On the Structural Validity, Measurement and Advanced Statistical Modelling of the Stereotype Content Model

Dissertation

zur Erlangung des Doktorgrades der Philosophie (Dr. phil.)

Fachbereich 8: Psychologie

Universität Koblenz-Landau

Deutschland

vorgelegt von

Maria-Therese Friehs (geb. Wiemer), M. Sc.

Landau in der Pfalz, 28.05.2021

Berichterstatter*innen:

Prof. Dr. Ulrich Wagner, Fachbereich Psychologie, Philipps-Universität Marburg, Deutschland

Prof. Dr. Julia Karbach, Fachbereich Psychologie, Universität Koblenz-Landau, Deutschland

Vorsitzende*r der Promotionskommission:

Prof. Dr. Melanie Steffens Fachbereich Psychologie, Universität Koblenz-Landau, Deutschland

Vom Promotionsausschuss des Fachbereichs 8: Psychologie der Universität Koblenz-Landau zur Verleihung des akademischen Grades Doktor der Philosophie (Dr. phil.) genehmigte Dissertation

Datum der wissenschaftlichen Aussprache: 19.07.2021

Acknowledgements

The present work aims at qualifying me to bear the title *doctor philosophiae*. And although this qualification is addressed at only one person, it would be a grave mistake to think that the underlying work was done single-handedly. This dissertation is a result of many hands and minds, and to all of them I express my deep, heartfelt and sincere gratitude and appreciation!!!

First of all, there are those who gave me the opportunity to write this thesis: I emphatically thank Ulrich Wagner and Julia Karbach for supervising me and allowing me to focus on my interests and projects. To both of you, let me say that it was great to know that I could always call on you when I needed your support, advice and guidance, but that otherwise I could work freely and choose to occupy myself with whatever interested me. This allowed me to orientate myself and grow, both professionally and as a person!

Then, of course, there are my brilliant co-authors, who set out with me to conduct this intriguing and fascinating research. Ann-Kristin, Felicia, Johanna, Tabea, Frank, Maarten, Patrick and Uli, it has been my great pleasure to collaborate with you, and this work could not have been done without you! My special thanks go to you, Patrick, for being my enduring supervisor, colleague, mentor, friend, (digital) lunch mate and fellow sufferer when commuting in the Deutsche Bahn (I have some fond recollections of the DB lounge in Hannover Central Station). Thanks for going through these projects with me, and let's have some more crazy ideas for the future!

Also, this dissertation would not have been possible without the assistance of many others. For one, these are the authors of all the studies we re-analysed. I sincerely thank you for your trust, openness and willingness to collaborate! I thank the Fachgruppe Sozialpsychologie of the Deutsche Gesellschaft für Psychologie, the Fachbereich Psychologie at the Philipps-Universität Marburg, and the Institute for Psychology at the Chemnitz University of Technology for providing the funds to conduct our research. Also, I thank Adelina Kletke, Annemarie Juchler, Corinne Lamadé, Julia Nolting, Lea Müller, Leonie Schmidt, Rebecca Cramer, Tobias Schmidt, and Vera Kaiser, who supported us in writing the article manuscripts as student assistants. And of course, I appreciate the assistance of all those who proof-read this work: My thanks to Alina Rinn, Leonie Schmidt, Maike Trautner and Patrick Kotzur! Lastly, there are those who have supported my work through inspiring discussions, intelligent questions and continuous interest, and for this, I thank Adrian Stanciu, Oliver Christ, Peter Schmidt, and many others.

I could not have managed to undertake this project without the constant and genial support of the brilliant people with whom I have the honour and pleasure of sharing my life. First and foremost, my husband Thilo, who never tired of discussing my projects and questions, and who always lifted me up and kept me grounded in exactly the right measure and precisely when I needed it. Where would I be without you? My great thanks also to my family, who encouraged me to be curious, inquisitive and ambitious. You laid the base for more than 20 years of scholarly and academic achievements! My heartfelt thanks also go to my splendid friends, with whom I could enjoy my studies, endless intelligent conversations, and infinite fun and support! Especially, these thanks go to you, Maike, for sharing so many aspects of my life (PhD included)! And finally, though this might be anthropomorphising, I acknowledge the effect of our dog Trixi, whose stern and critical gaze in my back (and occasional snort) accompanied and somewhat affected my work for more than a year now. You, too, earned your stripes!

Writing this dissertation has been one major objective of my life for more than three years now. It has been an amazing, inspiring, challenging and grounding task with ups and downs, and now, in the end, it seems this work leaves more questions open than it could answer. Nonetheless, and actually because of all this, I am deeply grateful that I have undertaken this adventure, and I emphatically thank all of you, whether called by name or not, for encouraging, accompanying, mentoring and supporting me on this way!!!

Table of Contents

Table of Figures

(not including Figures in manuscripts)

List of Tables

(not including Tables in manuscripts)

Short Summary

The *Stereotype Content Modell* (SCM; Fiske et al., 2002) proposes two fundamental dimensions of social evaluation: *Warmth*, or the intentions of the target, and Competence, or the ability to enact these intentions. The practical applications of the SCM are very broad and have led to an assumption of universality of warmth and competence as fundamental dimensions of social evaluation.

This thesis has identified five mainly methodological shortcomings of the current SCM research and literature: (I) An insufficient initial scale development; (II) the usage of varying warmth and competence scales without sufficient scale property assessment in later research; (III) the dominant application of first-generation analytical approaches; (IV) the insufficient definition and empirical proof for the SCM's assumption of universality; and (V) the limited application of the SCM for some social targets. These shortcomings were addressed in four article manuscripts strictly following open science recommendations.

Manuscript # 1 re-analysed published research using English SCM measures to investigate the measurement properties of the used warmth and competence scales. It reported the scales' reliability, dimensionality and comparability across targets as well as the indicator-based parameter performance in a (multiple group) confirmatory factor analysis framework. The findings indicate that about two thirds of all re-analysed scales do not show the theoretically expected warmth and competence dimensionality. Moreover, only about eleven per cent allowed meaningful mean value comparisons between targets. Manuscript # 2 presents a replication of Manuscript # 1 in the national and language of German(y) generating virtually identical results as Manuscript # 1 did. Manuscript # 3 investigated the stereotype content of refugee subgroups in Germany. We showed that refugees was generally perceived unfavourably in terms of warmth and competence, but that the stereotype content varied based on the refugees' geographic origin, religious affiliation, and flight motive. These results were generated using a reliability-corrected approach to compare mean values named alignment optimisation procedure. Manuscript # 4 developed and tested a high-performing SCM scale assessing occupational stereotypes a number of exploratory and confirmatory factor analyses.

Summary

Social perception is an essential process in everyday life which includes the differentiation and evaluation of different targets as well as the generation of cognitive, affective and behavioural reactions to them. In social evaluation research, the *Stereotype Content Modell* (SCM) presented by Fiske, Cuddy, Glick and Xu in 2002 has become quite prominent. The SCM proposes two fundamental and independent dimensions on which social evaluation processes take place. These dimensions are warmth, that is the benevolent or harmful intentions of the target, and competence, meaning the ability to enact these intentions. Together, warmth and competence perceptions predict affective and behavioural reactions to the evaluated target. Though the SCM has been proposed for the evaluation of several societally relevant groups, its practical applications are far broader and have led to an assumption of universality of warmth and competence as fundamental dimensions of social evaluation.

This thesis has identified five mainly methodological shortcomings of the current SCM research and literature. These are: (I) An insufficient initial scale development, which was followed by (II) the usage of varying warmth and competence scales without sufficient scale property assessment in later research; (III) the dominant application of first-generation analytical approaches instead of more reliable and valid advanced analytical strategies; (IV) the insufficient definition and empirical proof for the SCM's assumption of universality; and finally (V) the limited application of SCM research for some social targets which demand for more research activities. These shortcomings have been addressed with four article manuscripts, all of which strictly follow open science recommendations to increase the research's transparency and replicability.

Manuscript # 1 re-analysed published research using English SCM measures to investigate the measurement properties of the used warmth and competence scales. It reported the scales' reliability, dimensionality and comparability across targets as well as the indicator-based parameter performance in a (multiple group) confirmatory factor analysis framework. The findings indicate that about two thirds of all re-analysed scales do not show the theoretically expected warmth and competence dimensionality, even though the average reliabilities of the scales were acceptable.

Moreover, only about eleven per cent of all scales showed the preconditions for meaningful mean value comparisons between targets, which is the most frequent analytical application of the SCM. The manuscript also informs of the performance of different warmth and competence indicators, discusses the implications of the findings and presents ideas how to improve the validity and reliability of future SCM research. These ways forward include the thorough development of new SCM scales using state-of-the art methodology, the standard application of confirmatory factor analysis in applied research to test the scales' dimensionality and comparable measurement properties across targets as a precondition for meaningful mean value comparisons, the application of more advanced analytical procedures and the usage of larger samples. The manuscript calls for collective and collaborative efforts of different SCM researchers to implement these changes.

Manuscript # 2 presents a replication of Manuscript # 1 in another national and language context by reanalysing SCM data collected in German(y). It generated virtually identical results as Manuscript # 1 did, thus emphasising the issue of insufficiently validated warmth and competence scales with questionable measurement properties in SCM research and all construct validity threats associated with these issues.

Manuscript # 3 investigated the stereotype content of refugee subgroups in Germany, a social group of at-the-time high societal relevance but about which existed little previous knowledge in Germany. We showed that the social category refugees was generally perceived rather unfavourable in terms of warmth and competence, but that the stereotype content varied between different refugee subgroups based on their geographic origin, religious affiliation, and flight motive. These results were generated using a reliability-corrected approach to compare mean values named alignment optimisation procedure, which relied both on the test of dimensionality of the warmth and competence scales and on the establishment of comparable measurement properties between refugee subgroups when comparing warmth and competence mean values. Nonetheless, our analyses were severely limited by the unacceptable dimensionality of the used warmth and competence scale in six out of 16 cases.

These limitations were addressed in Manuscript # 4, which aimed at developing a high-performing SCM scale using state-of-the-art methodology. We developed warmth and competence scales suited explicitly to the assessment of occupational stereotypes and demonstrated their high performance with a number of exploratory and confirmatory factor analyses. Subsequently, we analysed the warmth and competence perceptions of different occupational groups using an identical methodology as Manuscript # 3 did.

The series of studies presented in this dissertation demonstrates comprehensively the methodological challenges and issues associated with the currently published SCM literature. What is more, it both theoretically discusses and empirically demonstrates ways to overcome these issues. This dissertation thus gives impulses for more reliable, valid and meaningful future research on social evaluation using the Stereotype Content Model.

Deutsche Zusammenfassung

Soziale Wahrnehmung ist ein grundlegender alltäglicher Prozess, der die Differenzierung und Bewertung verschiedener Objekte sowie die Ermittlung gedanklicher, emotionaler und verhaltensbezogener Reaktionen gegenüber diesen Objekten beinhaltet. Im Bereich der Forschung zu sozialen Bewertungsprozessen hält das *Stereotype Content Modell* (SCM) von Fiske, Cuddy, Glick und Xu aus dem Jahr 2002 eine prominente Rolle inne. Das SCM schlägt zwei fundamentale und unabhängige Dimensionen vor, auf denen soziale Bewertungsprozesse stattfinden. Diese Dimensionen sind einerseits Wärme, also die freundliche oder feindliche Intention des Bewertungsobjekts, und andererseits Kompetenz, also die Fähigkeit, besagte Intentionen in die Tat umzusetzen. Das Zusammenspiel aus Wärme- und Kompetenzwahrnehmung bedingt die emotionalen und verhaltensbezogenen Reaktionen gegenüber dem Bewertungsobjekt. Obwohl das SCM zur Bewertung mehrerer sozialer Gruppen vorgeschlagen wurde, geht die praktische Anwendung des Modells weit darüber hinaus und hat zu der Annahme geführt, Wärme und Kompetenz seien universelle Dimensionen der sozialen Bewertung.

Die vorliegende Doktorarbeit hat fünf vor allem methodologische Schwächen der SCM-Forschung und -Literatur identifiziert. Diese sind: (I) eine unzureichende anfängliche Skalenentwicklung, der (II) eine Nutzung diverser und variierender Skalen ohne hinreichende Prüfung der Skalenperformanz in der anschließenden SCM-Forschung folgte; (III) die vorherrschende Nutzung von Analysemethoden der ersten Generation anstatt der Anwendung fortgeschrittener Analysestrategien mit höherer Reliabilität und Validität; (IV) die unzureichende Definition und empirische Testung der Universalitätsannahme des SCM; und schlussendlich (V) die eingeschränkte Anwendung des SCM in Bezug auf einige soziale Gruppen, für die großer Bedarf an Informationen zur sozialen Wahrnehmung besteht. Diese Schwächen wurden in vier wissenschaftlichen Artikelmanuskripten aufgegriffen, welche allesamt streng den Empfehlungen der offenen Wissenschaft folgten, um die Transparenz und Replizierbarkeit der Forschung zu erhöhen.

Manuskript # 1 nutzte veröffentlichte Daten von englischen SCM-Skalen, um die Eigenschaften der genutzten Wärme- und Kompetenz-Skalen zu reanalysieren. Das Manuskript berichtet die Skalenreliabilität, Dimensionalität und Vergleichbarkeit der Skalen über verschiedene Bewertungsobjekte hinweg sowie verschiedene Performanz-Parameter der einzelnen Indikatoren auf im Rahmen einer (Multi-Gruppen-) konfirmatorischen Faktor-Analyse.

Die Befunde zeigen, dass zwei Drittel aller reanalysierten Skalen nicht die theoretisch angenommene Wärme- und Kompetenz-Dimensionalität haben, obwohl die mittleren Reliabilitäten der Skalen akzeptabel waren. Weiterhin zeigten nur elf Prozent aller Skalen die Voraussetzungen für aussagekräftige Mittelwertsvergleiche zwischen Bewertungsobjekten, was die häufigste Form der Auswertung von SCM-Daten ist. Das Manuskript präsentiert außerdem die Performanz verschiedener Wärme- und Kompetenz-Indikatoren, diskutiert die Folgerungen aus unseren Ergebnissen und zeigt Ideen zur Verbesserung der Validität und Reliabilität in zukünftiger SCM-Forschung auf. Diese Ideen beinhalten die gründliche Entwicklung neuer SCM-Skalen unter der Nutzung modernster Methoden, die standardmäßige Anwendung von konfirmatorischen Faktorenanalysen zur Prüfung der Skalendimensionalität und der Vergleichbarkeit der Messeigenschaften über verschiedene Bewertungsobjekte als Voraussetzung für aussagekräftige Mittelwertsvergleiche, die Anwendung fortgeschrittener Datenanalyse-Strategien und die Nutzung größerer Stichproben. Das Manuskript wirbt für die gemeinschaftliche und kollegiale Anstrengungen verschiedener SCM-Forschenden, um diese Veränderungen zu realisieren.

Manuskript # 2 beschreibt eine Replikation von Manuskript # 1 in einem anderen nationalen und Sprachkontext durch die Reanalyse von deutsch(sprachig)en SCM-Daten. Dieses Manuskript zeigt nahezu identische Ergebnisse wie Manuskript # 1 und stellt so das Problem der unzureichend validierten Wärme- und Kompetenzskalen mit fraglichen Messeigenschaften in der SCM-Forschung sowie die damit einhergehenden Gefährdungen der Validität heraus.

Manuskript # 3 untersucht die soziale Wahrnehmung von Subgruppen von Geflüchteten in Deutschland als eine soziale Gruppe, die zu seiner Zeit hohe gesellschaftliche Relevanz in Deutschland besaß, über die aber kaum Wissen vorhanden war. Wir konnten zeigen, dass die soziale Kategorie der Geflüchteten generell unvorteilhaft in Bezug auf Wärme und Kompetenz wahrgenommen wurde, aber dass die soziale Wahrnehmung der Subgruppen von Geflüchteten sich in Abhängigkeit der geografischen Herkunft, der religiösen Zugehörigkeit und der Fluchtgründe unterschied. Diese Befunde wurden durch die Nutzung eines Reliabilitäts-korrigierenden Verfahrens zum Vergleich von Mittelwerten namens Alignment-Optimierung generiert. Dieses beruhte sowohl auf dem Test der Dimensionalität der Wärme- und Kompetenz-Skalen, als auch auf der Etablierung vergleichbarer Messeigenschaften zwischen den unterschiedlichen Subgruppen von Geflüchteten beim Vergleich von Wärme- und Kompetenz-Mittelwerten. Nichtsdestotrotz wurde der Umfang unserer Analysen stark eingeschränkt, das sechs von insgesamt 16 Subgruppen von Geflüchteten keine angemessene Dimensionalität in den genutzten Wärme- und Kompetenzskalen aufwiesen.

Diese Schwäche wurde in Manuskript # 4 thematisiert, welches auf die Entwicklung einer leistungsfähigen SCM-Skala unter Nutzung modernster Analysemethoden abzielte. Wir entwickelten Wärme- und Kompetenzskalen, welche speziell auf die Erfassung von Stereotypen von Berufsgruppen abgestimmt sind, und zeigten ihre hohe Leistungsfähigkeit in einer Reihe explorativer und konfirmatorischer Faktorenanalysen. Anschließend werteten wir die Wärme- und Kompetenzwahrnehmung verschiedener Berufsgruppen mit derselben Methodik aus, die wir auch in Manuskript # 3 genutzt haben.

Die Serie von Studien, die in dieser Doktorarbeit präsentiert wird, zeigt umfassend die methodologischen Herausforderungen und Probleme auf, die in der derzeitig veröffentlichten SCM-Literatur vorliegen. Außerdem wird theoretisch diskutiert und praktisch demonstriert, wie diese Schwächen überkommen werden können. Somit gibt diese Doktorarbeit Impulse für eine reliablere, validere und aussagekräftigere zukünftige Soziale-Wahrnehmungs-Forschung unter Nutzung des Stereotype Content Modells.

I. Theoretical and Empirical Background

Social Perception and Social Evaluation

Social perception is an essential process of everyday life. Whenever we choose service providers such as medical or legal professionals, cast votes for political representatives, or simply assess an approaching pedestrian on the walkway, information (or assumptions) about who a person is or what he or she intends to do are highly relevant to navigate the social world. Social perception includes three processes: For one, social stimuli need to be differentiated and categorised (e.g., the differentiation between medical practitioners and other forms of health-related professionals). Further, the social stimuli are evaluated along important dimensions (e.g., assessing whether consulting a particular medical practitioner sustains one's well-being). Finally, people generate cognitive, affective and behavioural reactions to these targets (e.g., the decision whether or not to make an appointment; Zebrowitz, 1990). These processes apply to the evaluation of oneself and other individuals as well as larger social groups or societies (Abele et al., 2021).

