealed that warmth and competence scales were reliable on average ($M\omega_{\text{Warmth}}$ = .841, $SD\omega_{\text{Warmth}}$ = .088, min = .481, max = .977; $M\omega_{\text{Competence}}$ = .833, $SD\omega_{\text{Competence}}$ = .085, min = .411, max = .980; Raykov & Marcoulides, 2011). Regarding dimensionality, *K* = 204 models from 43 datasets showed satisfactory model-fit (34.81% of all targets), indicating that SCM scales demonstrated the theoretically expected dimensionality (one warmth factor, one competence factor) in about one third of all analysed cases. Summarised information on accepted CFA models are presented in ***Table 2***. Detailed tables with model-fit results for each target and dataset are provided in OSM-7.

To determine the SCM scales' comparability across targets, which is the prerequisite of valid mean-value comparison (RQ2), we assessed the datasets' MI up to (partial) scalar level. Fifteen datasets showed adequate CFA model-fit for only one target, which logically forbids MI testing; thus, 35.90% of all datasets with a total of *K* = 189 CFA models qualified for MI assessment. Summarised results are presented in ***Table 2*** and ***Figure 1***. Detailed tables including the model-fit parameters for the different levels of MI per dataset are provided in OSM-8.

Before scalar MI could be assessed, warmth and competence scales had to fulfil the criteria for configural and (partial) metric MI. Out of the 28 datasets that

DATASET	TOTAL # TARGETS	CFA		CONFIGURAL MI	METRIC MI	SCALAR MI		
		# ACCEPTABLE TARGETS	% OF TOTAL TARGETS	# TARGETS	# TARGETS (MI LEVEL)	# TARGETS (MI LEVEL)	% OF ACCEPTABLE TARGETS	% OF TOTAL TARGETS
1	3	0	00.0	/	/	/	/	/
2	20	9	45.0	9	Full (6)	Full (4)	44.4	20.0
3	12	1	8.3	/	/	/	/	/
4	4	1	25.0	/	/	/	/	/
5	12	3	25.0	3	/	/	/	/
6	2	2	100.0	2	Partial (2)	Partial (2)	100.0	100.0
7	4	1	25.0	/	/	/	/	/
8	4	0	00.0	/	/	/	/	/
9	5	0	00.0	/	/	/	/	/
10	4	0	00.0	/	/	/	/	/
11	8	0	00.0	/	/	/	/	/
12	8	0	00.0	/	/	/	/	/
13	16	0	00.0	/	/	/	/	/
14	3	3	100.0	3	Full (3)	Full (3)	100.00	100.0
15	5	4	80.0	4	Full (2)	Full (2)	50.00	40.00
16	4	0	00.0	/	/	/	/	/
17	30	20	66.7	20	Full (8)	Full (2)	10.0	6.7
18	30	23	76.7	23	Full (10)	Full (2)	8.6	6.7
19	22	14	63.6	14	Full (6)	/	/	/
20	2	0	00.0	/	/	/	/	/
21	2	0	00.0	/	/	/	/	/
22	2	0	00.0	/	/	/	/	/
23	2	1	50.0	/	/	/	/	/
24	61	13	21.3	13	Partial (13)	Partial (8)	61.5	21.3
25	10	10	100.0	10	Full (10)	Full (10)	100.0	100.0
26	20	9	45.0	Full	Full (2)	/	/	/
27	23	11	47.8	Full	Full (7)	/	/	/

(Contd.)

DATASET	TOTAL # TARGETS	CFA		CONFIGURAL MI	METRIC MI	SCALAR MI		
		# ACCEPTABLE TARGETS	% OF TOTAL TARGETS	# TARGETS	# TARGETS (MI LEVEL)	# TARGETS (MI LEVEL)	% OF ACCEPTABLE TARGETS	% OF TOTAL TARGETS
28	4	0	00.0	/	/	/	/	/
29	2	2	100.0	2	Full (2)	Full (2)	100.0	100.0
30	2	1	50.0	/	/	/	/	/
31	10	1	10.0	/	/	/	/	/
32	10	0	00.0	/	/	/	/	/
33	2	0	00.0	/	/	/	/	/
34	4	1	25.0	/	/	/	/	/
35	16	3	18.7	3	/	/	/	/
36	6	3	50.0	3	Partial (3)	Partial (3)	100.0	50.0
37	8	2	25.0	2	Full (2)	Full (2)	100.0	25.0
38	8	1	12.5	/	/	/	/	/
39	2	0	00.0	/	/	/	/	/
40	2	0	00.0	/	/	/	/	/
41	2	0	00.0	/	/	/	/	/
42	2	0	00.0	/	/	/	/	/
43	3	0	00.0	/	/	/	/	/
44	6	0	00.0	/	/	/	/	/
45	4	0	00.0	/	/	/	/	/
46	6	1	16.7	/	/	/	/	/
47	2	0	00.0	/	/	/	/	/
48	10	8	80.0	8	Partial (8)	Partial (2)	25.0	20.0
49	3	3	100.0	3	Full (2)	Partial (2)	66.7	66.7
50	6	3	50.0	3	Partial (2)	/	/	/
51	2	0	00.0	/	/	/	/	/
52	4	0	00.0	/	/	/	/	/
53	2	2	100.0	2	Full (2)	/	/	/
54	2	0	00.0	/	/	/	/	/
55	4	0	00.0	/	/	/	/	/
56	2	0	00.0	/	/	/	/	/
57	2	0	00.0	/	/	/	/	/
58	2	2	100.0	2	Full (2)	Partial (2)	100.0	100.0
59	2	1	50.0	/	/	/	/	/
60	2	1	50.0	/	/	/	/	/
61	2	0	00.0	/	/	/	/	/
62	35	8	22.8	8	Partial (6)	Partial (2)	25.0	5.7
63	4	1	25.0	/	/	/	/	/
64	2	0	00.0	/	/	/	/	/
65	2	0	00.0	/	/	/	/	/
66	25	15	60.0	15	Partial (13)	Partial (4)	26.7	16.0

(Contd.)

DATASET	TOTAL # TARGETS	CFA		CONFIGURAL MI	METRIC MI	SCALAR MI		
		# ACCEPTABLE TARGETS	% OF TOTAL TARGETS	# TARGETS	# TARGETS (MI LEVEL)	# TARGETS (MI LEVEL)	% OF ACCEPTABLE TARGETS	% OF TOTAL TARGETS
67	3	1	33.3	/	/	/	/	/
68	3	0	00.0	/	/	/	/	/
69	4	2	50.0	2	Full (2)	Partial (2)	100.0	100.0
70	5	0	00.0	/	/	/	/	/
71	3	0	00.0	/	/	/	/	/
72	6	4	66.7	4	Partial (4)	Full (2)	50.0	33.3
73	6	1	16.7	/	/	/	/	/
74	6	0	00.0	/	/	/	/	/
75	6	1	16.7	/	/	/	/	/
76	4	4	50.0	4	Full (4)	Partial (4)	100.0	100.0
77	4	3	75.0	3	Full (3)	Partial (3)	100.0	75.0
78	4	4	100.0	4	Full (4)	Partial (4)	100.0	100.0
Total	**586**	**204**	**34.8**	**189**	**77 Full / 51 Partial**	**29 Full / 38 Partial**	**69.9**	**56.5**

Table 2 Confirmatory Factor Analysis and Measurement Invariance Results for all Datasets in Study 1.