Extensive research asserts that social evaluation is not uni-dimensional (i.e., globally negative versus positive), but that these processes are more differentiated and include the evaluation on (at least) two fundamental dimensions (e.g., Asch, 1946; Markey, 2002; Rosenberg et al., 1968). These two dimensions assess (I) how the target of evaluation forms social bonds (i.e., social-emotional criteria, often referred to as *communion*, *warmth*, or *getting along*), and (II) how they accomplish tasks (i.e., task- and performance-related criteria, often referred to as *agency*, *competence* or *getting ahead*; Abele et al., 2021; Koch et al., 2021). Such two-dimensional models of social evaluation have been put forth by a variety of psychological disciplines, including evolutionary theorising (Chan et al., 2019; Ybarra et al., 2008), cultural psychology (Markus & Kitayama, 1991), self-presentation research (Paulhus, 2019), motivation research (Locke & Schattke, 2018), developmental psychology (Erikson, 1950), gender perception (Eagly & Steffen, 1984; Sczesny et al., 2019), personality psychology (Saucier, 2009) and face perception research (Willis & Todorov, 2006). Social psychology, too, has a long tradition of multi-dimensional social evaluation research and a variety of theoretical frameworks have been proposed and substantially supported by empirical findings (e.g., Abele & Wojciszke, 2007; Fiske et al., 2002; Koch et al., 2016; Leach et al., 2007; Yzerbyt et al., 2005).

The Stereotype Content Model

One of these social psychological frameworks of social perception is the *Stereotype Content Model* (SCM; Fiske et al., 2002), which investigates the cognitive facet of social evaluation processes (i.e., stereotypes; Dovidio et al., 2010). Rather than explaining processes that lead to the emergence and persistence of stereotypes, the SCM aims at studying their meaning or content. It thus became one of the most prominent social psychological theoretical frameworks of social evaluation, counting more than 340 published research articles in the Web of Science database which mentioned the SCM in the title, abstract or keywords by the end of May 2021. The SCM builds on the above-mentioned research by proposing two fundamental dimensions of social evaluation, namely *warmth*, meaning "the intentions of the other person or group" (Fiske et al., 2007, p. 77), and *competence,* referring to the "ability to act on those intentions" (Fiske et al., 2007, p. 77). Warmth can be described using traits such as friendliness, helpfulness, sincerity, trustworthiness or morality, and is predicted negatively by the socio-structural aspects of competition and perceived threat (Fiske et al., 2002; Cuddy et al., 2008; Kervyn et al., 2015). Competence includes traits like intelligence, skill, creativity and efficacy (Fiske et al., 2007), and is positively predicted by perceived status as a socio-structural determinant (Fiske et al., 2002; Cuddy et al., 2008). Warmth and competence assessments are theorised to be independent from one another (i.e., orthogonal dimensions), but empirical findings often report correlations between the two constructs (Durante et al., 2013; Kervyn et al., 2010, 2015).

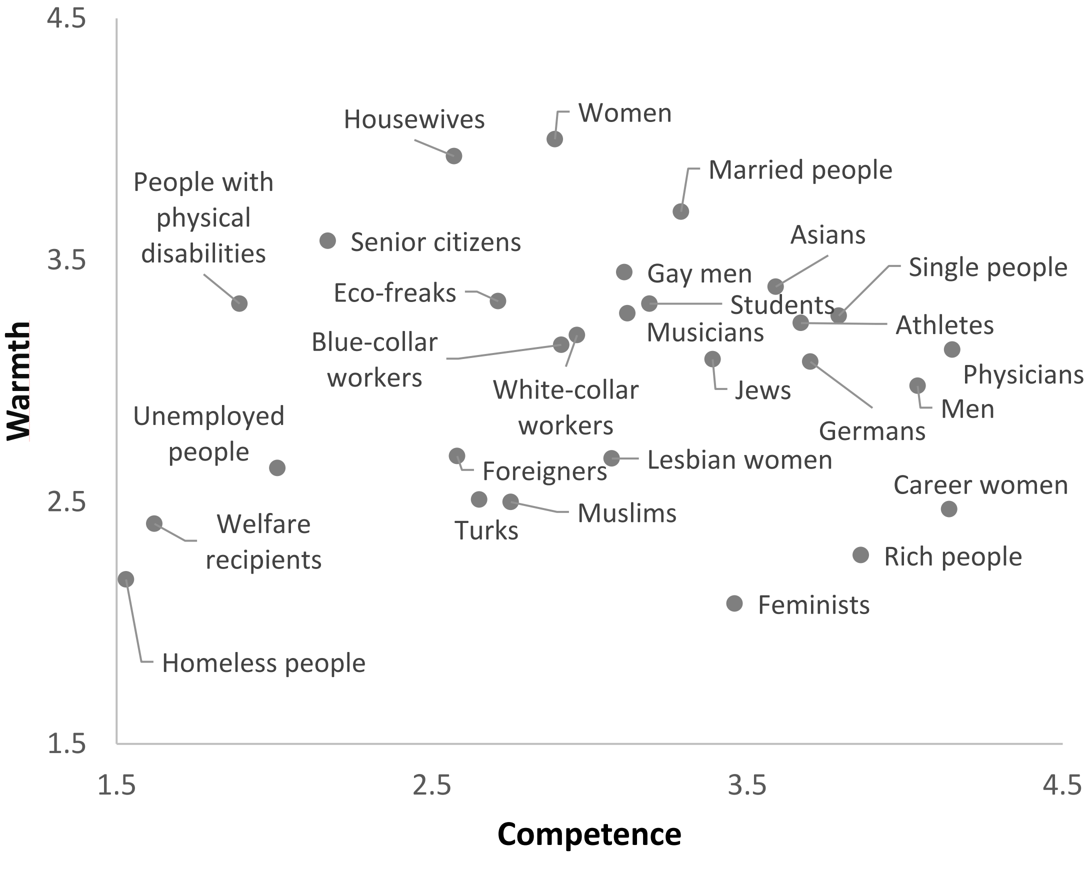
The SCM is traditionally applied to evaluate a number of different social groups from a societal perspective (i.e., rating target groups[1] on warmth and competence "as viewed by society" rather than asking for the individual perceptions of the survey participants; Fiske et al., 2002, p. 896).

[1] In the following, the stimuli that are subjected to warmth and competence evaluations will consistently be referred to as 'target groups'. The target groups could include individuals, social groups, organizations, or any other kind of stimuli.

Often, the SCM applications differentiate the evaluated social groups into different clusters on the base of to their relative warmth and competence ratings (Fiske et al., 2002). According to Fiske (2018), and as displayed in Figure 1, these clusters might include: (I) Target groups that are perceived as high in both warmth and competence, such as societal prototypes (e.g., Christians and Middle class in the US) evoking emotional reactions of pride and admiration; (II) target groups that are rated high in warmth, but low in competence (e.g., elderly or disabled people in the US), which evoke paternalistic prejudice and emotions of pity and sympathy; (III) target groups that are evaluated as low in warmth, but high in competence (e.g., rich people and Asians in the US) and which evoke emotional reactions of envy and jealousy; and (IV) target groups that are perceived as low in both warmth and competence (e.g., poor or homeless people in the US) and that evoke disgust and contempt.

Figure 1

Stereotype Content of Social Groups in Germany (Figure Adapted from Asbrock, 2010, p. 78)



Note. The answering scales ranged from 1 *(not at all)* to 5 *(completely)*.

As an extension of the SCM, the *Behaviours from Intergroup Affect and Stereotypes (BIAS) map* expanded the cognitive warmth and competence evaluations and the associated emotional reactions to also include behavioural (intentional) reactions (Cuddy et al., 2007, 2008). According to the BIAS map, behaviour patterns might be differentiated between active and passive behaviours as well as facilitating and harming behaviours. Warmth perceptions are theorised to determine active facilitating (in case of high warmth perceptions; e.g., helping) or active harming behaviours (in case of low warmth perceptions; e.g., harassing). At the same time, competence perceptions should determine passive facilitation (in case of high competence perceptions; e.g., associating with) and passive harm behaviours (in case of low competence perceptions; e.g., neglecting; Cuddy et al., 2007, 2008).

In the past 20 years, the SCM has been applied in numerous and multi-facetted ways in a variety of social and research contexts. These applications include, but are not limited to, identification and evaluation of societally relevant groups in more than 50 countries (e.g., Asbrock, 2010; Fiske, 2019; Fiske et al., 2002; The Fiske Lab, n.d.), examinations of transcultural variations in stereotype content (e.g., Cuddy et al., 2009, Durante et al., 2013, 2017) and regional- or sample-dependent differences (e.g., Binggeli et al., 2014a; Stanciu et al., 2017), analyses of diverging (sub)group perceptions as a function of the used labels (e.g., Binggeli et al., 2014b; Fröhlich & Schulte, 2019; Kotzur et al., 2017; Lee & Fiske, 2006), and the analysis of the stereotype content of individuals (e.g., Janda et al., 2019) and non-human target groups including brands (e.g., Kervyn et al., 2014), animals (e.g., Sevillano & Fiske, 2019) and geometric forms (e.g., Oldmeadow, 2018). Consistent emergence of warmth and competence dimensions in a great variety of geographic, temporal, societal and research contexts lead to the conclusion that „the two dimensions of intergroup perception appear to be universal across more than 30 nations (…) and 75 years (…), as well as targets that are individuals, subgroups, groups, nations, corporations, and species" (Fiske & North, 2015, p. 688).

When comparing the different models of social perception (e.g., the SCM; the Dual Perspective Model, Abele & Wojciszke, 2007, 2014, 2019; the Agency Beliefs Communion (ABC)

Model, Koch et al., 2016; the Behavioural Regulation Model, Ellemers, 2017; Leach et al., 2007; the Dimensional Compensation Model, Kervyn et al., 2010; Yzerbyt, 2018; Yzerbyt et al., 2005), it becomes apparent that all models agree in proposing (at least) two dimensions of social perception[2]. Nonetheless, the theoretical frameworks diverge considerably with regard to the labels and definition of relevant dimensions of social perception, as well as the dimensions' relationship, organisation and priority (for further elaboration, see Abele et al., 2021; Koch et al., 2021). What is more, the different theoretical frameworks have been developed for different targets of evaluation (e.g., self and specific other individuals for Abele & Wojciszke, 2007). Compared to the other mentioned models of social perception, the SCM is uniquely adapted to assess the social evaluation of several different target groups (e.g., more than 20 groups in Fiske et al., 2002) from a societal, non-individual perspective. This conceptual area of application, combined with the elaborate theoretical reasoning and the wide range of empirical research in different contexts, led us[3] to select the SCM as theoretical basis for the following work.

Selected Shortcomings of the Stereotype Content Model and The Related Literature

Despite the prominent position of the SCM, some aspects have received little attention in the development as well as the application of the model and can thus be criticised. A selection of these issues will be described in the following.

Shortcoming 1: Insufficient Initial Scale Development

The development of the SCM measurements was a step-wise and enduring process which has been documented in Fiske et al. (1999, 2002). Initially, on the basis of a comprehensive literature review including more than 85 years of empirical psychological research (e.g., Allport, 1954; Asch, 1946; Bakan, 1956; Conway et al., 1996; Eagly, 1987; Gilbert, 1951; Katz & Braly, 1933), Fiske et al. (1999) generated a pool of positively and negatively framed trait adjectives relating to warmth and competence. From this pool, 27 traits were selected to assess the social perception of 17 target

[2] Koch et al. (2016) propose the additional dimension *conservative versus progressive beliefs*.

[3] Although this dissertation constitutes my individual work of qualification, the underlying scientific work was conducted not single-handedly, but by a group of authors. Thus, I decided to use the personal pronoun 'we' in its academic sense throughout the dissertation.

groups. Subsequently, the authors conducted oblique exploratory factor analyses for each target group separately, identifying ten traits[4] that related most consistently to the warmth and competence factors across target groups. Importantly, the results of the exploratory factor analysis supported the item selection in most, but not in all cases. In subsequent research, Fiske et al. (2002, study 1) used nine out of the ten originally identified indicators and included as well as excluded further indicators for both warmth and competence scales (e.g., Fiske et al., 2002, study 2).

We agree with Halkias and Diamantopoulos (2020), who have strongly criticised the above-described approach from a psychometric perspective due to a number of reasons. These issues include very low sample sizes in all studies using the computationally demanding exploratory factor analyses of warmth and competence ($n_{\text{Study 1}}$ = 42 in Fiske et al., 1999; $n_{\text{Study}\mathit{1}}$ = 74, n_{Study2} = 148 with a sample split in Fiske et al., 2002). Such low sample sizes risk providing unstable results (MacCallum et al., 1999), and this fact might account for the exploratory factor analyses revealing up to three more scale dimensions than theoretically assumed in at least some of the target groups (Fiske et al., 2002, studies 1 and 2). Also, for some target groups, the chosen warmth and competence items did not relate exclusively to the theoretically expected factors (Fiske et al., 1999). What is more, Fiske and colleagues (1999) computed scale reliabilities and item inter-correlations by aggregating the data across target groups, thus effectively reducing the sample size to K = 17 groups. This might have masked potential differences in the target groups' scale reliability and/or the relationship of warmth and competence within the groups. Furthermore, the participants were presented with very high numbers of items (e.g., in Fiske et al., 1999, study 1, 27 traits x 17 target groups = 459 items), which might have triggered fatigue or satisficing effects and thus potentially threatened the validity of the results (Krosnick, 1991, 1999; Podsakoff et al., 2012; Shadish et al., 2002). Additionally, Fiske et al. (2002) did not present any explanation why certain items were dropped and did not disclose the origin of the newly integrated scale items, which resulted in low transparency in the scale development process. Lastly, research building on the initial scale development (e.g., Cuddy et al.,

[4] Warmth indicators: *likeable, sincere, good-natured, warm, tolerant*; Competence indicators: *competent, intelligent, confident, competitive, independent.*

2009) did not use more robust, informative and conclusive confirmatory techniques (e.g., confirmatory factor analysis). Such analyses would have added valuable information because they allow for a more comprehensive examination of scale performance properties and a more thorough development of warmth and competence scales (Brown, 2015). All the mentioned issues can be summarised as *questionable measurement practises*, which are defined as “decisions researchers make that raise doubts about the validity of the measures, and ultimately the validity of study conclusions” (e.g., non-transparency regarding the measurement due to the usage of measurements without reference to the source, lacking evidence of construct validity, or unjustified measurement flexibility; Flake & Fried, 2020, p. 456). Such questionable measurement practises should not be equalised with intentional scientific misconduct. But in the case of the SCM, the results of the questionable measurement practises during the scale development process are warmth and competence measurements with debatable features due to the low transparency in item selection and the absence of adequate analytical techniques to assess item- and scale performance.

Shortcoming 2: Insufficient Scale Performance Assessment and Varying Scales in Later Applications

In addition to the initial scale development process, we observed two dynamics which can also be categorised as questionable measurement practises: On the one hand, the critical scale development process notwithstanding, subsequent SCM research referred to the initial SCM scales to justify their choices of measurements. For instance, in the German SCM research context, Eckes (2002) used a German translation of the items from Fiske et al. (1999). A shorter version of these scales including three items per subscale was later used by Asbrock (2010), and this short version was applied for instance by Hansen et al. (2017, 2018), Hellmann et al. (2015), Kemme et al. (2020), Kotzur et al. (2017, 2020) and Kotzur, Schäfer et al. (2019). All of these German applications either conducted only exploratory approaches to scale performance (e.g., principal component analysis in Asbrock, 2010) or only reported the scales’ internal consistency (e.g., Eckes, 2002; Hansen et al., 2017, 2018; Hellmann et al., 2015; Kotzur et al., 2017; Kotzur, Schäfer et al., 2019). This lack of more advanced scale performance analyses might have concealed weaknesses resulting from the initial scale development. This assumption was empirically supported by Kotzur and colleagues (2020), who conducted confirmatory factor analyses and, as a consequence, were forced to severely limit their principal analyses due to the fact that the warmth and competence items could not empirically be summarised into scales with acceptable dimensionality. To summarise, an SCM scale with insufficiently tested performance was sustained and perpetuated through cross-referencing, potentially transmitting the critical issues of the scale development process presented above to successive applications.

On the other hand, in subsequent applications of the SCM, it could be observed that the list of items measuring warmth and competence was remarkably unstable and broad. As in the initial scale development process, items were often included and excluded seemingly arbitrarily in applied research without giving any reasoning or origin of the newly-used items (e.g., Caprariello et al., 2009, Meagher, 2017). As a consequence, the landscape of SCM research is strongly fragmented concerning the precise item content used to measure warmth and competence. We acknowledge that this practise is quite prevalent in (social) psychology (Flake & Fried, 2020), but nonetheless, we consider this issue problematic because the variations in measurement instruments might have resulted in unintended variations in the theoretical conceptualisation of warmth and competence and measurement validity across different studies (Halkias & Diamantopoulos, 2020). Moreover, the assessment of scale properties lacked rigor in most studies, thus scales of unknown content and structural validity (i.e., construct validity based on acceptable psychometric properties of the measurement, such as item performance analysis, scale reliability, scale dimensionality and differential item functioning; Flake et al., 2017) were applied. There are a few and fairly recent exceptions in the SCM literature which investigated the scale dimensionality applying confirmatory factor analyses to SCM measurements (Grigoryan et al., 2020; Hackbart et al., 2020; Halkias & Diamantopoulos, 2020; Janssens et al., 2015; Kotzur et al., 2020; Stanciu, 2015; Stanciu et al., 2017; Vauclair et al., 2016). But on the whole, we cannot be certain whether different SCM studies measured the same theoretical ideas of warmth and competence and how well they measured these concepts in general. These issues hinder a valid interpretation and comparison (Flake & Fried, 2020) as well as a broader theoretical integration of the findings and result in low generalisability

(Maruyama, 1997). Additionally, this practise of varying measurement instruments poses serious challenges to cumulative science (in the SCM, e.g., Cuddy et al., 2009; Durante et al., 2013, 2017; see also Schimmack, 2010). To initiate steps towards an improved measurement practise, Halkias and Diamantopoulos (2020) recently presented a thoroughly developed and tested German-language scale of Stereotype Content in the marketing context.

Shortcoming 3: Lack of Advanced Analytical Approaches to Analyse Principal Research Questions

It is noticeable that SCM research has used mainly basic analytical approaches, not only for the assessment of scale properties, but also for the analysis of the general research questions. Given the distinct application of the SCM to assess the social perception of several different target groups (e.g., comparing the social perception of the targets *Welfare recipients* and *Rich people*; Abele et al., 2021; Fiske et al., 2002; Koch et al., 2021), it is not surprising that most research has essentially compared the average warmth and competence ratings of different target groups. This is mostly done by either comparing observed warmth and competence means in *t*-tests and analyses of variance (e.g., Hansen et al., 2017, 2018; Hellmann et al., 2015; Janda et al., 2019; Kotzur et al., 2017) or by identifying patterns of social perception in cluster analysis (e.g., Asbrock, 2010; Cuddy et al., 2009; Eckes, 2002; Fiske et al., 2002; Fröhlich & Schulte, 2019). Less common research applications use warmth and competence assessments in correlational or regression-based research (e.g., whether the warmth and competence perceptions of the target group *Welfare recipients* relate differently to emotional and behaviour-intentional reactions compared to the warmth and competence perceptions of the target group *Rich people*; e.g., Cuddy et al., 2007; Durante et al., 2013, 2017; Kotzur, Schäfer et al., 2019). In summary, the overwhelming majority of these SCM applications use first-generation analytical techniques based on observed item or scale scores. These techniques have the great disadvantage of not accounting for measurement error (i.e., being less reliable; Brown, 2015). Moreover, in some cases, the analyses are conducted for the two SCM dimensions independently (e.g., in ANOVAs) without accounting for the potential covariation between these two dimensions, which potentially biases the findings (Kline, 2015). Advanced statistical alternatives to these techniques which use latent variable modelling to compensate for these drawbacks, such as latent mean value comparisons in the framework of multi-group confirmatory factor analysis (Davidov et al., 2014) or alignment optimization (Asparouhov & Muthén, 2014), latent class/latent profile analysis instead of cluster analysis (Berlin et al., 2013; Lubke & Muthén, 2005) or structural equation and path modelling instead of observed regression-based or correlational analyses (Kline, 2015). However, the application of such methods is extremely rare in SCM research (exceptions are Kotzur, Schäfer et al., 2019; Kotzur et al., 2020). Consequently, we identify the need for more applications of advanced analytical techniques to improve the overall reliability and validity of SCM research.

Shortcoming 4: Insufficiently Examined Assumption of Universality

As outlined above, the original authors proposed that the Stereotype Content Model's dimensions "warmth and competence [are] universal dimensions of social judgment, across perceivers, stimuli, cultures, and time" (Cuddy et al., 2008, p. 137). This assumption is based on the fact that targets could be compared along warmth and competence dimensions in a variety of contexts, for instance both in individualistic and collectivistic countries (Cuddy et al., 2009), for target groups that are human groups (e.g. Fiske et al., 2002), individuals (e.g., Hansen et al., 2017; 2018; Janda et al., 2019), animals (e.g., Sevillano & Fiske, 2019) or non-human objects like countries (e.g., Crandall et al., 2011), brands (e.g., Kervyn et al., 2014) or geometric figures (Oldmeadow, 2018), or in an archival study examining newspaper articles published between 1938 and 1943 in Fascist Italy (Durante et al., 2010). We agree with the SCM's authors that this repetitive pattern of warmth and competence dimensions is impressive, but for multiple reasons, we question whether this pattern indeed implies universality. One reason is that the limited scale performance examinations reported in most SCM research (see above) potentially facilitated finding such a pattern, as more rigorous tests (e.g., of the scales' dimensionality using confirmatory factor analyses) which might have uncovered differing scale performances were mostly omitted. What is more, we observe that although the claim of universality is quite prominent in many of the original authors' publications (e.g., Cuddy et al., 2008, 2009; Fiske, 2017; Fiske & North, 2015; Fiske et al., 2007), none of these publications defined what precisely is meant by the term 'universal'. Only relatively recently, Fiske

(2017) elaborated more strongly on the underlying concept of universality in the SCM by explicating that, although usage of the fundamental dimensions warmth and competence was omnipresent, the average warmth and competence assessment might vary considerably between cultures for some, but not all target groups.

We criticise that none of the above-mentioned publications relied on the extensive intercultural psychological theorising (e.g., Hui & Triandis, 1985; van de Vijver & Poortinga, 1982) to define and examine the SCM's universality. As a consequence, we aim to take this perspective when examining the SCM in the following. From an intercultural psychological perspective, in order to avoid systematic bias when comparing the results of different cultural context (for SCM applications, see e.g., Cuddy et al., 2009; Durante et al., 2013, 2017), it is (beyond other aspects) essential to ensure that psychological constructs such as warmth and competence are defined and understood equally in all cultural contexts. If that is the case, it is also important that measurement instruments assessing warmth and competence show equivalent measurement properties in different cultural contexts.

We further argue that in the context of the SCM, the said methodological preconditions of universality should not only be applied to studies carried out in different cultural contexts, but also to the different target groups that are being evaluated within one study. This is because such conceptual equivalence constitutes a fundamental precondition for meaningful interpretations of mean-value comparisons (e.g., Davidov et al., 2014), which are the most-frequent analytical strategy in SCM research. To put it more clearly, if for instance a study compares the social perception of the target groups *Welfare recipients* and *Rich people* (e.g., Asbrock, 2010; Fiske et al., 2002), then the survey participants' understanding of warmth and competence need to be identical in both groups to produce meaningful mean-value comparisons. Such conceptual equality could be examined using qualitative (e.g., cognitive interviewing or online probing techniques; Achbari & Davidov, 2019; Benítez & Padilla, 2014; Latcheva, 2011; Meitinger, 2017) or quantitative measures (e.g., factor analyses in a multi-group context; Meitinger et al., 2020). If the conceptual understandings of warmth and competence between the target groups were sufficiently similar, it would also be necessary to ascertain equal measurement properties between the target groups to rule out any systematic measurement bias or differential item functioning (Boer et al., 2018). Such comparable measurement properties can be assessed quantitatively through measurement invariance, which tests the equality of certain measurement parameters (e.g., observed item-latent factor relations, item thresholds, item error variances) between groups (Davidov et al., 2014). The amount and specification of equal measurement properties depends on the intended analysis: For unbiased mean value comparisons, which are the most prominent application of the SCM, equality of factor loadings and indicator intercepts of identical items across target groups is required.

To the best of our knowledge, such comprehensive empirical evidence supporting the proposed universality of the SCM has not been presented, neither within nor between research contexts. Most studies that compared samples from different cultural contexts only examined the scales' reliability (e.g., Cuddy et al., 2009; Durante et al., 2013, 2017), and the same applies to those studies comparing different target groups in one sample or cultural context (e.g., Asbrock, 2010; Eckes, 2002; Fiske et al., 2002). Only few studies presented an examination of equal conceptualisations and measurement properties across samples (e.g., Grigoryan et al., 2020; Halkias & Diamantopoulos, 2020) or across target groups within samples (e.g., Janssens et al., 2015; Kotzur et al., 2020; Stanciu, 2015). Their results showed mixed findings, ranging from very equal conceptualisations (e.g., Grigoryan et al., 2020; Halkias & Diamantopoulos, 2020; Stanciu, 2015) to partially (e.g., Kotzur et al., 2020) or completely differing conceptualisations (e.g., Janssens et al., 2015, Study 1). These results give no comprehensive answer to the question whether the SCM can really be considered universal from a measurement-theoretical and intercultural psychological perspective. We also point out that the above-mentioned questionable measurement practises, such as the insufficient scale development and especially the frequently changing warmth and competence measurements, amplify the issue of questionable universality. Thus, we call for a systematic examination of the SCM's measurement properties in order to empirically test and evaluate the assumed universality of warmth and competence.

Shortcoming 5: Limited Application for Some Categories of Target Groups

The applications of the SCM range very broadly with regard to the examined target groups, and include, beyond many others, the description of societally relevant target groups along warmth and competence dimensions (e.g., Asbrock, 2010; Cuddy et al., 2009; Durante et al., 2013, 2017) and the investigation of the effects different labels have for the perception of the same target group (Binggeli et al., 2014b; Eckes, 2002; Fröhlich & Schulte, 2019; Kotzur et al., 2017; Lee & Fiske, 2006). We identified, particularly in the German context, some areas in which the SCM could be used productively to gain scientific insights regarding the social perception of relevant target groups, but where it has not or only infrequently been put into practise.

One area is the social perception of refugees: The global numbers of forced migrants and refugees has been rising continuously over the last years, reaching an all-time high in 2019 (UNHCR, 2020). This development was also mirrored in Germany, which experienced the so-called "migration crisis" in 2015/16. In 2016, an all-time high of more than 745,000 applications for asylum was registered in Germany (Statista, 2021). Since that time, migrants in general and refugees in particular have become highly relevant and prominent social groups in Germany (Kotzur et al., 2021). In different contexts, discourses arose concerning for instance refugees' job- and labour-market integration, crime rate, or the increased influence of non-Christian religions in Germany (Infratest, 2016). Thus, social perception research is of high relevance, as it may assist in predicting the reactions of the receiving societies, facilitating acculturation processes and designing interventions to counteract negative stereotypes and prejudice. Interestingly, many discourses were accompanied by tendencies of subtyping, for instance when newspaper articles debate the different levels of refugee criminality depending on their country of origin (Hackensberger et al., 2016). The SCM has been applied previously to assess the social perception of subgroups of migrants in the US (Lee & Fiske, 2006), and recently also in Germany (Fröhlich & Schulte, 2019, Veit & Yemane, 2020). However, only little is known about the social perceptions of (specific subgroups of) refugees in Germany (but see Bansak et al., 2016; Ditlmann et al., 2016; Kotzur et al., 2017; Wyszynski et al., 2020). Consequently, we identify research on the social perception of refugee subgroups (which is also mindful of the critiques addressed above) as highly relevant and lacking.

Another area with little SCM applications is occupational groups. It has been found that occupation plays a vital role in individuals' lives because it defines an essential part of the self-conception and societal standing, it usually contributed to ensuring individuals' livelihood, and most people spend a considerable amount of their alert time pursuing their occupations (Crößmann & Günther, 2018). SCM research with the purpose of generally assessing the social perception of societally relevant groups has usually contained some occupational groups (such as white- and blue-collar workers, physicians, athletes; Asbrock, 2010). Thus, the social perception of some occupational groups has been replicated in multiple contexts and over a larger time span (e.g., workers; Cuddy et al., 2009; Durante et al., 2013; Janssens et al., 2015; The Fiske lab, n.d.; Stanciu, 2015). But research exclusively assessing the stereotype content of occupational groups is rare in Germany (but see Ihme & Möller, 2015; Imhoff et al., 2013; Lotzkat & Welpe, 2015) and was only recently published in the US context (He et al., 2019). However, even these recent findings might be outdated given the emergence of the COVID-19 pandemic, which triggered substantial debates about the societal and systemic relevance of different occupations in times of crisis (e.g., DGB Niedersachsen, 2020). Consequently, we perceive the necessity to further investigate the social perception of occupational groups in Germany. Moreover, for both the refugee subgroups and the occupational groups, we propose that the conducted research should be mindful of the above-mentioned methodological shortcomings.

II. The Present Research

The dissertation at hand has addressed each of the five shortcomings above in order to contribute to a more reliable and valid SCM research producing more robust, meaningful and generalisable findings. This goal has been pursued in four article manuscripts which applied two principal strategies (for an overview, see Table 1). For all manuscripts, we adhered strongly to the principles of open and reproducible science (APA Psychological Science Agenda, 2019; Sullivan et al., 2019), as we will outline below.

Table 1

Overview of the Research Aims, Study Designs, Statistical and Analytical Procedures and the Shortcomings in SCM Research Addressed by the Four Manuscripts Included in this Dissertation Project

Manu- script	Reference	Research Aim	Study Design	Statistical and Analytical Procedures	Addressed Shortcomings
# 1	Friehs, M.-T., Böttcher, J., Kotzur, P. F., Lüttmer, T., Wagner, U., Asbrock, F., & van Zalk, M. H. W. (2021). *Examining the structural validity of stereotype content measures – A preregistered re-analysis of published data and discussion of possible future directions*. Manuscript under review at the Personality and Social Psychology Bulletin.	Examination of the structural validity of English SCM measures	Re-analysis of 43 published SCM publications containing 78 datasets (*N* = 20,819) using English-language multi-item measures of warmth and competence and assessing at least two different target groups or samples	Examination of dimensionality using confirmatory factor analysis; Subsequent examination of measurement invariance up to (partial) scalar level using multiple-group confirmatory factor analysis	(2) Insufficient scale performance assessment and varying scales in later applications; (3) Lack of advanced analytical approaches; (4) Insufficiently examined assumption of universality
# 2	Friehs, M.-T.*, Kotzur, P. F.*, Zöller, A.-K. C., Wagner, U., & Asbrock, F. (2021). *A preregistered examination of scale properties of stereotype content measures: The German case.* Manuscript under review at the International Review of Social Psychology.	Replication of Manuscript # 1 in a different country and language context by examining the structural validity of German SCM measures	Re-analysis of 23 published SCM publications containing 29 datasets (*N* = 10,854) using German-language multi-item measures of warmth and competence measured in German samples and assessing at least two different target groups or samples	Examination of dimensionality using confirmatory factor analysis; Subsequent examination of measurement invariance up to (partial) scalar level using multiple-group confirmatory factor analysis	(2) Insufficient scale performance assessment and varying scales in later applications; (3) Lack of advanced analytical approaches; (4) Insufficiently examined assumption of universality

Table 1 (continued)

Manu- script	Reference	Research Aim	Study Design	Statistical and Analytical Procedures	Addressed Shortcomings
# 3	Kotzur, P. F.*, Friehs, M.-T.*, Asbrock, F., & van Zalk, M. H. (2019). Stereotype content of refugee subgroups in Germany. *European Journal of Social Psychology, 49*(7), 1344-1358. https://doi.org/10.1002/ejsp.2585	Assessment of the stereotype content of 16 refugee subgroups in Germany using latent (i.e., reliability-corrected) mean values	Cross-sectional online-survey (*N* = 264) using a German student sample and assessing the SCM with the scale proposed by Asbrock (2010)	Examination of scale dimensionality using confirmatory factor analysis; Establishment of partial metric and scalar measurement invariance and comparison of latent means using the alignment optimization procedure	(2) Insufficient scale performance assessment; (3) Lack of advanced analytical approaches; (5) Limited application for some categories of target groups
# 4	Friehs, M. T., Aparicio Lukassowitz, F., & Wagner, U. (2020). *Stereotype content of occupational groups in Germany.* Manuscript under review at the International Review of Social Psychology.	Development of a German SCM measure to assess occupational stereotypes with acceptable measurement properties; Assessment of the stereotype content of 13 occupational groups in Germany using latent (i.e., reliability-corrected) mean values	Cross-sectional online-survey (*N* = 425) using a heterogeneous German sample and assessing the SCM with the newly-developed scale	Examination of scale properties using exploratory and confirmatory factor analysis and reliability assessment; Establishment of partial metric and scalar measurement invariance and comparison of latent means using the alignment optimization procedure	(1) Insufficient initial scale development; (2) Insufficient scale performance assessment; (3) Lack of advanced analytical approaches; (5) Limited application for some categories of target groups

Notes. * The authors share first authorship. SCM = Stereotype content model. *N* = Sample Size.

As a first strategy, we used published data to systematically re-examine the measurement properties of previously published SCM literature in different language contexts in a cumulative science approach. Data re-analysis is a rarely used, but powerful scientific tool which allows to independently replicate previous findings and generate new cumulative results in an economical manner, thus increasing the transparency and accountability of reported research results (Hargreaves & Davey, 2015). Our data re-analysis resulted in two manuscripts: For Manuscript #1 (Friehs, Böttcher et al., 2021), we identified and gained access to published SCM data using English measurements in English-speaking samples. We re-analysed 78 datasets from 43 publications (total *N* = 20,819) in line with the original research questions and designs applying confirmatory factor analyses on the warmth and competence measures used to assess the different target groups. If feasible, we then conducted measurement invariance assessments using multiple-group confirmatory factor analyses to examine whether warmth and competence scores of different target groups could be meaningfully and validly compared as intended in the original study. Additionally, we reported the scales' internal consistency and evaluated the overall item and scale performance to comprehensively inform readers about the psychometric properties of previously published English SCM scales and to thus assist future SCM scale development attempts. The manuscript concludes with a number of directions and good practise suggestions for improving the structural validity of future SCM research, addressing beyond others the aspects of sample sizes, analytical strategies, and the specification of the SCM's universality claim. Consequently, Manuscript # 1 addresses the above-mentioned shortcomings of (2) insufficient scale performance assessment, (3) the lack of application of advanced analytical strategies, and (4) the examination of the universality assumption. Additionally, in line with open science good practise recommendations (e.g., APA Psychological Science Agenda, 2019, Sullivan et al., 2019), we pre-registered the study plan and analytical procedure of Manuscript #1 on the Open Science Framework and will provide open code when the manuscript is accepted for publication. Additionally, we have published a preprint of the current version of the manuscript on PsyArxiv.

To exclude the possibility of singular language- or context-specific findings (Sechrest et al., 1972), we conducted a second data re-analysis replicating the approach of Manuscript # 1 with SCM data collected in 29 datasets from 23 published studies using German-language warmth and competence measures in German samples (total *N* = 10,854). We presented the results of confirmatory factor analyses, measurement invariance assessment and internal consistency examination in Manuscript # 2 (Friehs, Kotzur et al., 2021). Thus, Manuscript # 2 provides an independent confirmation and replication of our findings of Manuscript # 1 while addressing the same shortcomings of SCM literature. Manuscript # 2 also adheres to open science good practise recommendations (APA Psychological Science Agenda, 2019; Sullivan et al., 2019) by presenting a pre-registered study plan and analysis procedure, providing open code once the final version of manuscript is accepted, publishing a preprint on PsyArxiv, and being submitted for publication to an open-access journal.

Our second strategy consisted in adjusting the traditional analytical approach of SCM data to integrate more internally valid and reliable advanced analysis strategies while at the same time collecting new SCM data on societally relevant target groups. In Manuscript # 3 (Kotzur, Friehs et al., 2019), we examined the applicability of a widely-used German SCM scale (Asbrock, 2010) to assess the social perception of 16 refugee subgroups which were differentiated along the dimensions flight motive, religious affiliation and country of origin. We collected data in a student sample (*N* = 264) and applied confirmatory factor analysis, which resulted in the exclusion of a substantial number of refugee subgroups for which we could not establish acceptable warmth and competence scales. Afterwards, we compared the remaining refugee subgroups' warmth and competence perceptions using latent (i.e., reliability-corrected) mean values while at the same time establishing the required partial metric and partial scalar measurement invariance using the alignment optimisation procedure (Asparouhov & Muthén, 2014). Manuscript # 3 gives important indications concerning the social perception of different refugee subgroups in Germany and presents a novel and statistically advanced approach to analysing SCM data. Thus, it addresses the shortcomings of (2) insufficient scale performance assessment, (3) the lack of application of advanced analytical strategies, and (5) limited research on some categories of target groups described above. In line with open science good practise recommendations (APA Psychological Science Agenda, 2019; Sullivan et al., 2019), we provide open code and open materials for Manuscript # 3 in the Open Science Framework. Additionally, we have published Manuscript # 3 with open access.

Manuscript # 4 (Friehs et al., 2020) transferred the approach of Manuscript # 3 to assess the social perception of different occupational groups in a heterogeneous German sample (*N* = 425). In order to avoid repeated data exclusion due to unacceptably performing warmth and competence scales (see Manuscript # 3), Manuscript # 4 developed a new German SCM scale specifically adjusted to the context of occupational stereotypes. The new warmth and competence measures' performance was assessed using explorative and confirmatory factor analysis as well as internal consistency examinations. The scale allowed for the comparison of all 13 investigated occupational groups along latent warmth and competence scales using the alignment optimisation procedure. Consequently, Manuscript # 4 addressed the above-described shortcomings of (1) insufficient scale development, (2) insufficient scale performance assessment, (3) the lack of application of advanced analytical strategies, and (5) limited research on some categories of target groups. Adhering to the open science good practise recommendations (APA Psychological Science Agenda, 2019; Sullivan et al., 2019), we have pre-registered the study plan and data analysis approach at the Open Science Framework, we have submitted Manuscript # 4 to an open-access journal and will provide open data, open code and open material when the manuscript is accepted for publication.