Note: CFA = Confirmatory Factor Analysis. / = The testing of this level of MI was not possible due to the number of target groups with acceptable model fit being below 2. [1] Of accepted CFA models (column Total CFA Fit). The dataset number refers to the numbering system in OSM-6; for full references, see OSM-6.

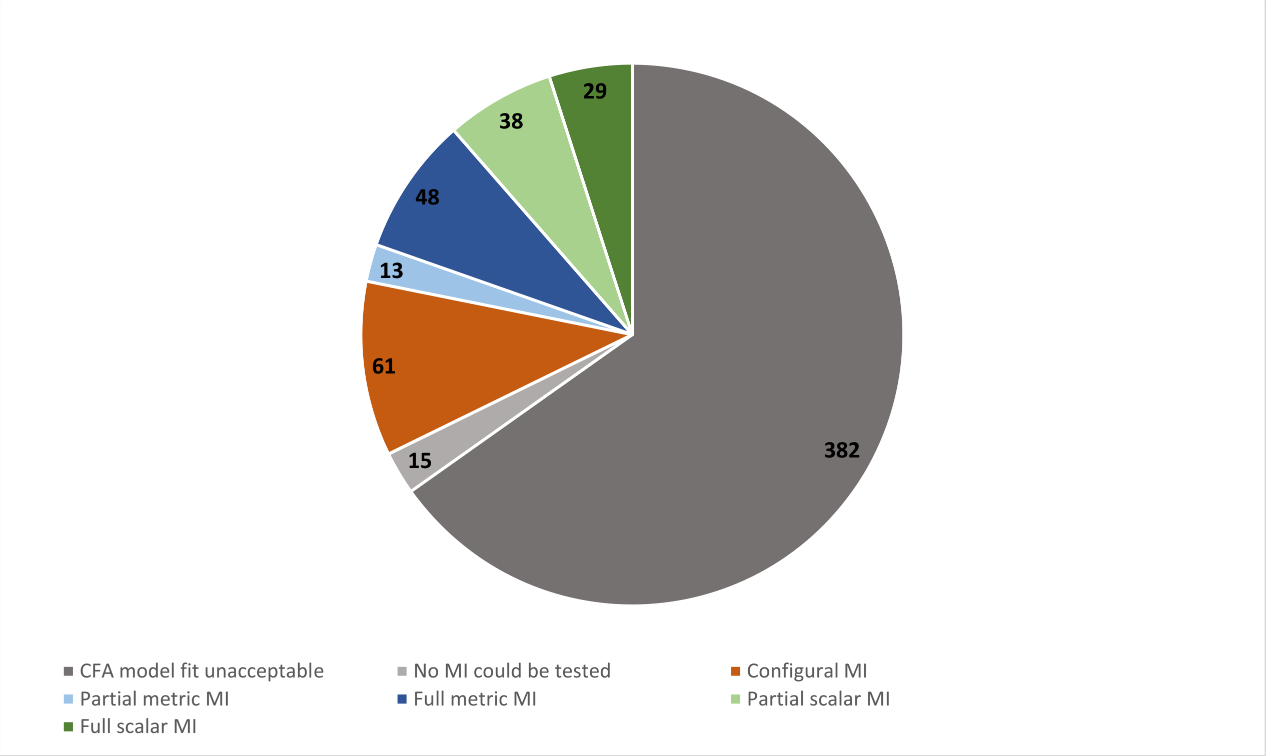

Figure 1 Highest Level of Established Measurement Invariance Based on CFA Models in Study 1.

Note: The figure shows the highest level of measurement equivalence in which the target dropped out from analyses, thus the numbers may vary compared to the descriptions in the text, which describes the results on the level of datasets. MI = Measurement invariance. The total number of CFA models K = 586.

qualified for MI testing, all held up to configural MI. In the next step, we constrained factor-loadings of identical items to be equal across all targets within each dataset to test metric MI. The full metric model showed acceptable fit in 18 datasets, the partial metric model in further eight datasets. Thus, a total of 26 datasets (33.33% of all datasets) held up to the standards of (partial) metric MI, allowing for comparative correlational analyses. Two datasets had to be excluded from further analyses because we could not establish (partial) metric MI for at least two targets. Finally, we tested (partial) scalar MI by constraining identical indicator-intercepts to be equal across targets within each dataset. Nine datasets held up to full scalar MI, a further 12 to partial scalar MI. This means that out of the 78 re-analysed datasets, a total of 21 datasets (26.92%), including *K* = 67 targets (11.43% of all targets), showed cross-group comparability in the form of (partial) scalar MI and thus allowed for meaningful and valid mean-value comparison between targets. Of those, three datasets achieved full scalar MI for all targets examined in the original publication, which means that in the remaining 18 datasets either parameters had to be freed (introducing partial MI) or targets had to be excluded.

STUDY 2: GERMAN CONTEXT

Our final set of data included 29 datasets from 23 publications (N = 10,854 participants; see OSM-4). Detailed information about each dataset can be found in OSM-9. We applied CFA to the SCM scales of *K* = 507 targets to determine the scales' reliability and to confirm the dimensionality of the SCM measurement models as proposed in the original publications (RQ1). Scale reliability across *K* = 497 CFA models (excluding 10 models with implausible parameter estimates or which did not converge) was good ($M\omega_{\text{Warmth}}$ = .849, $SD\omega_{\text{Warmth}}$ = .068, min = .553, max = .969; $M\omega_{\text{Competence}}$ = .809, $SD\omega_{\text{Competence}}$ = .078, min = .474, max = .969; Raykov & Marcoulides, 2011). Regarding dimensionality, *K* = 178 CFA models from 20 original publications achieved acceptable model-fit, indicating that we found evidence for the theoretically expected warmth and competence scales for 35.10% of all CFA models. Summary information about the SCM scales' dimensionality per dataset is presented in ***Table 3***. Detailed tables with model-fit results for each target and dataset are provided in OSM-10.

To determine the SCM scales' comparability across targets (RQ2), we inspected MI up to (partial) scalar level for *K* = 160 targets from 17 datasets (excluding *K* = 18 CFA models because only 1 target per dataset showed acceptable model-fit). Summarised results are presented in ***Table 3*** and ***Figure 2***. Detailed tables including the model-fit parameters for the different MI levels per dataset are provided in OSM-11.

All 17 datasets held up to configural MI. When constraining factor-loadings of identical items to be equal across targets to test for metric MI, the full metric MI model showed acceptable fit in nine datasets, the partial metric model in a further seven datasets. Thus, 55.17% of all datasets held up to the standards of (partial) metric MI, allowing for comparative correlational analyses. Two datasets had to be excluded from further analyses because we could not establish (partial) metric MI for at least two targets. When testing (partial) scalar MI by constraining identical indicator-intercepts to be equal across targets, seven datasets held up to full scalar MI, a further eight to partial scalar MI. To summarise, out of the 29 re-analysed datasets, 48.27% of all datasets, including *K* = 58 targets (11.44% of all targets), held up to the criteria of (partial) scalar MI and thus allowed for meaningful and valid mean-value comparison between targets. Of those, two datasets achieved full scalar MI for all targets examined in the original publication; in the case of all other datasets, either parameters had to be freed (introducing partial MI) or targets had to be excluded.

We also conducted extensive item and measurement performance analyses and some exploratory analysis regarding the relation of sample-size and model-fit, which we report in detail in OSM-12 for both studies.