Manuscript # 1: Examining the structural validity of stereotype content measures – A preregistered re-analysis of published data and discussion of possible future directions

Friehs, M.-T., Böttcher, J., Kotzur, P. F., Lüttmer, T., Wagner, U., Asbrock, F., & van Zalk, M. H. W. (2021). *Examining the structural validity of stereotype content measures – A preregistered re-analysis of published data and discussion of possible future directions.* Manuscript under review.

Submitted on May 17, 2021 to the Personality and Social Psychology Bulletin

Contributorship according to the CRediT system:

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

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

PFK: Conceptualization, Methodology, Project administration, Supervision, 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

Open Science Practises:

Preregistration created on March 16, 2020, see https://osf.io/gqmvz/?view_only=9a8fc0053b634ace8ea8941b6c9423b7

Open analysis code will be provided upon acceptance of the final manuscript version

Online supplementary materials are provided on the Open Science Framework, see https://osf.io/srh36/

A preprint of the manuscript was published on PsyArxiv, see https://psyarxiv.com/dej4m/

Manuscript # 2: A preregistered examination of scale properties of stereotype content measures: The German case

Friehs, M.-T.*, Kotzur, P. F.*, Zöller, A.-K. C., Wagner, U., & Asbrock, F. (2021). *A preregistered examination of scale properties of stereotype content measures: The German case.* Manuscript under review.

* Shared first authorship

Submitted on May 17, 2021 to the International Review of Social Psychology

Contributorship according to the CRediT system:

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

PFK: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing – Original draft

AKCZ: Data curation, Formal analysis, Methodology, Writing – Original draft

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

FA: Conceptualization, Resources, Writing – Review & editing

Open Science Practises:

Preregistration created on May 21, 2019, see https://osf.io/486h7/?view_only=e1b25da1084f4e248a621be36b31a153

Open analysis code will be provided upon acceptance of the final manuscript version

Online supplementary materials are provided on the Open Science Framework, see https://osf.io/jqzet/

A preprint of the manuscript was published on PsyArxiv, see https://psyarxiv.com/fa39w/

Manuscript # 3: Stereotype content of refugee subgroups in Germany

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

* Shared first authorship

Submitted on August 30, 2018 to the European Journal of Social Psychology

Accepted on March 11, 2019

Published online on October 29, 2019

Contributorship according to the CRediT system:

PFK: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Validation, Writing – Original draft

MTF: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – Original draft

FA: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – Review & editing

MHWVZ: Resources, Writing – Review & editing

Open Science Practises:

Open analysis code and Open materials provided on the Open Science Framework, see https://osf.io/5j7t6/

Published in an open-access issue of the European Journal of Social Psychology

Manuscript # 4: Stereotype content of occupational groups in Germany

Friehs, M. T., Aparicio Lukassowitz, F., & Wagner, U. (2020). *Stereotype content of occupational groups in Germany.* Manuscript under review.

Submitted on September 24, 2020 to the International Review of Social Psychology

Contributorship according to the CRediT system:

MTF: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Methodology, Project administration, Resources, Supervision, Visualization, Writing – Original draft

FAL: Conceptualization, Data curation, Investigation, Validation, Writing – Review & editing

UW: Conceptualization, Supervision, Writing – Review & editing

Open Science Practises:

Preregistration created on April 27, 2020, see https://osf.io/pmjgf/?view_only=353ea72e07fc4cb7b33b0beba8fe4842

Open data and open materials provided on the Open Science Framework, see https://osf.io/gxz49/

Open analysis code will be provided upon acceptance of the final manuscript version on the Open Science Framework

III. General Discussion

The Stereotype Content Model is one of the most prominent models of social perception (Abele et al., 2021). In light of expressed methodological concerns and shortcomings with regard to the model's applications, the present dissertation project aimed at contributing to a more reliable, valid and meaningful SCM research practise. We did so by describing the extent of methodological issues concerning the dimensionality and comparability of SCM measures between target groups in published research (Manuscripts # 1 and # 2) and by presenting alternative strategies and well-performing measures to answer SCM-typical research questions (Manuscripts # 3 and # 4). At the same time, we adhered strongly to open science good practise recommendations (e.g., APA Psychological Science Agenda, 2019; Sullivan et al., 2019) to increase our findings' transparency and reproducibility. In the following, we will first present an overarching discussion of our findings across manuscripts before addressing potential limitations of our approach as well as an outlook for future SCM research.

Summary and Implications of Results Regarding the Systematic Re-Analysis of Measurement Properties in Published SCM Literature (Manuscripts # 1 and # 2)

Prior SCM research has most often either failed to provide strong confirmatory evidence regarding the used scales' dimensionality and other psychometric properties (Halkias & Diamantopoulos, 2020) or has presented ambivalent evidence on that matter (e.g., Janssens et al., 2015; Kotzur et al., 2020; Stanciu, 2015). Consequently, we conducted systematic re-analyses of published SCM data in English (Manuscript #1) and German (Manuscript # 2). We investigated the warmth and competence measures' dimensionality, internal consistency and measurement invariance as precondition for meaningful (latent) mean-value comparisons (both manuscripts) and presented further item and scale performance parameters (only Manuscript # 1). In total, our cumulative re-analyses included more than 60 published articles with over 100 datasets and more than 31,000 participants. Manuscript # 1 additionally proposes methodological and statistical ways to strengthen the named measurement properties in future SCM research.

Manuscript # 1, focussing on English-language SCM measures from different contexts, and Manuscript # 2, using German-language SCM measures exclusively collected in German samples, show strongly similar results. In both manuscripts, in about 65% of all cases, the expected two-dimensional structure of the warmth and competence measures could not be confirmed. This finding indicates that, in the majority of cases, the used SCM measures did not allow to be validly summarised into a two-dimensional warmth and competence scales. From a construct validity perspective, it thus remains unclear which construct was measured exactly under the labels of warmth and competence and how concurrent these scales were with the underlying theoretical assumptions (Flake et al., 2017). This finding evidently biases all attempts for a valid interpretation of results, because "if the 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 into the primary results" (Flake et al., 2017, p. 370; see also Flake & Fried, 2020). Moreover, in both manuscripts, only about 11% of all target groups showed the measurement properties required for meaningful (latent) mean-value comparisons (i.e., partial or full scalar measurement invariance; Davidov et al., 2014). These measurement properties assure that similar conceptual definitions of warmth and competence were used between target groups with comparable item difficulties, thus increasing the construct validity of findings and additionally introducing one possible interpretation of universality with regard to measurement properties to the SCM (van der Vijver & Poortinga, 1982). Consequently, given that the vast majority of re-analysed cases did not present such measurement properties, the assumption of the SCM's universality (e.g., Cuddy et al., 2008, 2009; Fiske, 2017; Fiske & North, 2015; Fiske et al., 2007), at least with respect to its measurement properties, is not tenable on the basis of our results.

Overall, our findings indicate a large-scale problem concerning the operationalisation and measurement of the SCM as well as its assumption of universality. It seems that the published and applied SCM scales do not validly measure warmth and competence and that mean-value comparisons using these scales are often biased due to incompatible measurement properties or the different functioning of the target groups' scales. These issues do not seem to be limited to one research context, as we provided evidence from two languages and a multitude of national contexts as well as a huge variety of target groups and scale indicators. Nonetheless, we also propose ways to avoid these hidden validity issues in future SCM research. These options include the thorough development of new SCM scales, the standard application of confirmatory factor analysis and measurement invariance assessment (if required), the usage of larger sample sizes and advanced methodological approaches to answer research questions. We expect that these strategies may increase the meaningfulness, reliability and validity of the findings of SCM studies (which we also demonstrate in Manuscripts # 3 and # 4).

Thus, Manuscripts # 1 and # 2 effectively and comprehensively addressed the critiques regarding the insufficiently tested and varying measurements used in SCM research (shortcoming 2) and regarding the insufficiently examined assumption of universality (shortcoming 4) using advanced confirmatory analytical approaches (shortcoming 3). As an additional strength, both manuscripts did so by using a pre-registered procedure and presenting open code, which increases the transparency and reproducibility of the findings and excludes the possibility of adaptation of methodological and statistical procedures based on desired results (APA Psychological Science Agenda, 2019).

Summary and Implications of Results Regarding Innovative Analytical Approaches and Improved SCM Measurements (Manuscripts # 3 and # 4)

This dissertation project also aimed to demonstrate how SCM data can be collected and analysed without falling prey to the construct validity concerns mentioned above by using advanced statistical methodology. To this end, Manuscript # 3 used an established and frequently used (but in hindsight insufficiently validated) German SCM scale (Asbrock, 2010) to assess the social perception of refugee subgroups, a social category of high societal relevance during the so-called “migration crisis” that was not systematically investigated in former research. From a methodological perspective, we established the scale’s dimensionality and compared the evaluation of the different refugee subgroups using the alignment optimisation approach to generate latent (i.e., reliability-corrected) warmth and competence scores. Regarding their content, the results showed large effects of the additional information regarding the refugees’ religious affiliation, geographic origin and flight motive on the respective latent warmth and competence ratings. These were largely in line with at-the-time societal discourses, the social evaluation presented in the media, and previous research findings.

Building on the experiences we made with Manuscript # 3, Manuscript # 4 presented a new German SCM measure especially adapted to the assessment of the social perception of occupational groups. It also replicated the successful application of the alignment optimisation procedure to compare reliability-corrected warmth and competence means of different occupational groups, which are another social category with limited previous consideration in SCM research and which were subject of intense societal debate during the initial phase of the COVID-19 pandemic in spring 2020. Again, our findings regarding the latent warmth and competence ratings of different occupational groups largely confirmed our expectations. Thus, both manuscripts described the social perception of relevant societal target groups about whom there existed limited previous knowledge, thereby responding to shortcoming 5.

Manuscript # 3 impactfully demonstrated the limitations and challenges of using insufficiently validated SCM scales with uncertain measurement properties in practical research: Though we aimed at analysing the social perception of 16 target groups, the lack of dimensionality of the used SCM scale forced us to exclude a substantial portion of the data (i.e., six targ groups) from the main analysis. Consequently, we were strongly limited in our possibilities to draw meaningful conclusions from the data. Thus, Manuscript # 3 serves as an empirical display of the contemplations on the practical consequences of unknown psychometric SCM scale properties presented both in Manuscript # 1 as well as in shortcoming 2. This limitation was addressed in Manuscript # 4, which presented a new, context-adapted and high-performing SCM scale constructed on the base of state-of-the-art scale development procedures (Bandalos, 2018; Brown, 2015). This scale was developed using a rigorous approach based on internal consistency, exploratory and confirmatory factor analyses and showed favourable psychometric properties with regard to the dimensionality and measurement invariance. Thus, Manuscript # 4 presents a good practise response to the critiques concerning the initial SCM scale development process outlined in shortcoming 1 (for a further example, see Halkias & Diamamtopoulos, 2020).

Both manuscripts additionally used the alignment optimisation approach (Asparouhov & Muthén, 2014), a relatively recent statistical approach to compare latent mean values between target groups while simultaneously generating a data-driven measurement invariance pattern. Establishing measurement invariance using this procedure is swifter and less cumbersome than the multiple-group confirmatory factor analyses we applied in Manuscripts # 1 and # 2. Moreover, it is less dependent on individual decisions of researchers within the process (e.g., concerning the introduction of partial measurement invariance), thus creating more objective and reproducible solutions. In both manuscripts, this approach generated partial metric and partial scalar measurement invariant results with low levels of measurement non-invariance. Although this analytical approach has never been used before in SCM research (but see Seddig et al., 2020, for a related application to general stereotypes), the generated results were mostly in accordance with our expectations and in line with previous research results. Thus, we conclude that using advanced second-generation statistical approaches (thus addressing the critique expressed in shortcoming 3) did not produce undue deviations or methods bias in our results. We assume that our findings using alignment optimisation are comparable with the results that would have been generated with the first-generation statistical approaches often applied in published SCM research, such as ANOVA or cluster analysis. Thus, we recommend the usage for alignment optimisation for future SCM research. In summary, Manuscript # 3 and # 4 explored novel ways to generate reliable, meaningful and valid knowledge on social perception of different relevant target groups by using advanced methods of latent mean value comparison. At the same time, Manuscript # 3 showed that research insights with regards to content might be severely limited by the detrimental effects of insufficiently validated and inadequately performing SCM scales. This disadvantage was successfully addressed in Manuscript # 4, which demonstrated how novel SCM scales could be constructed and applied.

Limitations and Future Research

The successful implementation and comprehensive approach of this dissertation project notwithstanding, our approach is limited in various aspects, which will be elaborated in the following. These limitations offer pathways for future and alternative research approaches, which we will outline simultaneously.

Alternative Methodological Approaches

There is a variety of methods that have been used previously and which might be used in future to analyse SCM data, each with its different strengths and weaknesses. The methodological approaches we used in our manuscripts to investigate the psychometric properties and measurement-based universality of the SCM dimensions as well as to compare latent warmth and competence means across target groups were just one of many promising alternatives: Firstly, the measurement property analyses in Manuscripts # 1 and # 2 could also have been conducted and could be extended using other methods, for instance item response models and differential item functioning (Penfield & Camilli, 2007; Wetzel & Roberts, 2020). These methods differ somewhat conceptually and concerning the results they produce compared to the applied (multiple-group) confirmatory factor analysis approaches (Tay et al., 2015). We chose a confirmatory factor-analytical approach because (partial) measurement invariant solutions can easily be translated into further advanced analysis procedures such as structural equation modelling. Having established the knowledge about the SCM's measurement properties, item response theory-based approaches could provide us with additional valuable information about the effect size of measurement non-invariance in future research. What is more, the analysis of nomological networks might inform us in what ways the measurements might be understood differently by the participants (Thielmann & Hilbig, 2019; Welzel et al., 2021). We assume that our overall results are to some extent methods-dependent, as was also demonstrated with the data from Manuscript # 3, which were also re-analysed in Manuscript # 2. Compared to the alignment optimisation approach applied in Manuscript # 3, the multiple-group confirmatory factor analysis in Manuscript # 2 yielded less favourable results, as two more refugee subgroups needed to be discarded from measurement invariance analysis (for similar patterns, see e.g., Magraw-Mickelson et al., 2021; Seddig et al., 2020). Equally, the (mostly fairly recent) approaches based on item response theory or confirmatory factor analysis approaches using Bayesian estimation (i.e., approximate measurement invariance, Muthén & Asparouhov, 2013) might yield different results (Wetzel & Roberts, 2020). Consequently, we encourage future research to apply and compare different approaches when analysing the measurement properties and comparability of SCM scales.

Secondly, the criteria we applied to establish measurement invariance in Manuscripts # 1 and # 2, which are the interplay of acceptable global model fit according to the criteria of Schermelleh-Engel and colleagues (2008), a non-significant Satorra-Bentler chi-square difference test (Satorra & Bentler, 2001) and relative changes in model fit in accordance with the criteria defined by Chen (2007), were very strict and lead to a high exclusion rate of target groups in the measurement invariance assessments. This was mostly due to the Chen (2007) criteria, which were often not met even though absolute model fit and changes in chi-square value were acceptable. This issue is exacerbated by the fact that unlike in *p*-value-based hypothesis testing, there exist no well-defined conventions in structural equation modelling regarding which model fit indicators should be used with which cut-off criteria and how they should be weighed against each other. Thus, other researchers might choose different criteria to define acceptable SCM measurement models and measurement invariance and would consequently find different results. Future applied SCM research that wishes to establish dimensionality and measurement invariance as a precondition of valid further analyses might consider other model fit criteria and alternative approaches within the multiple-group confirmatory factor analysis framework to assess measurement invariance. One instance is the top-down approach to assess measurement invariance, which just aims at establishing a model with certain equality restrictions according to the research question without considering the model fit change compared to less restricted models (e.g., Horn & McArdle, 1992) instead of the bottom-up approach we used in Manuscripts # 1 and # 2. Such a top-down-approach poses less requirements to establishing the preconditions for valid (latent) mean value comparison and might thus be easier and more convenient to implement, and are therefore recommendable especially for more application-oriented research.

Thirdly, we limited our analyses of measurement-based universality exclusively to quantitative approaches. Qualitative approaches such as cognitive interviewing or online probing (Achbari & Davidov, 2019; Benítez & Padilla, 2014; Latcheva, 2011; Meitinger, 2017) might deliver valuable additional and complementary information in future research investigating the causes and underlying mechanisms of measurement non-invariance between specific target groups. Such information might be equally useful when constructing a new SCM measurement, as they deliver quite different information about the participants' understanding of the fundamental dimensions of social perception and their indicators.

Lastly, we wish to point out that alignment optimisation is only one of multiple options to analyse SCM data using advanced second-generation statistical methods. These different methods depend highly on the underlying research questions: For latent mean value comparisons, multiple-group confirmatory factor analysis would have provided a (more cumbersome) alternative to alignment optimisation. Results from alignment optimisation cannot easily be transferred to other analyses, for instance in a structural equation framework. Consequently, in such cases, multiple-group confirmatory factor analysis should be preferred. Alignment optimisation is equally unsuitable if researchers wish to generate the typical clustering which is often used in SCM research (e.g., Asbrock, 2010; Eckes, 2002; Cuddy et al., 2009; Fiske et al., 2002). Nonetheless, second-generation methodological alternatives to cluster-analysis exist, such as latent class and latent profile analysis (Berlin et al., 2013; Lubke & Muthén, 2005). However, these require a very high number of target groups to be included in the analysis, which complicates SCM data collections due to the high strain on participants and/or the high number of participants in case of split-sample approaches.

Open Questions Concerning the Causes for Heterogeneity in Measurement Dimensionality and Comparability

This dissertation project aimed at describing the extent of structural validity in SCM measures by examining the measurement properties and comparability between target groups. However, it did not extend to investigating the underlying causes for the mixed results we found. Therefore, we cannot yet answer questions such as why about two thirds of all re-analysed target groups did not show an acceptable two-dimensional warmth and competence factorial structure, what determines whether a selection of target groups becomes measurement invariant or what it means conceptually if the warmth and competence constructs are not equivalent between target groups. All these questions require additive and systematic research which might use some of the methods outlined above. At this point, we can only outline some tentative potential directions of inquiry.

One such direction might be the question whether the SCM measurements' dimensionality and comparability is affected by the specific target group or sample under scrutiny. For one, the participants‘ understanding of warmth and competence for the same target group might differ between contexts. For instance, participants rating the target group *Refugees* on the competence dimension might think about them in the context of their flight associated with a number of hardships and obstacles to overcome. Or participants might think about refugees trying to integrate into the labour market of the receiving country. These different contexts might strongly affect the average competence ratings. What is more, one might assume that evaluating warmth and competence is only a feasible task if the participants are sufficiently familiar with the assessed target group. Whenever a target group is unknown to the survey participants or they have limited knowledge and experiences with regard to that target group, the task of filling in different warmth and competence indicators might be overly difficult. In such cases, participants might rather rely on simpler good-or-bad heuristics or sources of information irrelevant to the measure at hand. The SCM-specific customs of having an independent sample nominating societally relevant (and therefore relatively familiar) target groups and asking participants to rate the indicators from the societies' rather than the individuals' perspective (Cuddy et al., 2008, 2009; Fiske et al., 2002) might compensate such unfamiliarity only to a limited extent. Relatedly, the level of involvement and identification of the participants with the target groups might affect the results based on ingroup preference or other intergroup dynamics. One such instance might be when a student is asked to evaluate the social group *Students* (e.g., Asbrock, 2010). Also, temporal influences might additionally affect the SCM measurements due to the current salience or prominence of a specific target group in public discourses. Thus, is might be a fruitful route to explore the target groups' context, level of familiarity, identification and perceived current relevance as potential explanatory variables for the differences in the SCM's psychometric properties.

Another potential line of inquiry is methods bias potentially rooted in different styles of presenting SCM indicators in surveys. In principle, the SCM measures could be presented either in a comprehensive indicator x target group matrix, one indicator at a time which is rated with relation to all target groups, or one target group at a time which is rated on all warmth and competence indicators simultaneously. These different strategies might bias the results due the inherently comparative and relative nature of SCM findings (i.e., in most cases, a target group's evaluation as high in warmth and high in competence is not based on absolute scale values but rather on the target group's position in comparison to the other evaluated groups). For such comparisons, participants need to have an overview which target groups they will evaluate on which indicators to generate meaningful responses. Following this logic, also the number of target groups that are being rated (i.e., single-group rating versus the comparison of multiple target groups), the order in which target groups and/or indicators are presented, the specific target groups' labels and the composition of target groups within one survey as well as the individual or societal perspective of evaluation might influence how the participants perceive and process the survey questions (e.g., Binggeli et al., 2014b; Eckes, 2002; Lee & Fiske, 2006; Kotzur et al., 2017, 2020). This could potentially affect the measurement properties of the resulting SCM scales; therefore, investigating these potential methodological biases is highly relevant to support future researchers in conducting meaningful, reliable and valid SCM research.

Impact of the Findings on The Theoretical Framework

As for now, it remains somewhat open what our empirical results concerning the operationalisation and measurement of warmth and competence imply for the theoretical underpinning of the stereotype content model. This dissertation project has demonstrated comprehensively that the current measurement and analysis approaches in SCM research are prone to bias due to structural validity threats. This bias causes subsequent issues concerning the valid interpretation of a study's finding (Flake et al., 2017, 2020), and our results indicate that the theoretical assumption of universality (if it can indeed be understood as a comparable conceptual definition of warmth and competence and equivalent measurement properties) is empirically not supported. But do these findings mean that we have falsified the Stereotype Content Model as a theoretical framework?

For various reasons, this is not the conclusion we draw from this dissertation project. One reason is that the SCM's substantial validity, meaning its embedding in research literature and alignment with previous research findings (Flake et al., 2017), is quite extensive and comprehensive with regard to the general existence of two dimensions of social perception and evaluation (Cuddy et al., 2008; Fiske et al., 1999, 2002). If we generalise to *agency* and *communion* as more general dimensions of social evaluation, the theoretical and empirical evidence is truly overwhelming and stemming from a multitude of different methodological approaches (Abele & Wojciszke, 2019; Abele et al., 2021). Another reason is that this dissertation project focused on only one model out of the broad spectrum of theories in social perception research, so it remains unclear if the methodological and measurement issues we outlined for the SCM extend to other theoretical frameworks, like for example the agency-and-communion framework (Abele & Wojciszke, 2007, 2014, 2019; Abele et al., 2021). Though the SCM is the only theoretical model with such a prominent claim of universality, the requirements relating to measurements' dimensionality and comparability apply also to the other theoretical frameworks, as all of them habitually compare the social perception of different targets (which might be individuals or groups). Insufficient measurement properties have been reported from many areas of social and personality psychology (Hussey & Hughes, 2020), so we might suspect that similar deficits would be identified in systematic re-analysis of data from other theoretical models. Only future research can investigate if our findings are singular to the SCM or whether they extend to the broader social perception literature.

Nonetheless, in this dissertation project, we identified many starting points for the further theoretical development of the Stereotype Content Model and improvements of the wider social perception research. Some of these starting points are exemplified in the following questions: How can the universality of the SCM be formally defined and how can it be empirically tested? Or alternatively, do we need to restrict the area of application of the SCM to certain geographic or language regions or particular target groups? Should we really expect that the warmth and competence dimensions apply equivalently in concept and understanding to all investigated target groups (as we implied with the assessment of measurement invariance), or might it be wiser to assess these dimensions using more selected and context-specific measurements (e.g., Halkias & Diamantopoulos, 2020, developed an SCM scale for the usage in the marketing context)? How does the lack of comparability of SCM measurements between target groups influence the content-wise findings of a study? For reasons of scope, we could not answer these questions in the context of this dissertation project, but we have presented a number of promising directions and strategies for future research. What is more, we believe that the response to the issues we outlined can only be a collective one. Well-known social perception researchers have only recently illustrated how to design processes of integrating contradicting or even adversarial positions and findings to engage in cooperative theory building and cumulative science (Ellemers, 2021; Ellemers et al., 2020). Therefore, with the findings of this dissertation project, we hope to stimulate animated, diverse, respectful and fruitful discussions striving to collaboratively and constructively revise and improve research on the fundamental dimensions of social perception.

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. https://doi.org/10.1037/rev0000262

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. https://doi.org/10.1037/0022-3514.93.5.751

Abele, A. E., & Wojciszke, B. (2014). Communal and agentic content. A dual perspective model. *Advances in Experimental Social Psychology*, *50*, 195-255. https://doi.org/10.1016/B978-0-12-800284-1.00004-7

Abele, A. E., & Wojciszke, B. (Eds.). (2019). *The Agency – Communion framework*. Oxford, UK: Routledge.

Achbari, W., & Davidov, E. (2019, July). *Re-assessing the radius of generalized trust: Measurement invariance, think aloud protocols, and the role of education* [Conference presentation]. Conference of the European Survey Research Association (ESRA), Zagreb, Croatia.

Allport, G. (1954). *The nature of prejudice*. Reading, MA: Addison-Wesley.

APA Psychological Science Agenda (2019). *An introduction to open science – How to incorporate best practices into your research.* https://www.apa.org/science/about/psa/2019/02/open-science

Asbrock, F. (2010). Stereotypes of social groups in Germany in terms of warmth and competence. *Scoial Psychology, 41*(2), 76-81. https://doi.org/10.1027/1864-9335/a000011

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*(3), 258–290. https://doi.org/10.1037/h0055756

Asparouhov, T., & Muthén, B. O. (2014). Multiple-group factor analysis alignment. *Structural Equation Modelling, 21(4)*, 1-14. https://doi.org/10.1080/10705511.2014.919210

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

Bansak, K., Hainmueller, J., & Hangartner, D. (2016). How economic, humanitarian, and religious concerns shape European attitudes toward asylum seekers. *Science, 354(*6309*),* 217-222. https://doi.org/10.1126/science.aag2147

Bakan, D. (1956). *The Duality of Human Existence: An Essay on Psychology and Religion.*, Oxford, UK: Rand McNally.

Benítez, I., & Padilla, J. L. (2014). Analysis of nonequivalent assessments across different linguistic groups using a mixed methods approach: Understanding the causes of differential item functioning by cognitive interviewing. *Journal of Mixed Methods Research*, *8*(1), 52–68. https://doi.org/10.1177/1558689813488245

Berlin, K. S., Williams, N. A., and Parra, G. R. (2013). An introduction to latent variable mixture modeling (Part 1): Overview and cross-sectional latent class and latent profile analyses. *Journal of Pediatric Psychology, 29*(2), 174-187. https://doi.org/10.1093/jpepsy/jst084

Binggeli, S., Krings, F., & Sczesny, S. (2014a). Perceived competition explains regional differences in the stereotype content of immigrant groups. *Social Psychology, 55*(1), 62-70. https://doi.org/10.1027/1864-9335/a000160

Binggeli, S., Krings, F., & Sczesny, S. (2014b). Stereotype content associated with immigrant groups in Switzerland. *Swiss Journal of Psychology, 73*(2), 123-133. https://doi.org/10.1024/1421-0185/a000133

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. https://doi.org/10.1177/0022022117749042

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

Caprariello, P. A., Cuddy, A. J., & Fiske, S. T. (2009). Social structure shapes cultural stereotypes and emotions: A causal test of the stereotype content model. *Group Processes & Intergroup Relations*, *12*(2), 147-155. https://doi.org/10.1177/1368430208101053

Chan, T., Wang, I., & Ybarra, O. (2019). Connect and strive to survive and thrive: The evolutionary meaning of communion and agency. In A. E. Abele & B. Wojciszke (Eds.), *Agency and Communion in Social Psychology* (pp. 13-24). New York: Routledge.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464-504. https://doi.org/10.1080/10705510701301834

Conway, M., Pizzamiglio, M. T., & Mount, L. (1996). Status, communality, and agency: Implications for stereotypes of gender and other groups. *Journal of Personality and Social Psychology, 71*(1)*,* 25–38. https:// doi.org/10.1037/0022-3514.71.1.25

Crandall, C. S., Bahns, A. J., Warner, R., & Schaller, M. (2011). Stereotypes as justifications of prejudice. *Personality and Social Psychology Bulletin*, *37*(11), 1488-1498. https://doi.org/10.1177/0146167211411723

Crößmann, A., & Günther, L. (2018). *Datenreport 2018. Ein Sozialbericht für die Bundesrepublik Deutschland* [Data report 2018. A social report for the Federal Republic of Germany] (pp. 149–165). Bundeszentrale für politische Bildung. https://www.bpb.de/system/files/dokument_pdf/dr2018_bf_pdf_ganzes_buch_online.pdf

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology, 92*(4), 631-648. https://doi.org/10.1037/0022-3514.92.4.631

Cuddy, A. j. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology, 40*, 61-149. https://doi.org/10.1016/S0065-2601(07)00002-0

Cuddy, A. J. C., Fiske, S. T., Kwan, V. S. Y., Glick, P., Demoulin, S., Lexens, J.-P., Bond, M. H., Croizet, J.-C., Ellemers, N., Sleebos, E., Htun, T. T., Kim, H.-J., Maio, G., Perry, J., Petkova, K., Todorov, V., Rodríguez- Bailón, R., Morales, E., Moya, M.,... Ziegler, R. (2009). Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology, 48*(1), 1-33. https://doi.org/10.1348/014466608X314935

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

DGB Niedersachsen (2020, April 02). *Systemrelevante Berufe: Kostenloser Applaus reicht nicht* [System-relevant occupations: Applause is not enough]. https://niedersachsen.dgb.de/themen/++co++76b18518-74b6-11ea-8b82-52540088cada

Ditlmann, R., Koopmans, R., Michalowski, I., Rink, A., & Veit, S. (2016). *Verfolgung vor Armut: Ausschlaggebend für die Offenheit der Deutschen ist der Fluchtgrund* [Persecution before poverty: Decisive for the openness of Germans is the reason for flight], WZB Mitteilungen, 151, 1-27.

Dovidio, J. F., Hewstone, M., Glick, P., & Esses, V. M. (2010). Prejudice, stereotyping and discrimination: Theoretical and empirical overview. In J. D. Dovidio, M. Hewstone, P. Glick & V. M. Esses (Eds.), *The SAGE Handbook of Prejudice, Stereotyping and Discrimination* (pp. 3-29). London: SAGE Publications Ltd. https://doi.org/10.4135/9781446200919.n1

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

Durante, F., Fiske, S. T., Kervyn, N., Cuddy, A. J., Akande, A. D., Adetoun, B. E., Adewuyi, M. F., Tserere, M. M., Al Ramiah, 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. https://doi.org/10.1111/bjso.12005

Durante, F., Volpato, C., & Fiske, S. T. (2010). Using the stereotype content model to examine group depictions in Fascism: An archival approach. *European Journal of Social Psychology, 40*(3), 465-483. https://doi.org/10.1002/ejsp.637

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation*. Hillsdale, NJ: Erlbaum.

Eagly, A. H., & Steffen, V. J. (1984). Gender stereotypes stem from the distribution of women and men into social roles. *Journal of Personality and Social Psychology, 46*(4), 735-754. https://doi.org/10.1037/0022-3514.46.4.735

Eckes, T. (2002). Paternalistic and envious gender stereotypes: Testing predictions from the stereotype content model. *Sex Roles*, *47*(3-4), 99-114. https://doi.org/10.1023/A:1021020920715

Ellemers, N. (2017). *Morality and the regulation of social behavior.* Milton Park, UK: Routledge / Taylor & Francis. https://doi.org/10.4324/9781315661322

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

Ellemers, N., Fiske, S. T., Abele, A. E., Koch, A., & Yzerbyt, V. (2020). Adversarial alignment enables competing models to engage in cooperative theory building toward cumulative science. *Proceedings of the National Academy of Sciences of the United States of America, 117*(14), 7561-7567. https://doi.org/10.1073/pnas.1906720117

Erikson, E. H. (1950). *Childhood and society*. New York: Norton.

Fiske, S. T. (2017). Prejudices in cultural contexts: Shares stereotypes (gender, age) versus variable stereotypes (race, ethnicity, religion). *Perspectives on Psychological Science, 12*(5), 791-799. https://doi.org/10.1177/1745691617708204

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science, 27*(2), 67-73. https://doi.org/10.1177/0963721417738825

Fiske, S. T. (2019). Warmth and competence as parallels to communion and agency: Stereotype content model. In A. E. Abele & B. Wojciszke (Eds.), *Agency and Communion in Social Psychology* (pp. 39-51). New York: Routledge.

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. https://doi.org/10.1016/j.tics.2006.11.005

Fiske, S. T., Cuddy, A. J., 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. https://doi.org/10.1037//0022-3514.82.6.878

Fiske, S. T., & North, M. S. (2015). Social psychological measures of stereotyping and prejudice. In J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of Personality and Social Psychological Constructs.* Oxford, UK: Academic Press.

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. https://doi.org/10.1111/0022-4537.00128

Flake, J., & 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. https://doi.org/10.1177/2515245920952393

Flake, J., 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. https://doi.org/10.1177/1948550617693063

Friehs, M. T., Aparicio Lukassowitz, F., & Wagner, U. (2020). *Stereotype content of occupational groups in Germany*. Manuscript submitted for publication.

Friehs, M.-T., Böttcher, J., Kotzur, P. F., Lüttmer, T., Wagner, U., Asbrock, F., & van Zalk, M. H. W. (2021). Examining the structural validity of stereotype content measures – A preregistered re-analysis of published data and discussion of possible future directions. Manuscript submitted for publication.

Friehs, M.-T.*, Kotzur, P. F.*, Zöller, A.-K. C., Wagner, U., & Asbrock, F. (2021*). A preregistered examination of scale properties of stereotype content measures: The German case.* Manuscript submitted for publication.

Fröhlich, L., & Schulte, I. (2019). Warmth and competence stereotypes about immigrant groups in Germany. *PLoS ONE, 14*(9), Article e0223103. https://doi.org/10.1371/journal.pone.0223103

Gilbert, G. M. (1951). Stereotype persistence and change among college students. *Journal of Personality and Social Psychology, 46*(2)*,* 245–254. https://doi.org/10.1037/h0053696

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. https://doi.org/10.1177/0146167219881434

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.

Hackensberger, A., Kalnoky, B., & Smirnova, J. (2016, June 09). Warum Flüchtlinge aus diesen Ländern oft kriminell werden [Why refugees from these countries often become delinquent]. *Die Welt*. https://www.welt.de/politik/ausland/article156077320/Warum-Fluechtlinge-aus-diesen-Laendern-oft-kriminell-werden.html

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. https://doi.org/10.1016/j.ijresmar.2020.02.004

Hansen, K., Rakic, T., & Steffens, M. C. (2017). Competent and warm? How mismatching appearance and accent influence first impressions. *Experimental Psychology, 64*(1), 27-36. https://doi.org/10.1027/1618-3169/a000348

Hansen, K., Rakic, T., & Steffens, M. C. (2018). Foreign-looking native-accented people: More competent when first seen rather than heard. *Social Psychological and Personality Science, 9*(8), 1001-1009. https://doi.org/10.1177/1948550617732389

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

He, J. C., Kang, S. K., Tse, K., & Toh, S. M. (2019). Stereotypes at work: Occupational stereotypes predict race and gender segregation in the workforce. *Journal of Vocational Behavior,115,* Article 103318. https://doi.org/10.1016/j.jvb.2019.103318

Hellmann, J. H., Berthold, A., Rees, J. H., & Hellmann, D. F. (2015). “A letter for Dr. Outgroup”: On the effects of an indicator of competence and changes of altruism toward a member of a stigmatized out-group. *Frontiers in Psychology, 6*, 1422. https://doi.org/10.3389/fpsyg.2015.01422

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18* (3-4), 117-144. https://doi.org/10.1080/03610739208253916

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology, 16*(2), 131-152. https://doi.org/10.1177/0022002185016002001

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science, 3*(2), 166-184. https://doi.org/10.1177/2515245919882903

Ihme, T. A., & Möller, J. (2015). „He who can, does; he who cannot, teaches?”: Stereotype threat and preservice teachers. *Journal of Educational Psychology, 107*(1), 300-308. https://doi.org/10.1037/a0037373

Imhoff, R., Woelki, J., Hanke, S., & Dotsch, R. (2013). Warmth and competence in your face! Visual encoding of stereotype content. *Frontiers in Psychology, 4*, 386. https://doi.org/10.3389/fpsyg.2013.00386

Infratest (2016). *Sorgen über die Folgen der Flüchtlingszuwanderung nach Deutschland* [Worries about the consequences of refugee immigration to Germany]. Infratest dimap. http://www.infratest-dimap.de/umfragenanalysen/bundesweit/ard-deutschlandtrend/2016/maerz/

Janda, C., Asbrock, F., Herget, M., Kues, J. N., & Weise, C. (2019). Changing the perception of premenstrual dysphoric disorder – An online-experiment using the Stereotype Content Model. *Women & Health, 59*(9), 967-984. https://doi.org/10.1080/03630242.2019.1584599

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

Katz, D., & Braly, K. W. (1933). Racial stereotypes of 100 college students. *Journal of Abnormal and Social Psychology, 28*(3)*,* 280–290. https://doi.org/10.1037/h0074049

Kemme, S., Essien, I., & Stelter, M. (2020). Antimuslimische Einstellungen in der Polizei? Der Zusammenhang von Kontakthäufigkeit und -qualität mit Vorurteilen gegenüber Muslimen [Anti-Muslim attitudes in the police force? The relationship of frequency and quality of contact with prejudice towards Muslims]. *Monatsschrift für Kriminologie und Strafrechtsreform, 103*(2), 129-149. https://doi.org/10.1515/mks-2020-2048

Kervyn, N., Chan, E., Malone, C., Korpusik, A., & Ybarra, O. (2014). Not all disasters are equal in the public’s eye: The negativity effect on warmth in brand perception. *Social Cognition, 32*(3), 256-275. https://doi.org/10.1521/soco.2014.32.3.256

Kervyn, N., Fiske, S., & Yzerbyt, V. (2015). Forecasting the primary dimension of social perception. *Social Psychology, 46*(1), 36-45. https://doi.org/10.1027/1864-9335/a000219

Kervyn, N., Yzerbyt, V., & Judd, C. M. (2010). Compensation between warmth and competence: Antecedents and consequences of a negative relation between the two fundamental dimensions of social perception. *European Review of Social Psychology, 21*(1), 155-187. https://doi.org/10.1080/13546805.2010.517997

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). New York: Guilford Press.