DISCUSSION

The SCM (Fiske et al., 2002) proposes two fundamental dimensions of social perception of groups: warmth and competence. This comprehensive theoretical framework has stimulated important research in many different contexts and has been applied to various research questions. We have contributed to this body of research and its ongoing methodological and conceptual debates (Abele et al., 2021, 2021; David et al., 2018; Halkias & Diamantopoulos, 2020; Koch et al., 2020, 2021; Kotzur et al., 2020) by presenting systematic examinations of the structural validity of published SCM scales. Besides the often-reported scale reliability, we focused especially on the aspects of dimensionality and cross-group comparability, which were rarely examined in previous publications (see ***Table 1***), but which are essential for the meaningful and valid interpretation of findings. In two studies investigating the published literature from two different contexts (study 1: English SCM scales and English-speaking samples; study 2: German SCM scales and German samples), we re-analysed a total of 107 SCM datasets from 66 publications (N_{total} = 31,673) using (multiple-group) confirmatory factor analysis. Both studies showed remarkably consistent results, which can be summarised as follows: (I) the warmth and competence scales showed on average good reliability; (II) in contrast, only about 35% of the 1,093 analysed targets demonstrated acceptable CFA model-fit of the SCM scales. Thus, in approximately 65% of all cases, the

DATASET	TOTAL # TARGETS	CFA		CONFIGURAL MI	METRIC MI	SCALAR MI		
		# ACCEPTABLE TARGETS	% OF TOTAL TARGETS	# TARGETS	# TARGETS (MI LEVEL)	# TARGETS (MI LEVEL)	% OF ACCEPTABLE TARGETS	% OF TOTAL TARGETS
1	29	7	24.14	7	7 (Partial)	4 (Partial)	57.14	13.79
2	15	5	33.33	5	4 (Partial)	3 (Partial)	60.00	20.00
3	2	0	00.00	/	/	/	/	/
4	16	7	43.75	7	7 (Partial)	5 (Partial)	71.43	31.25
5	4	1	25.00	/	/	/	/	/
6	4	1	25.00	/	/	/	/	/
7	4	2	50.00	2	2 (Full)	0	00.00	00.00
8	6	2	33.33	2	0	0	00.00	00.00
9	6	3	50.00	3	0	0	00.00	00.00
10	3	1	33.33	/	/	/	/	/
11	4	1	25.00	/	/	/	/	/
12	32	19	59.38	19	14 (Partial)	2 (Full)	10.53	6.25
13	2	1	50.00	/	/	/	/	/
14	3	0	00.00	/	/	/	/	/
15	2	1	50.00	/	/	/	/	/
16	4	1	25.00	/	/	/	/	/
17	16	10	62.50	10	10 (Full)	8 (Partial)	80.00	50.00
18	12	4	33.33	2	2 (Full)	2 (Full)	50.00	16.67
19	18	12	66.67	8	6 (2 Full/ 4 Partial)	4 (2 Full/ 2 Partial)	33.33	22.22
20	20	15	75.00	12	2 (Partial)	2 (Full)	13.33	10.00
21	2	2	100.00	2	2 (Full)	2 (Full)	100.00	100.00
22	204	50	24.51	50	8 (Full)	5 (Full)	10.00	2.45
23	4	4	100.00	4	4 (Full)	4 (Full)	100.00	100.00
24	4	1	25.00	/	/	/	/	/
25	64	22	34.38	22	12 (Partial)	12 (Partial)	54.55	18.75
26	16	3	18.75	3	3 (Full)	3 (Partial)	100.00	18.75
27	5	1	20.00	/	/	/	/	/
28	3	0	00.00	/	/	/	/	/
29	3	2	66.67	2	2 (Full)	2 (Partial)	100.00	66.67
Total	**507**	**178**	**35.10**	**160**	**85**	**58**	**32.58**	**11.44**

Table 3 Confirmatory Factor Analysis and Measurement Invariance Results for all Datasets in Study 2.

Note: CFA = Confirmatory Factor Analysis, MI = Measurement Invariance. # = Number. / = The testing of this level of MI was not possible due to the number of target groups with acceptable model fit being below 2. Note that for Samples 18–20, measurement equivalence was tested for each target group separately across experimental conditions if prerequisites were met (see OSM-9 for details). The dataset number refers to the numbering system in OSM-9; for full references, see OSM-9.

theoretically assumed dimensionality of warmth and competence as two distinct factors was not supported. (III) About 11.40% of the 1,093 analysed targets presented (partial) scalar measurement invariance as an indication of cross-group comparability, which is essential for meaningful mean-value comparisons. Our findings indicate considerable structural validity concerns, especially regarding dimensionality and cross-group comparability, in existing SCM research. In the following section, we discuss our findings in more detail

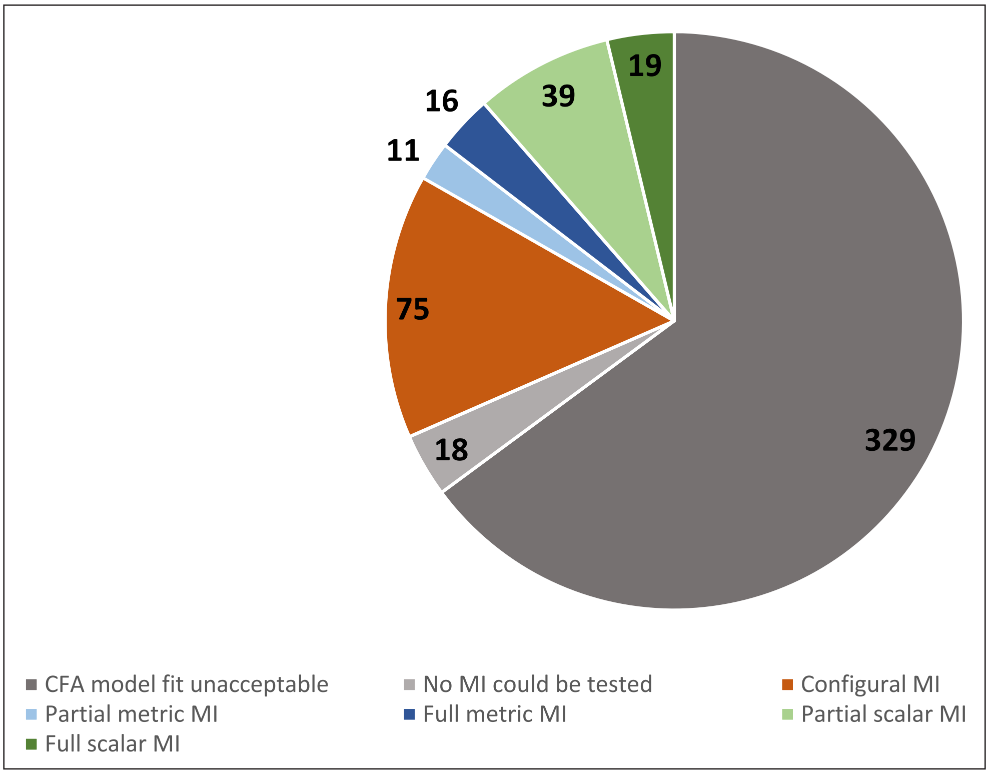

Figure 2 Highest Level of Established Measurement Invariance Based on CFA Models in Study 2.

Note: The figure shows the highest level of measurement equivalence in which the target dropped out from analyses, thus the numbers may vary compared to the descriptions in the text, which describes the results on the level of datasets. MI = Measurement invariance. The total number of CFA models K = 507.

and provide recommendations for future research.

RELIABILITY AND DIMENSIONALITY OF SCM SCALES

To address research question one, we analysed whether warmth and competence scales reliably and validly measured stereotype content as two distinct dimensions of social perception. Favourable dimensionality evidence was a prerequisite for the subsequent analyses that compared the measurement properties of warmth and competence scales between different targets. CFA results revealed that the scales showed on average good reliability, while at the same time showing they did not support the assumed scale dimensionality in about 65% of all cases. This lack of dimensionality implies that the items used to assess warmth and competence did not actually form valid scales on which we could meaningfully compare the stereotype content of different targets (Brown, 2015). Although these findings appear counter-intuitive, scales with acceptable reliability but unacceptable dimensionality have also been reported elsewhere (Hester et al., 2022) and are explicable because reliability and dimensionality are distinct features of scale performance (for more information, see Davenport et al., 2015; Green & Yang, 2015).

Indications regarding the issue of dimensionality have been reported sporadically in some previous SCM studies (e.g., Janssens et al., 2015; Kotzur et al., 2019, 2020; Stanciu, 2015). Moreover, our findings align with the recent critique of the SCM's initial scale development (Halkias & Diamantopoulos, 2020) and give empirical evidence to the critique of scale development and usage in the field of (social) psychology (Flake & Fried, 2020). As a consequence, one could have expected the existence of some unacceptable CFA models, but nonetheless, the extent of the issue is astounding. We therefore surmise that there exists a substantial gap between the well-founded theoretical framework of the SCM and appropriate operationalisations of the two dimensions of social perception, which calls for more careful scale construction efforts in the future (for an example, see Halkias & Diamantopoulos, 2020).

Our findings can also be interpreted in light of the ongoing debate regarding the number and meaning of the dimensions of social perception (e.g., Abele et al., 2021; Brambilla et al., 2021; Kervyn et al., 2013; Koch et al., 2016, 2020, 2021; Leach et al., 2007; Stanciu, 2015). Our results indicate that most of the SCM scales used cannot be employed without further analysis to assess and compare warmth and competence perceptions validly. This does not mean that other theoretical models of social perception should be preferred, as we do not know the extent to which these related models show adequate structural validity. We recommend taking this aspect into consideration for future applications of these theoretical models: the requirements of structural validity are not singular to the SCM but apply to all research applying scale-based measurement instruments.

CROSS-GROUP COMPARABILITY OF SCM SCALES

Research question two focused on the comparability of the SCM scales' measurement properties (i.e., measurement invariance) across targets. Evidence supporting this aspect is necessary to ensure that the underlying warmth and competence constructs are defined equally when comparing targets on SCM scales, and thus to ensure a valid and meaningful interpretation of mean differences. We subjected the targets that showed acceptable CFA model-fit to MI analysis based on MGCFA up to (partial) scalar level to fulfil the statistical requirements of unbiased mean-value comparisons. Our results indicated that meaningful mean-value comparison along the SCM dimensions was possible for only about 11.40% of all targets. The absence of (partial) scalar measurement invariance for most targets indicates that, usually, mean-value comparisons on SCM dimensions result in the figurative comparison of apples with oranges (Chen, 2008) because the targets' warmth and competence concepts are non-equivalent and their measurement properties non-comparable. This compromises a meaningful and valid interpretation of research findings (Boer et al., 2018; Flake et al., 2017; Hussey & Hughes, 2020).

In line with our results, other SCM studies (e.g., Janssens et al., 2015; Kotzur et al., 2020; Stanciu et al., 2017) and other scales in social and personality psychology (Hussey & Hughes, 2020) have also reported measurement non-invariance. For instance, Hussey and Hughes (2020) investigated the structural validity of 15 established scales in social and personality psychology (not including the SCM) and found only mixed or poor CFA results in 76% of cases, as well as poor MI results in 48% of cases. Though their methodological approach was different to ours and not without critique (Wetzel & Roberts, 2020), the results mirror our findings.

We cannot explain why we found such an extensive lack of cross-group comparability. Measurement non-invariance on scalar level may be caused by varying social desirability/social norm influences (i.e., method bias) or varying propensities to respond to specific items that do not represent 'true' differences in the underlying construct (i.e., item bias; Boer et al., 2018; Chen, 2008). We cannot theoretically argue why certain target assessments, compared to others, should be subject to these influences. But we might hypothesise that such response biases, if they occurred systematically, emerge when participants find some targets more difficult to evaluate on SCM dimensions than others.

Lastly, some SCM studies might focus on comparative correlational analyses of warmth and competence with other variables, although such research questions are less frequent and thus not the main focus of our study. For such comparative correlational studies, establishing metric MI is an equally relevant precondition for drawing meaningful conclusions as scalar MI is for valid mean-value comparisons. (Partial) metric MI was given more often than scalar MI, but still, it was more often absent than present. This was mainly due to lack of CFA fit. Metric non-invariance might indicate item bias or method bias in the measurement, for instance based on varying stimulus familiarity (Boer et al., 2018). Again, we cannot find any theoretical reasoning that would lead us to expect such biases in SCM research.

We do not mean to affront any member of the SCM research community with the interpretations, and our results should not be interpreted as a general claim that all SCM research is biased or invalid. We do not wish to devaluate the efforts of many researchers, nor to depict the field as 'inept and misguided' (Fiske, 2017: 653). Indeed, the research we authored suffers from the same structural validity concerns (Asbrock, 2010; Asbrock et al., 2011; Kotzur et al., 2017, 2019, 2020; Meyer & Asbrock, 2018). With this study, we wish to draw researchers' attention to the importance of structural validity in SCM research and to start a lively, productive and constructive discussion about how the SCM scales could be improved, to eventually advance the research on social perception by taking issues of structural validity into account. To initiate such a discourse, we present some suggestions with the aim of ensuring highly structurally valid future SCM research.

POSSIBLE FUTURE DIRECTIONS

Structural validity assessment as standard. It is erroneous to assume that the structural validity of SCM scales is a given in the absence of sufficient empirical tests. Therefore, we propose that reporting results of CFA and MI examinations (if applicable) becomes a standard for future SCM and related research in an acknowledgement of current psychometric standards and transparent science. Up until now, common practice has included only the report of reliability coefficients or the results of principal component or exploratory factor analyses (Halkias & Diamantopoulos, 2020; see also Hester et al. 2022), while CFA and MI assessment have rather been exceptions (see ***Table 1***). Future SCM research testing MI might consider using a top-down-approach (e.g., Horn & McArdle, 1992), which starts with the assumption of full scalar measurement invariance and relaxes equality constraints until overall model-fit is achieved. This procedure, compared to the bottom-up approach we chose in this manuscript, might reduce the effort required to run MI analyses.

Increased sample size. CFA-based analyses require larger sample sizes than many re-analysed datasets presented. Small sample sizes may affect the general computation and convergence of structural equation models, the model-fit, as well as the statistical power and significance of factor-loadings and structural relations between latent variables (Kenny et al., 2015; Wolf et al., 2013). Thus, in line with calls in the general field of psychology, we recommend the usage of larger samples and the a priori planning of sample-size. There exist a number of rather straightforward approaches to determine power in structural equation models (for further information, see Brown, 2015; Muthén & Muthén, 2002; Wang & Rhemtulla, 2021), which we encourage SCM researchers to use.