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. https://doi.org/10.1037/pspa0000046

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 B. Grawronski (Ed.), *Advances in Experimental Social Psychology*, (Vol. 63, pp. 1-68). https://doi.org/10.1016/bs.aesp.2020.11.001

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

Kotzur, P. F., Friehs, M.-T., Schmidt, P., Wagner, U., Pötzschke, S., & Weiß, B. (2021*). Attitudes towards refugees: Introducing a short three-dimensional scale.* Manuscript submitted for publication.

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. https://doi.org/10.1027/1864-9335/a000312

Kotzur, P. F., Schäfer, S., & Wagner, U. (2019). Meeting a nice asylum seeker: Intergroup contact changes stereotype content perceptions and associated emotional prejudice, and encourages solidarity-based collective action. *British Journal of Social Psychology, 58*(3), 668-690. https://doi.org/10.1111/bjso.12304

Kotzur, P. F., Veit, S., Namyslo, A., Holthausen, M.-A., 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. https://doi.org/10.1111/bjso.12375

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

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537-567. https://doi.org/10.1146/annurev.psych.50.1.537

Latcheva, R. (2011). Cognitive interviewing and factor- analytic techniques: A mixed method approach to validity of survey items measuring national identity. *Quality* & *Quantity*, *45*(6), 1175–1199. https://doi.org/10.1007/s11135-009-9285-0

Leach, C., Ellemers, N., & Barreto, M. (2007). Group virtue: the importance of Morality versus Competence and Sociability in the evaluation of in-groups. *Journal of Personality and Social Psychology, 93*(2), 234-249. https://doi.org/10.1037/0022-3514.93.2.234

Lee, T. L., & Fiske, S. T. (2006). Not an outgroup, not yet an ingroup: Immigrants in the stereotype content model. *International Journal of Intercultural Relations, 30*(6), 751-768. https://doi.org/10.1016/j.ijintrel.2006.06.005

Locke, E. A., & Schattke, K. (2019). Intrinsic and extrinsic motivation: Time for expansion and clarification. *Motivation Science, 5*(4), 277-290. https://doi.org/10.1037/mot0000116

Lotzkat, G., & Welpe, I. M. (2015). Gibt es Geschlechterunterschiede in der Wahrnehmung von Berufsgruppen? [Are there gender differences in the perception of occupational groups?]. In I. M. Welpe, P. Brosi, L. Ritzenhöfer, & T. Schwarzmüller (eds.), Auswahl von Männern und Frauen als Führungskräfte [Selection of males and females as leaders] (pp. 167-182). Wiesbaden: Springer Fachmedien. https:// doi.org/10.1007/978-3-658-09469-0_15

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*(1), 21–39. https://doi.org/10.1037/1082-989x.10.1.21

MacCallum, R., Widaman, K., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

Magraw-Mickelson, Z., Hermida Carrilo, A., Owuamalam, C. K., Weerabangsa, M. M., & Gollwitzer, M. (2021). *Comparing classic and novel approaches to measurement invariance*. PsyArxiv. https://doi.org/10.31234/osf.io/pz8u9

Markey, P. M. (2002). *The duality of personality: Agency and communion in personality traits, motivation, and behaviour* [Unpublished doctoral dissertation]. University of California Riverside.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: implications for cognition, emotion, and motivation. *Psychological Review, 98*(2), 224-253. https://doi.org/10.1037/0033-295X.98.2.224

Maruyama, G. (1997). *Basics of structural equation modeling*. Thousand Oakes: Sage Publications.

Meagher, B. R. (2017). Judging the gender of the inanimate: Benevolent sexism and gender stereotypes guide impressions of physical objects. *British Journal of Social Psychology*, 56(3), 537-560. https://doi.org/10.1111/bjso.12198

Meitinger, K. (2017). Necessary but insufficient. why measurement invariance tests need online probing as a complementary tool. *Public Opinion Quarterly*, *81*(2), 447–472. https:// doi.org/10.1093/poq/nfx009

Meitinger, K., Davidov, E., Schmidt, P., & Braun, M. (2020). Measurement invariance: Testing for it and explaining why it is absent. *Survey Research Methods, 14*(4), 345-349. https://doi.org/10.18148/srm/2020.v14i4.7655

Muthén, B., & Asparouhov, T. (2013, January 11). *BSEM measurement invariance analysis*. Mplus web notes no. 17. https://www.statmodel.com/examples/webnotes/webnote17.pdf

Oldmeadow, J. A. (2018). Stereotype content and morality: How competence and warmth arise from morally significant interactions. *British Journal of Social Psychology, 57*(4), 834-854. https://doi.org/10.1111/bjso.12262

Paulhus, D. P. (2019). The Big Two dimensions of desirability. In A. E. Abele & B. Wojciszke (Eds.), *Agency and Communion in Social Psychology* (pp. 79-89). New York: Routledge.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 125–167). Amsterdam: North-Holland.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63*(1), 539–569. https:// doi.org/10.1146/annurev-psych-120710-100452

Rosenberg, S., Nelson, C., & Vivekananthan, P. S. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology, 9*(4), 283–294. https://doi.org/10.1037/h0026086

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507-514. https://doi.org/10.1007/BF02296192

Saucier, G. (2009). What are the most important dimensions of personality? Evidence from studies of descriptors in diverse languages. *Social and Personality Psychology Compass, 3*(4), 620-637. https://doi.org/10.1111/j.1751-9004.2009.00188.x

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. (2010). What multi-method data tell us about construct validity. *European Journal of Personality, 24*(3), 241–257. https://doi.org/10.1002/per.771

Sczesny, S., Nater, C., & Eagly, A. H. (2019). Agency and communion: Their implications for gender stereotypes and gender identities. In A. E. Abele & B. Wojciszke (Eds.), *Agency and Communion in Social Psychology* (pp. 103-116). New York: Routledge.

Sechrest, L., Fay, T. L., & Zaidi, S. H. (1972). Problems of translation in cross-cultural research. *Journal of Cross-Cultural Psychology*, *3*(1), 41-56. https://doi.org/10.1177/002202217200300103

Seddig, D., Maskileyson, D., & Davidov, E. (2020). The comparability of measures in the ageism module of the fourth round of the European Social Survey, 2008-2009. *Survey Research Methods, 14*(4), 351-364. https://doi.org/10.18148/srm/2020.v14i4.7369

Sevillano, V., & Fiske, S. T. (2019). Stereotypes, emotions, and behaviors associated with animals: A causal test of the stereotype content model and BIAS map. *Group Processes & Intergroup Relations*, *22*(6), 879-900. https://doi.org/10.1177/1368430219851560

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin.

Stanciu, A. (2015). Four sub-dimensions of stereotype content: Explanatory evidence from Romania. *International Psychology Bulletin, 19*, 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. https://doi.org/10.1080/00224545.2016.1262812

Statista (2021). Anzahl der Asylanträge (insgesamt) in Deutschland zwischen 1995 und 2021 [Total number of applications for asylum in Germany between 1995 and 2021]. https://de.statista.com/statistik/daten/studie/76095/umfrage/asylantraege-insgesamt-in-deutschland-seit-1995/

Sullivan, I., DeHaven, A., & Mellor, D. (2019). Open and reproducible research on open science framework. *Current Protocols Essential Laboratory Techniques, 18*(1), e32. https://doi.org/10.1002/cpet.32

Tay, L., Meade, A. W., & Cao, M. Y. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*(1), 3–46. https://doi.org/10.1177/1094428114553062

The Fiske Lab. (n.d.). *Intergroup relations, social cognition, and social neuroscience.* https://www.fiskelab.org/cross-cultural-wc-maps/

Thielmann, I., & Hilbig, B. E. (2019). Nomological consistency: A comprehensive test of the equivalence of different trait indicators for the same construct. *Journal of Personality, 87*, 715-730. https://doi.org/10.1111/jopy.12428

UNHCR (2020). *Global trends: Forced displacement in 2019*. https://www.unhcr.org/5ee200e37.pdf

van de Vijver, F. J. R., & Poortinga, Y. H. (1982). Cross-cultural generalization and universality. *Journal of Cross-Cultural Psychology, 13*(4), 387-408. https://doi.org/10.1177/0022002182013004001

Vauclair, C. M., Hanke, K., Huang, L. L., & Abrams, D. (2016). Are Asian cultures really less ageist than Western ones? It depends on the questions asked. *International Journal of Psychology*, *52*(2), 136-144. https://doi.org/10.1002/ijop.12292

Veit, S., & Yemane, R. (2020). *Judging without knowing: How people evaluate others based on phenotype and country of origin – Technical report*. Wissenschaftszentrum Berlin für Sozialforschung. https://www.econstor.eu/handle/10419/215833

Welzel, C., Brunkert, L., Kruse, S., & Inglehart, R. F. (2021). Non-invariance? An overstated problem with misconceived causes. *Sociological Methods & Research*. Advance online publication. https://doi.org/10.1177/0049124121995521

Wetzel, E., & Roberts, B. W. (2020). Commentary on Hussey & 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. https://doi.org/10.1177/2515245920957618

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*(7), 592-598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

Wyszynski, M. C., Guerra, R., & Bierwiaczonek, K. (2020). Good refugees, bad migrants? Intergroup helping orientations towards refugees, migrants, and economic migrants in Germany. *Journal of Applied Social Psychology, 50*(10), 607-618. https://doi.org/10.1111/jasp.12699

Ybarra, O., Chan, E., Park, H., & Stanik, C. (2008). Life's recurring challenges and the fundamental dimensions: An integration and its implications for cultural differences and similarities. *European Journal of Social Psychology, 38*(7), 1083-1092. https://doi.org/10.1002/ejsp.559

Yzerbyt, V. Y. (2018). The dimensional compensation model: reality and strategic constraints on Warmth and Competence in intergroup perceptions. In A. E. Abele & B. Wojciszke (Eds.), *The Agency-Communion framework* (pp. 126-141). London, UK: Routledge. https://doi.org/10.4324/9780203703663-11

Yzerbyt, V. Y., 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. https://doi.org/10.1177/1368430205053944

Zebrowitz, L. A. (1990). *Mapping social psychology series: Social perception.* Buckingham: Open University Press.

Appendix

Preregistration for Manuscript # 1

The preregistration was competed using an AsPredicted template on March 16, 2020.

OSF preregistration link: https://osf.io/gqmvz/?view_only=9a8fc0053b634ace8ea8941b6c9423b7

Data collection

Have any data been collected for this study already? Note: 'Yes' is a discouraged answer for this preregistration form.

- Yes
- **It's complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid pre-registration nevertheless.**
- No

Hypothesis

The current project focusses on two research questions, which are very similar to those outlined in Friehs, Kotzur, Zöller, Wagner, and Asbrock (2020).

1. To what extent do English language scales of the stereotype content model (SCM; Fiske, Cuddy, Glick, & Xu, 2002) consistently measure warmth and competence as two separate dimensions of social perception across different social groups? This research question is examined using confirmatory factor analyses (CFA) of the SCM measures, which require acceptable baseline model fit as defined by the model fit criteria outlined below.
2. To what extent do English language SCM measures allow for the latent mean value comparisons of different social groups that were described in the original research articles? This research question is examined by assessing measurement invariance (up to scalar level, if feasible in the stepwise process outlined below) in the comparisons described in the original research articles (e.g., across different social groups in classic SCM research, or across conditions in experimental research).

Dependent variable

For each social group, we will analyse the warmth and competence measures of the original research articles. Warmth is commonly defined as the social group's intention concerning their goals, and competence as the capability to achieve those goals (Fiske et al., 2002). As we are conducting a reanalysis of existing data, the wording and number of items per scale will vary between studies.

Conditions

How many and which conditions will participants be assigned to?

The number and nature of conditions or comparisons across social groups varies between studies and is dependent on the original studies' design.

Analyses

Whereas most of the original studies focused on the content of stereotypes or its relationship with other constructs, in this reanalysis, we will focus on the aspect of the scales' measurement properties/structural validity. In order to pursue our research questions, we will investigate the methodical question of whether the scales assessing warmth and competence for different social groups or conditions consistently measure stereotype content as two fundamental dimensions of social perception (research question 1) using CFA of the proposed measurement models. As a second step, for those measurement models which fulfil the model fit criteria we defined for research question 1, we will examine whether these measures allow for latent mean value comparison of those comparisons reported in the original research articles by testing for (at least partial) scalar measurement invariance using a Multiple group confirmatory factor analysis (MGCFA). A similar approach was applied in Friehs et al. (2020). For all analyses described in the following, example Mplus analysis syntaxes are provided in the OSF project's files.

Regarding research question 1, for each data set, we will assess the measurement models' properties in CFA using the Mplus software. In accordance with the measurement models presented in the original studies, the measurement model is specified by a latent warmth and competence factor predicting the corresponding warmth and competence items, respectively. No cross-loadings or covarying indicator residuals are allowed (unless specified differently in the original study), the factors warmth and competence correlate with each other (see also Friehs et al., 2020). A robust maximum likelihood estimator (MLR) will be used.

Regarding research question 2, we will conduct a hierarchical step-up approach of testing measurement invariance (Brown, 2015), continually evaluating the model fit of increasingly restrictive models. The complete analytical procedure for testing both research questions is described in the following:

1. For all compared groups, the baseline model fit will be evaluated individually by applying CFA to the measurement models as described in the original studies. A model fit of RMSEA ≤ .08, SRMR ≤ .10, CFI ≥ .95 (Schermelleh-Engel, Moosbrugger, & Müller, 2003) is deemed acceptable. For groups/conditions that do not fulfil one or more of these criteria, the model will be rejected, and they will be excluded from subsequent analyses. Additionally, we will assess parameter performance for all baseline models, which is deemed satisfactory when the following criteria are fulfilled: |Latent factor correlation| ≤ 0.8, |standardized factor loadings| ≥ 0.4, significant factor loadings in the unstandardized solution, and absence of implausible parameter values (i.e., Heywood cases; Brown, 2015). Poor parameter performance will not be an exclusion criterion for further analyses, but will provide additional information about the quality of the measurement models.
2. Equal form (i.e., configural invariance) will be evaluated by testing the model fit for all groups/conditions with acceptable baseline model fit simultaneously. If the model fit is not acceptable according to the above-mentioned criteria, the group with the highest individual χ2-value in the configural model will be stepwise excluded from the analyses until the model fit is acceptable. Please note that measurement invariance assessment of within-person comparisons, such as in classical SCM research, are usually conducted using longitudinal measurement invariance modelling, which requires very large samples due to the high number of parameters that need to be estimated. Given that most of the reviewed studies have a rather

small sample size which would not permit such analyses, we decided to conduct the multiple-group confirmatory factor analysis approach (MGCFA) in all cases, which is usually applied for between-person comparisons (such as the comparison of experimental conditions; B. Muthén, personal communication, March 06, 2020).

3. The remaining groups will then be tested for equal factor loadings (i.e., metric invariance). We assess the model fit of the metric model using the model fit criteria outlined above, the Santorra-Bentler corrected χ2-difference test which compares the more restricted metric model to the more freely estimated configural model and which must not be significant ($p > .05$), and based upon the criteria for change in model fit indices proposed by Chen (2007): For $N \leq 300$, changes are $\geq -.005$ for CFI, $\leq .010$ for RMSEA, $\leq .025$ for SRMR; for $N > 300$, changes are $\geq -.010$ for CFI, $\leq .015$ for RMSEA, $\leq .030$ for SRMR.
4. If these criteria apply, the groups will be tested for equal indicator intercepts (i.e., scalar invariance), applying the same criteria as above, except for SRMR, which will be satisfactory if changes are $\leq .005$ for $N \leq 300$ and $\leq .010$ for $N > 300$ (Chen, 2007). Resulting final (partial) scalar invariance models will again be examined for parameter performance as described above.
5. If the model does not hold up to either the equal factor loadings or equal indicator intercepts assumption, it will be tested for partial measurement invariance (PMI; Byrne, Shavelson, & Muthén, 1989). Therefore, parameter restrictions of individual indicators are freed in individual groups. At least two indicators per latent factor are required to remain completely constrained to equality across all compared groups (Byrne et al., 1989). A precondition of this approach is that acceptable baseline model fit and configural MI are obtained. We will use modification indices (i.e., χ2-difference tests on each constrained or fixed parameter in CFA with one single degree of freedom) as principal source of information about predicted model fit changes when introducing PMI (Saris, Satorra, & van der Veld, 2009). We will only

allow for modifications in the PMI model that do not impair the interpretability of the final model (i.e., we will allow freely estimated factor loadings and indicator intercepts, but we will not allow for residual covariations of cross-loadings). For models that don't hold up to (partial) metric measurement invariance (i.e., equal loadings), equal intercepts will not be tested. If partial metric or scalar measurement invariance is not achieved, we will exclude groups in a step-wise approach in which the groups with the highest χ2 value in the fully restricted model (i.e., either the metric or scalar model before introducing the partial models) will be excluded. Thus, we will test if there is a subset of groups that does hold up to (partial) scalar measurement invariance.

Outliers and Exclusions

We will check data sets for implausible values (i.e., values that are beyond scale range). Implausible values will be coded as missing values. Other than that, we will submit the data as we received them from the original authors to our substantial analyses (which might result in different sample sizes than in the original studies). Since we use the robust MLR-estimator for our analyses, our findings are unlikely to be biased by uni- or multivariate outliers.

Sample Size

We will reanalyse existing data sets that have been collected and served as a basis for publications by the time we analyse them. Thus, we will not determine sample size.

Other

Data collection

The described project is an exploratory re-analysis of existing and published Stereotype Content Model (SCM; Fiske et al., 2002) data. All data have already been collected. While the authors of the original studies analysed the data with different research foci, the given data have not yet been analysed by the authors of this preregistration to test the specific research questions outlined in this preregistration.

Re-Analysis Inclusion Criteria

We reviewed SCM articles and targeted studies that

- were published not later than January 2020,
- directly referred to the SCM and not to any related model of social perception,
- used English-language samples/SCM scales,
- assessed warmth and competence scales with at least two indicators each, and
- compared different social groups within data sets (e.g., classic SCM studies), and/or the same social group across conditions within the same study along the SCM dimensions using identical scales (e.g., experimental studies).

We contacted the corresponding authors of all relevant studies and asked for data access for the purpose of re-analysis.

Literature

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (2nd Edition). The Guilford Press. New York.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structure: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Friehs, M.T., Kotzur, P. F:, Zöller, A.-K., Wagner, U., & Asbrock, F. (2020). German published stereotype content model scales are more often structurally invalid than they are valid: Analyses and implications. *Manuscript in preparation.*

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. https://doi.org/10.4324/9781315187280

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16,* 561-82. https://doi.org/ 10.1080/10705510903203433

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* (Vol. 8). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.509.4258&rep=rep1&type=pdf

Name

Re-Analysis of the Structural Validity of English Stereotype Content Data

Finally

- Class project or assignment
- Experiment
- Survey
- **Observational/archival study**
- Other (please describe)
- Other
- No response

Manuscript # 1

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

Maria-Therese Friehs[1] (ORCID: 0000-0002-5897-8226),

Johanna Böttcher[2],

Patrick F. Kotzur[3] (ORCID: 0000-0002-5193-3359),

Tabea Lüttmer,

Ulrich Wagner[4] (ORCID: 0000-0001-6716-9212),

Frank Asbrock[5] (ORCID: 0000-0002-6348-2946), and

Maarten H. W. van Zalk[2] (ORCID: 0000-0002-0185-8805)

[1] FernUniversität in Hagen, Germany

[2] University of Osnabrück, Germany

[3] Durham University, United Kingdom

[4] Philipps-University Marburg, Germany

[5] Chemnitz University of Technology, Germany

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

The authors declare no conflict of interests with respect to the authorship or the publication of this article. The authors thank Adelina Kletke und Corinne Lamadé for their assistance in preparing this manuscript.