SCM measurement. We also call for changes in the measurement of warmth and competence. Previous SCM research did not rely on one SCM scale but rather on a variety of context- or language-specific instruments. We saw issues in CFA and MI testing in nearly all scales in our analyses, many of which relied on the items used in the initial SCM publications (Fiske et al., 1999, 2002). Therefore, we believe that further construction and improvement efforts on SCM scales are required. Standardised scales would contribute to cumulative science projects (such as Durante et al., 2013, 2017) and hold great value for researchers who work on a smaller scope and would thus struggle with validating their own scales. In OSM-12, we provide item performance information for both contexts, which might be helpful to select well-functioning warmth and competence items. Moreover, Halkias and Diamantopoulos (2020) recently presented a diligent scale construction project for a German SCM measurement instrument. This could serve as a start for eventually developing validated SCM scales in multiple languages that hold up to the criteria of structural validity.

When constructing such scales, we recommend using more than three items for warmth and competence, because the more information provided in the measurement model, the more analysis options are available (e.g., a larger extent of partial measurement invariance). More items would also allow for more ad hoc model adjustments (e. g., by deleting indicators from the scale to increase CFA model-fit or MI, as in Kotzur et al., 2020) and usually result in higher reliabilities. We are aware that this recommendation has practical drawbacks, as SCM studies usually collect information about many targets at the same time. Increasing the number of items would naturally increase potential participant fatigue, which was criticised in the initial SCM scale development (Halkias & Diamantopoulos, 2020). Thus, balancing the number of items and targets in a study is essential, and one option might be to apply sample splits so that participants rate only a subset of targets (e.g., Fiske et al., 2002; Kotzur et al., 2019).

On a related note, scale development endeavours might incorporate recent findings, which propose the existence of subdimensions or alternative factor structures in the SCM and other models of social perception (Abele et al., 2016; Brambilla et al., 2021; Koch et al., 2016; Leach et al., 2007; Sayans-Jiménez et al., 2017; Stanciu, 2015). Scale development efforts might aim at differentiating SCM scales from those of other social evaluation models proposed in the literature and exploring sub-dimensions of warmth and competence. Using broader (i.e., including more indicators) or more specified (i.e., identifying sub-dimensions) scales for warmth and competence might also hold the advantage that deviations in measurement models might be indicative of different conceptualisations of the constructs, and therefore potentially qualitative differences in warmth and competence perceptions between targets, which would be very informative from a theoretical perspective (for a cross-cultural perspective, see e.g., Boehnke et al., 2014).

Exploration of structural non-validity findings. Knowing which target failed to produce acceptable model-fit or showed measurement non-invariance could be put to practical use. Future research could search for systematic patterns or explanatory variables for non-fit or non-invariance, for example by using complementary approaches, such as cognitive interviewing or online probing (Meitinger et al., 2020). If such patterns existed, they could be indicative of differential processes of social perception that might have been overlooked with the current methods and in the theoretical debates. To explain why some targets might differ in social perception, findings from a methodological, measurement-theoretical level could thus be related directly to qualitative research contents.

New analytical approaches. Future works might also broaden SCM research by focusing more strongly on the application of a broader range of methods (e.g., latent forms of analysis that ensure reliability-corrected estimation, confirmatory hypothesis testing and MI evaluation; Brown, 2015). Therefore, we believe it is worthwhile to apply structural equation modelling as an alternative to regression analysis or latent profile analysis instead of cluster analysis. Also, alternatives to the MGCFA-based MI assessment and latent mean-value comparison, such as alignment optimisation, might be applied (Asparouhov & Muthén, 2014; Kotzur et al., 2019). Other works that investigated the dimensionality of social perception employed data-driven approaches, such as multi-dimensional scaling (Koch et al., 2016) or network-analytical approaches (Grigoryev et al., 2019) instead of theory-driven approaches and therefore provide a different perspective on the data. Moreover, recent research has stressed the importance of focusing on the variation in SCM scales instead of the mean-value (Koch et al., 2020), which involves still other analytical approaches.

CRITICAL REFLECTIONS

Our research presents a number of strengths, including the broad extent of our re-analysis.

In two studies, we included data from 107 datasets and 66 studies with more than 31,000 participants.

We believe that this extent qualifies for a careful generalisation to the entire pool of SCM scales, based on the return rate of our data access requests (study 1: 36.1%, study 2: 71.9% of all eligible studies) and the careful selection of our studies. In study 1, we chose a language-specific context to ensure that translation issues did not confound our results (Sechrest et al., 1972). We acquired data from various countries (e.g., Australia, Canada, Great Britain and Northern Ireland, India, New Zealand, Pakistan, United States), which represent different societal contexts. This stimulated the somewhat narrower approach in study 2, which focusses exclusively on German-language studies conducted in Germany. It is also worth mentioning that the practical applications in our data far exceed the theoretically proposed application of the SCM as a model to evaluate a number of social groups: our data contain ratings of in- and outgroups, well-known and unfamiliar individuals, faces, names, companies, geometrical forms, animals and many other types of stimuli. Thus, our datasets include a huge variety of targets, sample characteristics and sizes, data collection modes, and measurement models, and our results are robust and surprisingly similar across different countries and languages. Nonetheless, we acknowledge that unknown features of the studies at hand might have affected the generalisability of our results and that, due to the broad research landscape, studies might have escaped our notice despite the extensive literature reviews we conducted.

Moreover, both the general idea of our research and its practical execution follow the recommendations of and contribute to transparent, open and reproducible science: data re-analysis is a rarely used but powerful scientific tool that allows independent replication of previous findings and the generation of new cumulative results in an economical manner, thus increasing the transparency and accountability of reported research results (Davey & Hargreaves, 2015). By pre-registering our research questions and analytical approach, providing open code and extensive online supplementary materials, as well as using standardised procedures and a diligently implemented four-eyes-principle in all analyses, we strengthened the transparency and reproducibility of our analysis as well as its accessibility for other researchers.

Nonetheless, we also wish to draw some critical attention to the methodology we chose: MGCFA, which is frequently applied for testing MI (Davidov et al., 2014), is a procedure requiring numerous individual decisions (e.g., which parameters to free when establishing PMI), potentially leading to non-reproducible MI solutions (Sass, 2011). We counteracted this by introducing transparent and pre-registered analytical procedures and applying a four-eyes-principle in all analyses. Moreover, the model-fit criteria and cut-off criteria we chose directly affected our results, because unlike in significance testing, there are various proposed indices and cut-off criteria with individual strengths and weaknesses, and which could be selected more conservatively or more liberally (Davidov et al., 2014; Sass, 2011). What is more, the usage of MGCFA for MI examination is not without critique: we chose this approach because it has been the first and most commonly used approach to testing MI (Davidov et al., 2014), it has the largest methodology and applied literature base to draw upon, and MGCFA solutions can be transferred with relative ease to further analyses, such as structural equation modelling, which is practical for applied researchers. Nonetheless, it was also argued that the criteria and methods used in MGCFA for testing MI may be too strict in case of smaller deviations from the assumption of comparability (Asparouhov & Muthén, 2014), and several new and more liberal approaches have been proposed (e.g., approximate measurement invariance, exploratory structural equation modelling or alignment optimization; Brown, 2015). We cannot say how the usage of these recent alternatives would have impacted our results.