Corresponding author: Maria-Therese Friehs, Faculty of Psychology, FernUniversität in Hagen, Germany, maria-therese.friehs@fernuni-hagen.de

Anonymized Preregistration:

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

Total Word count:

Abstract: 146

Main Text and Notes: 9,366

References: 3,620

Tables: 5

Figures: 3

Online Supplementary Materials: 2 (more will be made available upon the final acceptance of the manuscript as indicated throughout the manuscript)

Data availability statement: The used data are subjected to a re-analysis of published data, the data, materials and codebooks should be requested from the authors of the original publications. The analysis code will be made available in the OSF upon acceptance of the final manuscript version. Detailed information on all analyses conducted in this article are available upon request from the corresponding author of this article.

Abstract

The stereotype content model (SCM) plays a prominent role in social perception research when comparing the evaluation of different groups on warmth and competence dimensions. We examined the structural validity of SCM measures from publications based on data from English speaking participants. Re-analyzing 78 datasets from 43 published studies using confirmatory factor analyses and measurement invariance assessment, we found that 34.81% of the 586 re-analyzed SCM measurement models showed adequate scale dimensionality, implying a meaningful and valid warmth and competence assessment in one third of all cases. Regarding the scales' comparability as defined by measurement invariance, we found (partial) scalar invariance as precondition for meaningful mean-value comparisons in 11.43% of all cases. These findings indicate considerable validity concerns in published SCM research. We propose future directions to improve the measurement quality and validity in future SCM research and invite fellow researchers to constructively discuss these ideas.

Keywords: Stereotype Content Model, Structural Validity, Confirmatory Factor Analysis, Measurement Invariance, Social Perception

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

The Stereotype Content Model (SCM; Fiske et al., 2002) is one of the most prominent models of social perception of groups (Abele et al., 2021). It proposes two fundamental dimensions of social perception: *Warmth*, which is commonly defined as a group's perceived intention and as such is negatively predicted by the perceived competition and threat associated with a group (Fiske et al., 2007; Kervyn et al., 2015), and *competence*, which is defined as a group's capacity to enact those intentions and is positively predicted by a group's perceived status (Fiske et al., 2007). These two dimensions align with past research on person and group perception going back to the 1940s (for an overview, see Fiske et al., 2007) and are compatible with categories 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). A recent effort to align these different models as an overarching framework of social perception introduced two focal dimensions, namely a 'horizontal' (i.e., getting along, comparable to the warmth dimension) and a 'vertical' one (i.e., getting ahead, comparable to the competence dimension; Abele et al., 2021; Koch et al., 2021).

One of the SCM's distinctive features is that it allows to assess the social perception of multiple social groups simultaneously (Abele et al., 2021) and that it has an extensive record of research applications. These include the mapping of warmth and competence perceptions of social groups - hereafter called targets[1] - in various countries and contexts, comparing and explaining differences in the social perception of targets within and across contexts (e.g., Durante et al., 2013, 2017; Fiske, 2017b; The Fiske lab, n.d.), identifying regional or sample-specific variations of stereotype content (e.g., Binggeli, et al., 2014; Fiske, 2017b; Stanciu et al., 2017), and investigating the social perception of targets and subgroups depending on the used label (e.g., Lee & Fiske, 2006; Eckes, 2002; Meijs et al., 2019). Additionally, many studies varied targets as part of experimental manipulations and used the SCM's dimensions as dependent variables (e.g., Gul & Uskul, 2019). Other researchers used warmth and competence scales to rate non-human targets such as animals

(Sevillano & Fiske, 2016) and contexts of self vs. other ratings (Meijs et al., 2017). Consequently, the published applications and research questions go far beyond the initial conception of the model comparing several targets within one context. The SCM's prominence eventually lead to the conclusion that warmth and competence "appear to be universal across more than 30 nations (…) and 75 years (…), as well as targets that are individuals, subgroups, groups, nations, corporations, and species" (Fiske & North, 2015, p. 688; see also Cuddy et al., 2008, 2009; Fiske, 2018; Fiske et al., 2007). Hence, the SCM is well established and applied to diverse research questions across many different contexts. Indeed, it has shaped an entire branch of (social) psychological research and continues to do so.

Equivalence of Warmth and Competence

One major insight from reviewing the published SCM literature is that most studies have relied on the comparison of stereotype content between targets. In the SCM's most basic application, a researcher asks participants to rate targets A and B on equally worded items that operationalize warmth and competence, computes a mean score, applies a mean-value comparison (e.g., a *t*-test, ANOVAs, or cluster analyses) and concludes that target A is warmer/more competent than target B.

These comparisons can only be meaningful and valid if the warmth and competence scales represent precisely the same construct for all targets that are compared (Byrne, 2008), a core aspect of construct validity. Thus, it is an important and implicit assumption in operationalizations of existing SCM studies that the concrete warmth and competence scales meaningfully measure these two abstract constructs as fundamental dimensions of stereotypes. This critical assumption, however, has rarely been tested.

Given the definitions of warmth and competence, we argue that the cross-target-equivalence of these constructs might be less certain than assumed by presenting the following example roughly based on Fiske et al. (2002): Imagine a researcher wants to compare the targets *Welfare recipients* and *Rich people* using the SCM and applies classic items that measure warmth and competence, such as how 'competent' each target is perceived. From a participant's perspective, the term 'competent' might refer to different aspects for the two targets, e.g., for welfare recipients how able they are to secure their own livelihood, and for rich people how able they are in accumulating and multiplying wealth. While both goals pertain to money, participants might deduce substantially different abilities that the targets need to achieve these goals and rate competence with regard to these abilities (e.g., for welfare recipients: ability to identify possibilities of gainful employment or being informed about social security options vs. for rich people: ability of identifying ways to conserve and increase wealth or being informed about financial investment options). In consequence, the researcher might actually be measuring distinct conceptual meanings of 'competent' for both targets. If there are discrepancies in the operationalization of warmth and/or competence between these targets, e.g., because warmth and competence are not understood equivalently across targets, a comparison of warmth and competence scores between these two groups would not be meaningful. Thus, we argue that developing a valid operationalization of warmth and competence which is consistent with theory and assures equivalence in warmth and competence across targets is essential to validly interpret SCM results.

Statistically, such equivalence can be examined using confirmatory factor analysis (CFA) and measurement invariance assessment. CFA tests the factorial structure of the model and therefore its dimensionality (Brown, 2015). The SCM factorial structure, also known as the measurement model, consists of two latent factors (representing the underlying constructs, i.e., warmth and competence) and a number of observed items assessing warmth and competence which systematically relate to the factors and which make the latent construct measurable. Relating back to our example from above, in order to apply the scale, the researcher would first need to confirm that the warmth and competence items they used– which are based on theory – form warmth and competence scales of acceptable measurement properties (e.g., item performance, reliability, dimensionality; Flake et al., 2017). Thus, the researcher uses CFA informs to investigate whether the SCM items used to rate *Welfare recipients* and *Rich people* allow for the meaningful formation of warmth and competence scales[2].

Besides the factorial structure, the researcher further needs to ensure that the measurement properties of the SCM scale do not differ between *Welfare recipients* and *Rich people*, meaning that the warmth and competence scales should measure the same concept in an equal manner for both targets (Davidov et al., 2014). If that is the case, measurement invariance (MI) can be assumed and the researcher could validly interpret their results. Thus, when comparing different targets on warmth and competence scales, MI ensures that the concepts underlying the two SCM's dimensions are equal, which prevents the figurative comparison of apples and oranges (Chen, 2008).

To summarize, CFA and MI analyses (together with item performance and reliability assessments) constitute important aspects of structural validity, which is one essential component of construct validity (Flake et al., 2017)[3]. These aspects are essential for the interpretability of a study, for "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, p. 370).

Reviewing the SCM's Structural Validity

Halkias and Diamantopoulos (2020) have reviewed the initial scale construction process as depicted in Fiske et al. (1999, 2002) and identified a number of issues, eventually naming the entire process "highly problematic" (p. 719). Among others, the authors criticize small, homogenous samples, possible fatigue effects as a result of the high number of scale items and targets, a lack of robust methodology for scale performance assessment, and a non-transparent item selection process. Halkias and Diamantopoulos (2020) also point out that there is neither an established SCM scale which was used repeatedly by researchers across studies, nor a standard set of items, which has led to a very broad and diverse SCM item landscape. Researchers often generated or adapted an ad hoc measure, using a selection of items presented in the original articles (Fiske et al., 1999, 2002). Such practice, though problematic with regard to construct validity, is common in all areas of psychology (e.g., Flake & Fried, 2020). If items and measures were used repeatedly, solid scale construction and validity checks have rarely been presented. Only very recently, attempts have been made to systematically develop and validate SCM scales (Halkias & Diamantopoulos, 2020).

What is more, while most studies report high reliabilities (e.g., Fiske et al., 2002; Meagher, 2017; Oldmeadow & Fiske, 2012), only few publications have investigated the factorial structure and MI of the SCM. Fiske et al. (2002) conducted principal component analyses to explore the scale dimensionality of the used item pool. In all studies of the original SCM publication (Fiske et al., 2002), the principal component analyses per target revealed up to four more components than theoretically expected. Subsequent research projects often applied the same procedure. Whereas principal component analysis as an exploratory approach empirically generates a certain item-factor pattern, CFA tests are stricter due to their confirmatory nature (i.e., CFAs uses theoretically pre-defined expectations about item-factor relationships which are tested and (dis)confirmed; Brown, 2015). Though the SCM has been used in the title, abstract or keywords of about 350 publications according to the Web of Science (as in May 2021), we identified only nine SCM works which have reported CFA (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., 2016). Even fewer of these studies reported MI. Table 1 summarizes the strongly varying CFA and MI assessments in these SCM articles.

- Table 1 about here -

It can be concluded that structural validity, although highly relevant due to its many consequences for the quality of operationalizations, has not received much attention in SCM research. Moreover, the few existing findings regarding the SCM's structural validity vary substantially. Thus, in most published SCM articles that rely on mean value comparisons (e.g., when computing cluster analyses, ANOVA, or *t*-tests), warmth and competence scores were assumed to be equivalent in factorial structure and equivalence without explicitly testing for it, which might pose a threat to the meaningful and valid interpretation of study results.

Study Aims and Research Questions

To address the described shortcomings of existing SCM studies, we systematically examined the structural validity of SCM measures based on a data re-analysis. We focused on studies published in English language and based on English-speaking samples. We focused on only one language context to ensure comparability within our re-analysis and because a vast number of studies are conducted in English. Our analyses addressed the following research questions:

1. To what extent do SCM measures support the theoretically proposed dimensionality/factorial structure by assessing warmth and competence as two separate dimensions of social perception across different targets?
2. To what extent are SCM measures equivalent in their concepts of warmth and competence, thus allowing for meaningful and valid (latent) mean value comparisons of different targets?

In accordance with the definition of structural validity of Flake et al. (2017), we will also report the measurements' internal consistency and investigate the overall item and measurement performance of SCM measures. Contingent on our results, we will deduce concrete implications of our findings for SCM researchers as well as potential steps to uphold structural validity in future SCM research.

Methods

Inclusion Criteria, Data Requests and Datasets

The methods and procedure of this re-analysis were preregistered in the Open Science Framework (https://osf.io/gqmvz/?view_only=26cfcec4f651454e9b508f2bdc917a96). Data, materials and codebooks can be requested from the authors of the original study. The syntax and output files containing code and detailed results will be stored on the Open Science Framework once the final manuscript version is accepted.

Data eligibility and data access. For re-analysis, we chose research articles which complied with all of to the following criteria: Studies that

1. assessed warmth and competence with at least two items each which is a precondition to conduct CFA (Brown, 2015);
2. directly referred to the SCM and not to a related model of social perception;
3. published in English and used English language scales and/or datasets;

4. assessed and compared the social perception of at least two targets using the SCM dimensions with identical scale, as this was a precondition for measurement invariance assessments;
5. were published no later than January 2020, the start point of our project, in order not to un-intentionally compromise pending publications.

To identify eligible studies, we scanned various online platforms (e.g., Web of Science, Google Scholar, Fiske-Lab) for articles which either used SCM as a keyword or cited Fiske et al. (2002)'s initial SCM publication. First, we excluded double entries or studies that were not eligible based on the information in the abstract. Subsequently, we read the remaining studies and further excluded those that did not comply with our inclusion criteria. We then contacted the corresponding authors of eligible publications and sent two reminders in case of non-response. Parallel to this, we sent out a call for data in various professional mailing lists (e.g., SPSSI, EASP), which yielded one additional study. All data to which we gained access until end of December 2020 were included in our re-analysis. The study identification process is summarized in Figure 1.

- Figure 1 about here -

We excluded studies that focused exclusively on related concepts, such as 'agency' and 'communion' (Abele & Wojciszke, 2007). If other terms were used (e.g., 'social skills'; Lai & Babcock, 2013), we included them if the study referred directly to the SCM and argued that the constructs were equivalent to the SCM conceptualizations of warmth and competence. In some cases, studies measured additional dependent variables apart from warmth and competence. We included these studies in our re-analysis if the study presented the variables separately (i.e., not integrating the SCM and non-SCM dependent variables into a new variable). We screened for data collected in countries whose official language is not English if the studies were conducted in English or the language was not explicitly stated. In such cases, we inquired the survey language when contacting authors and excluded data from research which as not administered in English.

Our final set of data consisted of 78 datasets across 43 studies. Detailed information about each dataset can be found in Table 2. All included studies are marked in the references with an asterisk (*). We express our deepest gratitude to all researchers who responded to our inquiries and supported this project by providing their data and assistance.

- Table 2 about here -

Analysis

We prepared the data and obtained descriptive statistics using IBM SPSS Statistics, Version 26 (IBM Corporation, 2019). Confirmatory factor analyses and measurement invariance assessments were conducted in Mplus Version 8.3 (Muthén & Muthén, 1998-2019) using the robust maximum likelihood estimation, which restores missing data and is robust to non-normality (MLR; Muthén & Muthén, 1998-2019).

Analytical procedure. To comprehensively evaluate the structural validity of the data, we first assessed the factorial structure/dimensionality of the SCM measures for each target using CFA (see research question one). The CFA models also provided the reliability information required to compute the McDonald's omega internal consistency coefficient (Hayes & Coutts, 2020). We also examined the general item and scale performance in the CFA model focusing on the strength and significance of factor loadings and factor correlations as well as the existence of implausible estimates (i.e., Heywood cases).

Secondly, for each dataset separately, we assessed MI for all targets with adequate dimensionality (see research question two) to assess the scales' cross-target equivalence. We did this using multiple group confirmatory factor analysis (MGCFA), which is the most frequently used methodology in this field (van de Schoot et al., 2015). We chose the step-up hierarchical MI procedure outlined below[4]. Given the differing data structures of the re-analyzed datasets (i.e., in some cases, multiple datasets stem from one publication), we used datasets as the level of analysis.

In detail, our stepwise procedure included:

(1) Testing *CFA models* proposed in the original publication for each target separately to examine the SCM measures' factorial structure/dimensionality (see research question 1). We fitted a CFA based on the items and item-factor associations described in the original publications and assessed whether this SCM measurement model fitted the empirical data adequately (Brown, 2015).

The warmth and competence factors were modelled simultaneously with their latent variances fixed to one (Brown, 2015) and a freely-estimated factor correlation that reflected the correlations often found between the targets' warmth and competence scales (e.g., Durante et al., 2013; Kervyn et al., 2015). No item cross-loadings or residual covariances were modelled because these were not described in the original studies and because we saw no theoretical justification. An exemplary model is depicted in Figure 2. We accepted CFA models and subsequent MI models if the model fit met the following criteria: RMSEA ≤ .08, SRMR ≤ .10 and CFI ≥ .95 (Schermelleh-Engel et al., 2003). If two or more CFA models in one dataset showed acceptable model fit, we proceeded to testing MI with these models.

- Figure 2 about here -

(2) Testing c*onfigural* MI[5], that is, whether "the number of subscales (i.e., factors), the location of the items (i.e., pattern by which items load onto each factor), and postulated correlations among the subscales (i.e., existence of covariances)" (Byrne, 2008, p. 873) were equal for all targets. Configural MI was tested using MGCFA for all included targets simultaneously with freely estimated factor loadings or indicator intercepts. If configural MI was obtained, that signified that the theoretical constructs warmth and competence were associated with the same items across all targets. In the context of the SCM, configural MI would yield support that warmth and competence - as assumed universal dimensions of social perception - can indeed be universally measured for all targets.

(3) Testing *metric* MI, that is, whether the items were equally influenced by the factors across targets (Vandenberg & Lance, 2000). In the framework of the SCM, metric MI would mean that the measurement units of warmth and competence factors would be equal for all targets within one dataset. Metric MI is a precondition for (latent) correlational analyses (Steenkamp & Baumgartner, 1998). We introduced metric MI by constraining the factor loadings of identical items to be equal across targets. In addition to fulfilling the overall model fit criteria, metric MI models were accepted only if scaled Satorra-Bentler chi-square difference test (adapted to robust maximum likelihood estimation; Satorra & Bentler, 2001; Muthén & Muthén, 1998-2019) comparing the configural and

the metric model yielded a non-significant result. This indicated that the model fit of the metric MI model was not substantially worse than the configural model. We also observed the changes in model fit indicators, applying the criteria proposed by Chen (2007): Compared to the configural model, model fit changes should be ΔRMSEA ≤ 0.010, ΔCFI ≥ -0.005, ΔSRMR ≤ 0.025 for *N* ≤ 300, and ΔRMSEA ≤ 0.015, ΔCFI ≥ -0.010, ΔSRMR ≤ 0.030 for *N* > 300.

(4) Finally, testing *scalar* MI, that is, whether "subjects with the same latent factor score (…) have similar responses on average for an item (i.e., observed score) when the latent factor score is zero" (Sass, 2011, p. 349), or in other words, whether they had the same point of zero (Boer et al., 2018). This ensures that the items were equally difficult (Vandenberg & Lance, 2000). Scalar MI was examined by additionally constraining intercepts of identical observed variables to be equal across targets (Davidov et al., 2014). In the framework of the SCM, scalar MI would mean that there is no systematic item bias leading to over- or underestimation of any dimension between targets. Scalar MI is a precondition for meaningful mean comparisons (e.g., Bryne, 2008). We assumed scalar MI if scaled Satorra-Bentler chi-square difference test comparing the metric and the scalar model indicated a non-significant difference, and if the following model fit change criteria were met (Chen, 2007): ΔRMSEA ≤ 0.010, ΔCFI ≥ -0.005, ΔSRMR ≤ 0.005 for *N* ≤ 300, and ΔRMSEA ≤ 0.015, ΔCFI ≥ -0.010, ΔSRMR ≤ 0.010 for *N* > 300.