On a related note, to avoid confusion, it is worth noting that most methodological literature defines (partial) scalar MI as a precondition for *latent* (i.e., measurement-error-corrected) mean-value comparison (vis-à-vis comparisons of observed mean- or sum-scores; e.g., Brown, 2015; Davidov et al., 2014; Vandenberg & Lance, 2000). Nonetheless, in our work, we did not introduce this distinction between latent and observed scores. The reason is that when computing observed mean- or sum-scores, researchers implicitly introduce equality assumptions to the data that are similar to those of measurement invariance (e.g., in a mean-score, all scale indicators usually have the same weight across targets, which is logically equivalent to introducing equal factor-loadings across targets, i.e., metric MI; differences in indicator-intercepts between targets are not examined in observed scores, resulting in assumed equivalence assumptions between targets as found in scalar MI; Meredith & Millsap, 1992; Sörbom, 1978). Thus, we believe that issues of structural validity, and especially cross-group comparability, should be considered even for elementary statistical analysis such as observed mean- or sum-score analysis.

Lastly, we did not differentiate between within-sample, between-sample, or mixed comparisons in our analysis. In some cases, the data structure implies a repeated measurement of SCM dimensions of different targets within the same sample. Thus, multi-level or longitudinal measurement invariance testing (e.g., Kotzur et al., 2020), and not MGCFA, would have been a more suitable approach (Brown, 2015; Vandenberg & Lance, 2000). However, such analytical approaches would not have reported separate χ^2-values for the included targets, which would have rendered our strategy of excluding targets from analysis impossible. Moreover, these analyses would require all targets to be included in one analytical model, which substantially increases the number of estimated parameters, and thus sample size requirements (Brown, 2015). Few of the datasets we analysed presented the necessary sample size for this approach, which is why we chose MGCFA instead. Methodologically, this implies that we based our analyses only on a limited part of the observed variance-covariance-matrix by treating dependent data as independent. This approach potentially biases the measurement invariance assessment by increasing the χ^2-value and reducing the estimated standard errors (B. Muthén, personal communication, March 6, 2020). But given the fact that a high χ^2-value on its own was no criterion in our analysis, and that standard errors were not considered at all, we feel this bias is unfortunate but tolerable.

CONCLUSION

Despite these limitations, we are convinced of the relevance and critical impact of our findings on SCM research. Our results cast doubt on the valid and meaningful usability of current SCM scales (Flake et al., 2017; Flake & Fried, 2020; Hester et al., 2022; Hussey & Hughes, 2020) and thus uncover previously hidden structural validity concerns. Such critical analyses of established theories and published data contribute to the general idea of open and reproducible science and stimulate controversial discussions and, thereby, innovative scientific perspectives. In line with Ellemers (2021), we hope that our work will animate respectful, animated, and fruitful discourses striving to collaboratively and constructively revise and improve research on the Stereotype Content Model and the fundamental dimensions of social perception in general.

PREREGISTRATION

English context: *https://osf.io/gqmvz/?view_only=26cfcec4f651454e9b508f2bdc917a96*.

German context: *https://osf.io/486h7/?view_only=870dbff75b004b29aeffae7d27c62518*.

DATA ACCESSIBILITY STATEMENT

As the given manuscript is a re-analysis of published data, the data, materials and codebooks should be requested from the corresponding authors of the original publications. The analysis code is available in the OSF (see *https://osf.io/jqzet/*). Detailed information on all analyses conducted in this article are available in the online supplementary materials.

NOTES

[1] In the following, the term 'target' describes any kind of entity evaluated on the SCM's warmth and competence dimensions.

[2] Please be aware that, unlike the other two, this research question was not originally pre-registered. For further information on the deviations from the preregistrations, see OSM-2.

[3] In this article, we focus on scalar measurement invariance as the precondition for valid mean-value comparison. We do so because mean-value comparisons are the most frequent application in SCM research. Nonetheless, our analysis always includes metric invariance as the precondition for comparative correlational analysis, which is somewhat rarer in SCM research.

[4] An overview of the minor points in which we deviated from the preregistrations is given in OSM-2.

[5] McDonald's ω is preferred to Cronbach's αbecause it does not assume tau-equivalence of the different items (Hayes & Coutts, 2020).

ADDITIONAL FILE

The additional file for this article can be found as follows:

- **Online Supplementary Material.** OSM-1 to 12. DOI: *https://doi.org/10.5334/irsp.613.s1*

ACKNOWLEDGEMENTS

We express our profound gratitude to all authors who contributed to this study by sharing their data.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

MTF: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – Original draft

PFK: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing – Review & editing

JB: Data Curation, Formal analysis, Investigation, Validation, Visualization, Writing – Original draft

AKCZ: Data curation, Formal analysis, Investigation, Methodology, Writing – Review & editing

TL: Data Curation, Formal analysis, Investigation, Validation, Writing – Review & editing

UW: Conceptualization, Resources, Supervision, Writing – Review & editing

FA: Conceptualization, Resources, Supervision, Writing – Review & editing

MHWVZ: Resources, Writing – Review & editing

AUTHOR AFFILIATIONS

Maria-Therese Friehs *orcid.org/0000-0002-5897-8226*
FernUniversität in Hagen, Germany

Patrick F. Kotzur *orcid.org/0000-0002-5193-3359*
Durham University, United Kingdom

Johanna Böttcher *orcid.org/0000-0002-9086-8924*
University of Osnabrück, Germany

Ann-Kristin C. Zöller
Hannover Medical School, Germany

Tabea Lüttmer
Currently not affiliated to an academic institution, Germany

Ulrich Wagner *orcid.org/0000-0001-6716-9212*
Philipps-University Marburg, Germany

Frank Asbrock *orcid.org/0000-0002-6348-2946*
Chemnitz University of Technology, Germany

Maarten H. W. van Zalk *orcid.org/0000-0002-0185-8805*
University of Osnabrück, Germany

REFERENCES

Abele, A. E., Ellemers, N., Fiske, S. T., Koch, A., & **Yzerbyt, V.** (2021). Navigating the social world: Toward an integrated framework for evaluating self, individuals, and groups. *Psychological Review*, *128*(2), 290–314. DOI: *https://doi.org/10.1037/rev0000262*

Abele, A. E., Hauke, N., Peters, K., Louvet, E., Szymkow, A., & **Duan, Y.** (2016). Facets of the fundamental content dimensions: Agency with competence and assertiveness—Communion with warmth and morality. *Frontiers in Psychology*, *7*. DOI: *https://doi.org/10.3389/fpsyg.2016.01810*

Abele, A. E., & **Wojciszke, B.** (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology*, *93*(5), 751–763. DOI: *https://doi.org/10.1037/0022-3514.93.5.751*

Asbrock, F. (2010). Stereotypes of social groups in Germany in terms of Warmth and Competence. *Social Psychology*, *41*(2), 76–81. DOI: *https://doi.org/10.1027/1864-9335/a000011*

Asbrock, F., Nieuwoudt, C., Duckitt, J., & **Sibley, C. G.** (2011). Societal stereotypes and the legitimation of intergroup behavior in Germany and New Zealand. *Analyses of Social Issues and Public Policy*, *11*(1), 154–179. DOI: *https://doi.org/10.1111/j.1530-2415.2011.01242.x*

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

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Press.

Boehnke, K., Arnaut, C., Bremer, T., Chinyemba, R., Kiewitt, Y., Koudadjey, A. K., Mwangase, R., & **Neubert, L.** (2014). Toward emically informed cross-cultural comparisons: A suggestion. *Journal of Cross-Cultural Psychology*, *45*(10), 1655–1670. DOI: *https://doi.org/10.1177/0022022114547571*

Boer, D., Hanke, K., & **He, J.** (2018). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology*, *49*(5), 713–734. DOI: *https://doi.org/10.1177/0022022117749042*

Brambilla, M., Sacchi, S., Rusconi, P., & **Goodwin, G. P.** (2021). The primacy of morality in impression development: Theory, research, and future directions. In *Advances in Experimental Social Psychology*, *64*, 187–262. Elsevier. DOI: *https://doi.org/10.1016/bs.aesp.2021.03.001*

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (Second edition). The Guilford Press.

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), 456–466. DOI: *https://doi.org/10.1037/0033-2909.105.3.456*

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: *https://doi.org/10.1080/10705510701301834*

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. DOI: *https://doi.org/10.1037/a0013193*

Davenport, E. C., Davison, M. L., Liou, P.-Y., & **Love, Q. U.** (2015). Reliability, dimensionality, and internal consistency as defined by Cronbach: Distinct albeit related concepts. *Educational Measurement: Issues and Practice*, *34*(4), 4–9. DOI: *https://doi.org/10.1111/emip.12095*

Davey, C., & **Hargreaves, J.** (2015, July 23). *How re-analysing the data of scientific research can change the findings*. Retrieved from *http://theconversation.com/how-re-analysing-the-data-of-scientific-research-can-change-the-findings-44926*

David, D., Bizo, A., Cimpean, A. I., Oltean, H., Cardos, R., Soflau, R., & **Negut, A.** (2018). The effect of research method type on stereotypes' content: A brief research report. *The Journal of Social Psychology*, *158*(3), 379–392. DOI: *https://doi.org/10.1080/00224545.2017.1361375*

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & **Billiet, J.** (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*(1), 55–75. DOI: *https://doi.org/10.1146/annurev-soc-071913-043137*

Durante, F., Fiske, S. T., Gelfand, M. J., Crippa, F., Suttora, C., Stillwell, A., Asbrock, F., Aycan, Z., Bye, H. H., Carlsson, R., Björklund, F., Dagher, M., Geller, A., Larsen, C. A., Latif, A.-H. A., Mähönen, T. A., Jasinskaja-Lahti, I., & **Teymoori, A.** (2017). Ambivalent stereotypes link to peace, conflict, and inequality across 38 nations. *Proceedings of the National Academy of Sciences*, *114*(4), 669–674. DOI: *https://doi.org/10.1073/pnas.1611874114*

Durante, F., Fiske, S. T., Kervyn, N., Cuddy, A. J. C., Akande, A. D., Adetoun, B. E., Adewuyi, M. F., Tserere, M. M., Ramiah, A. A., Mastor, K. A., Barlow, F. K., Bonn, G., Tafarodi, R. W.,

Bosak, J., Cairns, E., Doherty, C., Capozza, D., Chandran, A., Chryssochoou, X., …, & Storari, C. C. (2013). Nations' income inequality predicts ambivalence in stereotype content: How societies mind the gap. *British Journal of Social Psychology*, *52*(4), 726–746. DOI: *https://doi.org/10.1111/bjso.12005*

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*. DOI: *https://doi.org/10.3389/fpsyg.2015.00621*

Ellemers, N. (2021). Science as collaborative knowledge generation. *British Journal of Social Psychology*, *60*(1), 1–28. DOI: *https://doi.org/10.1111/bjso.12430*

Fiske, S. T. (2017). Going in many right directions, all at once. *Perspectives on Psychological Science*, *12*(4), 652–655. DOI: *https://doi.org/10.1177/1745691617706506*

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. DOI: *https://doi.org/10.1016/j.tics.2006.11.005*

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902. DOI: *https://doi.org/10.1037/0022-3514.82.6.878*

Fiske, S. T., Xu, J., Cuddy, A. C., & Glick, P. (1999). (Dis) respecting versus (dis)liking: Status and interdependence predict ambivalent stereotypes of competence and warmth. *Journal of Social Issues*, *55*(3), 473–489. DOI: *https://doi.org/10.1111/0022-4537.00128*

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. DOI: *https://doi.org/10.1177/2515245920952393*

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

Green, S. B., & Yang, Y. (2015). Evaluation of dimensionality in the assessment of internal consistency reliability: Coefficient alpha and omega coefficients. *Educational Measurement: Issues and Practice*, *34*(4), 14–20. DOI: *https://doi.org/10.1111/emip.12100*

Grigoryan, L., Bai, X., Durante, F., Fiske, S. T., Fabrykant, M., Hakobjanyan, A., Javakhishvili, N., Kadirov, K., Kotova, M., Makashvili, A., Maloku, E., Morozova-Larina, O., Mullabaeva, N., Samekin, A., Verbilovich, V., & Yahiiaiev, I. (2020). Stereotypes as historical accidents: Images of social class in postcommunist versus capitalist societies. *Personality and Social Psychology Bulletin*, *46*(6), 927–943. DOI: *https://doi.org/10.1177/0146167219881434*

Grigoryev, D., Fiske, S. T., & Batkhina, A. (2019). Mapping ethnic stereotypes and their antecedents in Russia: The stereotype content model. *Frontiers in Psychology*, *10*, 1643. DOI: *https://doi.org/10.3389/fpsyg.2019.01643*

Hackbart, M., Rapior, M., & Thies, B. (2020). Wie werden Erziehungsberatende in Abhängigkeit von Geschlechts- und ethnischer Zugehörigkeit kognitiv repräsentiert? [How are educational consultants cognitively represented as a function of gender and ethnicity?]. *Zeitschrift Für Soziologie Der Erziehung Und Sozialisation*, *40*(2), 116–132. DOI: *https://doi.org/10.3262/ZSE2002116*

Halkias, G., & Diamantopoulos, A. (2020). Universal dimensions of individuals' perception: Revisiting the operationalization of warmth and competence with a mixed-method approach. *International Journal of Research in Marketing*, *37*(4), 714–736. DOI: *https://doi.org/10.1016/j.ijresmar.2020.02.004*

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But…. *Communication Methods and Measures*, *14*(1), 1–24. DOI: *https://doi.org/10.1080/19312458.2020.1718629*

Hester, N., Axt, J., & Hehman, E. (2022, January 14). Evaluating Validity Properties of 25 Race-Related Scales. DOI: *https://doi.org/10.31234/osf.io/vxbtg*

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: *https://doi.org/10.1080/03610739208253916*

Hussey, I., & Hughes, S. (2020). Hidden invalidity among fifteen commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 166–184. DOI: *https://doi.org/10.1177/2515245919882903*

Janssens, H., Verkuyten, M., & Khan, A. (2015). Perceived social structural relations and group stereotypes: A test of the Stereotype Content Model in Malaysia: Social structure and stereotypes. *Asian Journal of Social Psychology*, *18*(1), 52–61. DOI: *https://doi.org/10.1111/ajsp.12077*

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, 44(3), 486–507. DOI: *https://doi.org/10.1177/0049124114543236*

Kervyn, N., Fiske, S. T., & Yzerbyt, V. Y. (2013). Integrating the stereotype content model (warmth and competence) and the Osgood semantic differential (evaluation, potency, and activity): Integrating the SD and the SCM. *European Journal of Social Psychology*, *43*(7), 673–681. DOI: *https://doi.org/10.1002/ejsp.1978*

Kervyn, N., Fiske, S., & Yzerbyt, V. (2015). Forecasting the primary dimension of social perception: Symbolic and realistic threats together predict warmth in the Stereotype Content Model. *Social Psychology*, *46*(1), 36–45. DOI: *https://doi.org/10.1027/1864-9335/a000219*

Koch, A., Imhoff, R., Dotsch, R., Unkelbach, C., & Alves, H. (2016). The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology*, *110*(5), 675–709. DOI: *https://doi.org/10.1037/pspa0000046*

Koch, A., Imhoff, R., Unkelbach, C., Nicolas, G., Fiske, S., Terache, J., Carrier, A., & **Yzerbyt, V.** (2020). Groups' warmth is a personal matter: Understanding consensus on stereotype dimensions reconciles adversarial models of social evaluation. *Journal of Experimental Social Psychology, 89*, 103995. DOI: *https://doi.org/10.1016/j.jesp.2020.103995*

Koch, A., Yzerbyt, V., Abele, A., Ellemers, N., & **Fiske, S. T.** (2021). Social evaluation: Comparing models across interpersonal, intragroup, intergroup, several-group, and many-group contexts. In *Advances in Experimental Social Psychology, 63*, 1–68. Elsevier. DOI: *https://doi.org/10.1016/bs.aesp.2020.11.001*

Kotzur, P. F., Forsbach, N., & **Wagner, U.** (2017). Choose your words wisely: Stereotypes, emotions, and action tendencies toward fled people as a function of the group label. *Social Psychology*, *48*(4), 226–241. DOI: *https://doi.org/10.1027/1864-9335/a000312*

Kotzur, P. F., Friehs, M., Asbrock, F., & **Zalk, M. H. W.** (2019). Stereotype content of refugee subgroups in Germany. *European Journal of Social Psychology*, *49*(7), 1344–1358. DOI: *https://doi.org/10.1002/ejsp.2585*

Kotzur, P. F., Veit, S., Namyslo, A., Holthausen, M., Wagner, U., & **Yemane, R.** (2020). 'Society thinks they are cold and/or incompetent, but I do not': Stereotype content ratings depend on instructions and the social group's location in the stereotype content space. *British Journal of Social Psychology*, *59*(4), 1018–1042. DOI: *https://doi.org/10.1111/bjso.12375*

Leach, C. W., Ellemers, N., & **Barreto, M.** (2007). Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology*, *93*(2), 234–249. DOI: *https://doi.org/10.1037/0022-3514.93.2.234*

Meitinger, K., Davidov, E., Schmidt, P., & **Braun, M.** (2020). Measurement invariance: Testing for it and explaining why it is absent. *Survey Research Methods*, 345–349. DOI: *https://doi.org/10.18148/SRM/2020.V14I4.7655*

Meredith, W., & **Millsap, R. E.** (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika*, *57*(2), 289–311. DOI: *https://doi.org/10.1007/BF02294510*

Meyer, B., & **Asbrock, F.** (2018). Disabled or cyborg? How bionics affect stereotypes toward people with physical disabilities. *Frontiers in Psychology*, *9*, 2251. DOI: https://doi.org/10.3389/fpsyg.2018.02251

Muthén, L. K., & **Muthén, B. O.** (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(4), 599–620. DOI: *https://doi.org/10.1207/S15328007SEM0904_8*

Pettigrew, T. F. (2018). The emergence of contextual social psychology. *Personality and Social Psychology Bulletin*, *44*(7), 963–971. DOI: *https://doi.org/10.1177/0146167218756033*

Pettigrew, T. F. (2021). *Contextual social psychology: Reanalyzing prejudice, voting, and intergroup contact*. American Psychological Association. *https://www.jstor.org/stable/j.ctv1gd0vp8*. DOI: *https://doi.org/10.1037/0000210-000*

Raykov, T., & **Marcoulides, G. A.** (2011). *Introduction to psychometric theory*. Routledge. DOI: *https://doi.org/10.4324/9780203841624*

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, *29*(4), 347–363. DOI: *https://doi.org/10.1177/0734282911406661*

Sayans-Jiménez, P., Cuadrado, I., Rojas, A. J., & **Barrada, J. R.** (2017). Extracting the evaluations of stereotypes: Bi-factor model of the stereotype content structure. *Frontiers in Psychology*, *8*, 1692. DOI: *https://doi.org/10.3389/fpsyg.2017.01692*

Schermelleh-Engel, K., Moosbrugger, H., & **Müller, H.** (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23–74.

Schimmack, U. (2020). A meta-psychological perspective on the decade of replication failures in social psychology. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 364–376. DOI: *https://doi.org/10.1037/cap0000246*

Sechrest, L., Fay, T. L., & **Zaidi, S. M. H.** (1972). Problems of Translation in Cross-Cultural Research. *Journal of Cross-Cultural Psychology*, *3*(1), 41–56. DOI: *https://doi.org/10.1177/002202217200300103*

Sörbom, D. (1978). An alternative to the methodology for analysis of covariance. *Psychometrika*, *43*(3), 381–396. DOI: *https://doi.org/10.1007/BF02293647*

Stanciu, A. (2015). Four sub-dimensions of stereotype content: Exploratory evidence from Romania. *International Psychology Bulletin*, *19*(4), 14–20.

Stanciu, A., Cohrs, J. C., Hanke, K., & **Gavreliuc, A.** (2017). Within-culture variation in the content of stereotypes: Application and development of the stereotype content model in an Eastern European culture. *The Journal of Social Psychology*, *157*(5), 611–628. DOI: *https://doi.org/10.1080/00224545.2016.1262812*

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. DOI: *https://doi.org/10.1177/109442810031002*

Vauclair, C.-M., Hanke, K., Huang, L.-L., & **Abrams, D.** (2017). Are Asian cultures really less ageist than Western ones? It depends on the questions asked. *International Journal of Psychology*, *52*(2), 136–144. DOI: *https://doi.org/10.1002/ijop.12292*

Wang, Y. A., & **Rhemtulla, M.** (2021). Power Analysis for Parameter Estimation in Structural Equation Modeling: A Discussion and Tutorial. *Advances in Methods and Practices in Psychological Science*, 4(1), 251524592091825. DOI: *https://doi.org/10.1177/2515245920918253*

Wetzel, E., & **Roberts, B. W.** (2020). Commentary on Hussey and Hughes (2020): Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *3*(4), 505–508. DOI: https://doi.org/10.1177/2515245920957618

Wolf, E. J., Harrington, K. M., Clark, S. L., & **Miller, M. W.** (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *73*(6), 913–934. DOI: https://doi.org/10.1177/0013164413495237

Yzerbyt, V., Provost, V., & **Corneille, O.** (2005). Not competent but warm... Really? Compensatory stereotypes in the French-speaking world. *Group Processes & Intergroup Relations*, *8*(3), 291–308. DOI: https://doi.org/10.1177/1368430205053944

TO CITE THIS ARTICLE:
Friehs, M.-T., Kotzur, P. F., Böttcher, J., Zöller, A.-K. C., Lüttmer, T., Wagner, U., Asbrock, F., & van Zalk, M. H. W. (2022). Examining the Structural Validity of Stereotype Content Scales – A Preregistered Re-Analysis of Published Data and Discussion of Possible Future Directions. *International Review of Social Psychology*, 35(1): 1, 1–18. DOI: https://doi.org/10.5334/irsp.613

Submitted: 17 May 2021 Accepted: 13 December 2021 Published: 09 February 2022

International Review of Social Psychology is a peer-reviewed open access journal published by Ubiquity Press.