Due to the high number of analysis steps per data set, all analyses were carried out by one of the authors and independently examined and checked by another. Possible discrepancies were resolved by discussion.

Measurement non-invariance. In case we could not establish full scalar MI, we tried to introduce the less strict partial measurement invariance (Byrne et al., 1989). In partial MI, the equality restrictions of factor loadings and/or indicator intercepts are introduced for only some (at least two; Davidov et al., 2014), but not all parameters. Partial MI is thus easier to obtain, but also limits the comparability of the scales (Sass, 2011). It was tested by freeing single model constraints in a stepwise process under the precondition that at least two items per factor remained constrained to equality for all parameters in all targets (i.e., we only introduced partial MI in models which included three or more items per factor). To select parameters, we used the highest plausible modification index from a list of all indices > 4, the cut-off for significant model fit improvement. If no satisfactory partial solution was identified, targets were stepwise excluded from the analysis. We excluded the target with the highest chi-square contribution in the fully constrained model at the respective MI level that could not be established, and then recommenced the entire testing process at the configural level.

Parameter performance. We additionally examined the parameter performance of CFA and accepted (partial) scalar models to gain diagnostic information on potential problematic items and factor correlations and how they may impact model fit. In accordance with Brown (2015), we defined adequate CFA model parameter performance as requiring (I) statistically significant factor loadings in all items, (II) standardized factor loadings of $|\lambda| \geq 0.40$, and (III) a latent warmth-competence factor correlation of $|r| \leq 0.80$. Low or insignificant factor loadings imply that the item variance explained by the factor is too low for the item to be a good representation of the measured construct. Overly high latent correlations hint at a possibly more parsimonious one-factor measurement model describing global evaluation. Moreover, adequate parameter performance required (IV) the absence of implausible estimation values (i.e., Heywood cases), such as standardized factor loadings $|\lambda| > 1$, factor correlations $|r| > 1$ and/or negative residual variances. Such implausible estimation values undermine the meaningful interpretation of the entire CFA/MI model.

Results

Factorial Structure of SCM Scales

Research question one asked to what extent the SCM scales consistently measured warmth and competence as two distinct dimensions of social perception across different targets. This was examined using CFAs to assess the model fit for each target. Summarized information for CFA model fit can be viewed in Table 3. [Detailed tables with fit results for each target and dataset will be provided in the Online Supplementary Material upon acceptance of the final manuscript version. Also, the Mplus output files containing code and detailed results will be stored on the Open Science Framework once the final manuscript version is accepted.]

- Table 3 about here -

Across 78 datasets, we tested 586 CFA models, of which 204 models showed satisfactory fit (34.81% of all targets). This indicates that in about one third of all analyzed cases, SCM scales showed an adequate factorial structure of warmth and competence in accordance with the theoretical assumptions. The rest of the CFA models did not demonstrate acceptable dimensionality. Thirty-five datasets (44.87% of all datasets) showed no acceptable CFA model fit for any target (Mdn = 3 analyzed targets, min = 2, max = 16), 15 datasets (19.23% of all datasets) showed adequate CFA model fit in only one target (Mdn = 4 analyzed targets, min = 2, max = 12), and 28 datasets (35.9% of all datasets) in two or more targets (Mdn = analyzed targets, min = 2, max = 61). Consequently, 35.9% of all datasets with a total of 189 CFA models qualified for MI testing as a precondition for meaningful mean value comparison[6]. Omega statistics revealed that warmth and competence scales were reliable on average ($M\omega_{\text{Warmth}}$ = .840, $SD\omega_{\text{Warmth}}$ = .088, min = .481, max = .977; $M\omega_{\text{Competence}}$ = .833, $SD\omega_{\text{Competence}}$ = .085, min = .411, max = .980).

Equivalence of SCM Scales Across Targets

Research question two asked to which extent the SCM scales functioned equivalently across targets, thus holding up to prerequisites of mean value comparison, that is (partial) scalar measurement invariance. This was examined with the procedure described above. Summarized results are presented in Table 3 and in Figure 3. [We will provide detailed tables including the model fit parameters for the different levels of MI per dataset and Mplus output files containing code and detailed results upon acceptance of the final manuscript version.]

- Figure 3 about here -

Before scalar invariance could be assessed, however, warmth and competence scales had to fulfil the criteria for configural and metric invariance. Out of the 28 datasets that 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 satisfactory fit in 18 datasets (23.08% of all datasets; Mdn = 2 targets, min = 2, max = 10), the partial metric model in further eight datasets (10.26% of all datasets; Mdn = 5 targets, min = 2, max = 13).

This means a total of 26 datasets (33.33% of all datasets) held up to standards of (partial) metric MI, allowing for correlational analyses. Two datasets (2.56% of all 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 to establish whether the data allowed for mean value comparison by constraining identical indicator intercepts to be equal across targets within each dataset. Nine datasets (11.54% of all datasets; Mdn = 2 targets, min = 2, max = 10) held up to full scalar MI, further twelve (15.38% of all datasets; Mdn = 2.5 targets, min = 2, max = 8) to partial scalar MI. This means that, out of the 78 re-analyzed datasets, 21 datasets (26.92%) including 67 targets (11.43% of all targets) held up to criteria of (partial) scalar MI and thus allowed for meaningful and valid (latent) mean value comparison between targets. Of those, three datasets achieved full scalar MI in all targets examined in the dataset, which means that in the remaining 18 datasets, either parameters had to be freed (introducing partial MI) or targets had to be excluded.

Parameter Performance

Moreover, we assessed the parameter performance of CFA and accepted (partial) scalar models to gain diagnostic information on potentially problematic scales. Table 4 provides an overview of CFA results. Similar information for accepted (partial) scalar models can be found in the online supplementary materials (OSM-1).

- Table 4 about here -

In 15 datasets (19.23% of all datasets), adequate parameter performance as defined above was observed in all CFA models. In the other 63 datasets (80.77% of all datasets), there was at least one parameter that did not perform adequately. In order to gain diagnostic information, we tested whether the number of these non-adequate parameters differed between accepted and rejected CFA models. Thus, we conducted χ^2-tests for low factor loadings, high factor correlations and implausible parameter estimates separately: Low and/or insignificant factor loadings appeared significantly more frequently in rejected CFA models than in accepted ones, $\chi^2(1) = 9.56$, $p = .002$, as did factor correlations $|r| > 0.8$, $\chi^2(1) = 12.54$, $p < .001$. These results indicate a potential problem with CFA model fit when items perform problematically. There was no significant difference in the amount of implausible parameter estimates, i.e., negative residual variances and/or correlations |*r*| > 1, in accepted and rejected CFA models, $\chi^2(1) = 2.66$, *p* = .103.

Additional Exploratory Analyses

Inspired by our main analyses and with the goal of gaining more diagnostic information for the future use of SCM scales, we conducted further exploratory analyses in three focus areas.

Item performance. Firstly, we analyzed how often different items were used and how they performed with the aim of gaining information for future scale development. To assess item performance, we looked at the amount of non-significant or low factor loadings as defined above. Overall, we identified 107 different warmth and competence items. Table 5 gives a small selection of items that stood out because they either performed well or problematically (for a full overview of all items, please see OSM-2). In some cases, these parameters call for caution due to their low performance, e.g., when using items such as 'competitive' or even 'competent' and 'warm' (which were very frequently used in the re-analyzed studies). But there are also items that work relatively (e.g., 'intelligent', 'efficient', 'sincere') or exceptionally well (e.g., 'educated', 'good-natured', 'trustworthy') in the CFA models. These parameters can be a first point of reference for future scale construction efforts.

- Table 5 about here -

Sample size and model fit. We also examined the relation between sample sizes and relative CFA model fit, as sample size issues are an important area of discussion in the field of CFA (Kenny et al., 2015; Wolf et al., 2013). When correlating the sample sizes (*M* = 283.31, *SD* = 236.61, *min* = 20, *max* = 1046) with the relative frequency of acceptable CFA fit within each dataset, we found a small but significant correlation, *r* = .23, *p* = .04, indicating that higher sample sizes are positively related to a higher rate of acceptable model fits.

Study design and model fit. The third analysis examined whether there was a significant difference in the average relative number of acceptable CFA models in datasets that used within-subjects versus those that applied between-subjects designs (i.e., study designs in which participants rated multiple targets versus study designs in which they rated only one target). Using an independent *t*-test, we found that within-subjects designs displayed significantly higher relative CFA model fit (*M* = 0.421, *SD* = 0.415, *min* = 0, *max* = 1) than datasets using between-subjects (*M* = 0.189, *SD* = 0.272, *min* = 0, *max* = 1), *t*(70) = 2.86, *p* = .006. Thus, in study designs that required participants to rate multiple targets in SCM scales, more CFA models showed acceptable model fit.

Discussion

The SCM (Fiske et al., 2002) postulates that the social perception of groups is founded on evaluations on two basic dimensions: Warmth and competence. This comprehensive theoretical framework has stimulated important research on social perception in many different contexts and has been applied to various research questions. We have contributed to this body of research by systematically examining the SCM's structural validity, especially its factorial structure/dimensionality and measurement invariance as preconditions for the meaningful and valid interpretation of mean value comparisons. We applied (MG)CFA to re-analyze 78 SCM datasets from 43 publications. We found that less than 35% of all targets demonstrated acceptable CFA model fit of the SCM measurement model, indicating that the theoretically proposed two-dimensional factorial structure was not supported in the majority of cases. Moreover, only 21 datasets including 11.43% of all targets allowed for meaningful (latent) mean value comparisons within datasets. Our findings indicate severe problems of structural validity in existing SCM research. In the following, we will discuss our findings in more detail. We will also provide recommendations for future research and critical reflections how future studies can build on ours.

Factorial Structure of SCM Scales

In research question one, we analyzed whether warmth and competence items validly and reliably measured stereotype content as two distinct dimensions of social perception across targets. Evidence in favor was a prerequisite for all subsequent analyses that compared targets on warmth and competence scales. CFA results revealed that the measurement models proposed in the original publications showed satisfactory model fit in little more than one third of cases; or, putting it differently, in more than 65% of all cases, the items applied to measure stereotype content could *not* be summarized into warmth and competence scales with acceptable measurement properties. This lack of acceptable CFA model fit in the majority of SCM measures implies that the items that were used to assess warmth and competence did not actually form valid scales on which we could compare the social perception of different targets (Brown, 2015). Nonetheless, the scales showed on average good reliability. Although these findings appear counter-intuitive, scales with acceptable reliability but unacceptable dimensionality are explicable because reliability and dimensionality are distinct features of scale performance (for more information, see Davenport et al., 2015; Green & Yang, 2015).

Although we are the first to systematically demonstrate the extent of the problem, indications of the issue of dimensionality have been reported sporadically in a small number of SCM studies beforehand (e.g., Stanciu, 2015, Janssens et al., 2015, Kotzur et al., 2019, 2020). Moreover, our findings give empirical support for some of the points expressed in Halkias and Diamantopoulos' (2020) recent critique of the SCM's initial scale development. Flake and Fried (2020) point out some potential issues in scale development and usage for the field of psychology more generally. 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 the appropriate operationalization 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 of the number and meaning of the dimensions of social perception (e.g., Abele et al., 2021; Brambilla et al., 2011; Kervyn et al., 2013; Koch et al., 2016, 2021; Leach et al., 2007; Stanciu, 2015). Our results indicate that most of the used SCM measures cannot be used without further ado to validly assess and compare warmth and competence perceptions. 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 factorial structures. We recommend taking this aspect into consideration for future applications and comparisons of these theoretical models: The requirements of adequate factorial structure are not singular to the SCM but applies to all research applying scale-based measurements.

Equivalence of SCM Scales Across Targets

Research question two focused on the equivalence (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. We subjected the targets that showed acceptable CFA model fit to MI analysis based on MGCFA up to (full or partial) scalar level to fulfil the statistical requirements of unbiased (latent) mean value comparisons. Our results indicated that meaningful mean value comparison along the SCM dimensions was possible for only 11.43% of targets from 21 out of 78 datasets. The absence of (partial) scalar measurement invariance in most of the cases indicates that mean value comparisons of different targets on SCM dimensions result in the mentioned figurative comparison of apples with oranges (Chen, 2008) because the targets' warmth and competence concepts are non-equivalent. Both aspects compromise a meaningful and valid interpretation of research findings (Flake et al., 2017, Hussey & Hughes, 2018, Boer et al., 2018).

In line with our results, other SCM studies (e.g., Janssens et al., 2015; Kotzur et al., 2020; Stanciu et al., 2017) and other measures in social and personality psychology (Hussey & Hughes, 2020) have also reported a certain extent of measurement non-invariance. For instance, Hussey and Hughes (2020) investigated the structural validity of 15 established measures 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 say for certain why we found such an extensive lack of MI. Measurement non-invariance on scalar level may be caused by varying social desirability or social norm influences (i.e., method bias) and propensities to respond more strongly to specific items despite equal latent variable means or different reference points (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 hypothesize that these response biases, if they show some kind of systematic pattern, 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 (128 targets in 26 datasets), 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 which would lead us to expect such biases in SCM research.

Parameter Performance and Additional Explorative Analyses

We performed additional exploratory analyses aiming to gain further diagnostic information on ill- and well-fitting models, the degree to which certain items worked well, and the impact of sample size and study design on factorial structure and measurement invariance. We found that low or non-significant factor loadings as well as overly high factor correlations appeared more frequently in rejected than in accepted models. What is more, there was only one case of low/non-significant factor loadings in the scalar models, but more than half of scalar models had cases of high factor correlations. This can be interpreted as an issue of dimensionality, because in these cases it might be assumed that a unidimensional model of global evaluation fits the data equally well (Brown, 2015), and this would question SCM’s proposed two-dimensional structure significantly. This issue should be carefully inspected in future applications of the SCM.

When correlating sample size to relative CFA model fit, we found a small but significant correlation, indicating that a higher sample size is related to a higher share of acceptable CFA model fit. This finding is also in line with previous works on structural equation modeling outlining the challenges of small sample sizes (e.g., Kenny et al., 2015; Wolf et al., 2013). We acknowledge that most of the datasets we re-analyzed were probably not intended to be subjected to such demanding methods, and therefore, the sample sizes are mostly far from optimal for CFA and MI analyses. This might have affected our overall results, an aspect we discuss below.

We also found that higher CFA fit was achieved when researchers applied a within-subjects design compared to a between-subject design. This finding supports the recently stated strength of the SCM in assessing the social perception of several targets at once (Abele et al., 2021). Additional item performance analyses based on the magnitude and significance of the factor loading for a wide range of items yielded mixed results which might inform future scale development efforts. Items such as ‘educated’ and ‘likeable’ worked impeccably, and some of these well-performing items also overlap with recent systematic scale construction projects in the SCM context (Halkias & Diamantopoulos, 2020). Other indicators such as ‘competitive’ or ‘concerned with appearance’ often did not load highly or significantly onto their respective factors and may thus have impacted CFA model fit because they may not represent the concepts of competence and warmth appropriately.

Our overall interpretation of these results neither intends to affront any member of the SCM research community nor should it be interpreted as a general claim that all SCM research is biased and invalid. Neither do we aim at devaluating the efforts of many researchers, nor at depicting the field as “inept and misguided” (Fiske, 2017a, p. 653). Indeed, some of our own research suffers from the exact structural invalidities we outlined (e.g., anonymized for peer-review A, B, C, D, E, F). 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 of how the SCM’s measures could be improved, and to eventually advance the research on social perception by taking issues of structural validity into account. To initiate such a discourse, we present concrete suggestions with the aim of ensuring highly structurally valid future SCM research. These ideas focus on how SCM dimensions can be measured reliably and validly and how to ensure comparability in the SCM framework.

Possible Future Directions

Structural validity assessment as standard. It is erroneous to assume that the structural validity of SCM measures is given in the absence of sufficient empirical tests. Therefore, we propose that reporting CFA results of measurement models and MI examinations (if applicable) becomes a standard for future SCM applications and related theories of social perception in the spirit of open and transparent research. So far, common practice included only the report of reliability coefficients or results of principal component or explorative factor analyses (Halkias & Diamantopoulos, 2020), 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 acceptable overall model fit is achieved. This procedure, compared to the bottom-up approach we chose in this manuscript, might reduce both the effort required to run MI analyses and the number of excluded targets, as it does not apply the Chen (2007)-criteria of changes in model fit that we used in the study at hand. We also encourage replications of this systematic re-analysis in other cultural and language contexts to inform about the generalizability of our findings outside the English-language context. The supplementary materials we provide in the OSF might serve as initial orientation for such replication attempts.

Increased sample size. CFA-based analyses require larger sample sizes than many re-analyzed datasets presented. 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 are affected by small sample sizes (Kenny et al., 2015; Wolf et al., 2013). In line with this, we found that studies with lower sample sizes showed a lower share of acceptable CFA models. Determining an appropriate sample size in structural equation modelling is non-trivial and depends on various criteria (for a full overview, see Brown, 2015). Consequently, we cannot make any rule-of-thumb recommendations for future sample sizes, but we encourage SCM researchers to plan studies with higher sample sizes than in previous studies, preferably using (a priori) power analysis which is also available for CFA models (for further information, see Brown, 2015; Muthén & Muthén, 2009).

SCM scales. We call for changes in the measurement of warmth and competence. From what we saw in our re-analysis, previous SCM research did not rely on one measurement of the SCM dimensions, but rather on a variety of context-, nation- or language-specific measures. We saw measurement issues in CFA and MI testing in nearly all scales in our analyses, many of which relied largely on the items used in the initial SCM publications (Fiske et al., 1999, 2002). Therefore, we believe that further reconstruction and improvement efforts on SCM measures are required. Standardized scales would contribute to cumulative science projects (e.g., Durante et al., 2013, 2017) and hold great value for researchers that work on a smaller scope and would thus struggle with validating their own scales. The item performance information we presented in this manuscript might be helpful in this process to select well-functioning warmth and competence indicators. Moreover, Halkias and Diamantopoulos (2020) recently presented a diligent scale construction project for a German SCM measure. This could serve as starting points for eventually developing validated SCM scales in multiple languages which hold up to the criteria of structural validity.

When constructing these scales, we recommend including more than three indicators for warmth and competence, because the more information is provided in the measurement model, the more analysis options are available (e.g., a larger extent of partial measurement invariance, or the analysis of warmth and competence as separate factors; Brown, 2015). More indicators 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). We are aware that this recommendation has its drawbacks, because SCM studies usually collect information about many targets at the same time (Abele et al., 2021). Increasing the number of indicators would naturally increase potential participant fatigue, which was criticized 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; He et al., 2019; Kotzur et al., 2019).

On a related note, scale development endeavors 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., 2011, 2021; Koch et al., 2016; Leach et al., 2007; Sayans-Jiménez et al., 2017; Stanciu, 2015). Scale development efforts might aim at differentiating SCM measures 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) measures for warmth and competence might also hold the advantage that deviations in measurement models might indicate different conceptualizations 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 findings of structural non-validity. The knowledge which CFA 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 of measurement models or non-invariance, for example by using complementary approaches such as cognitive interviewing or online probing (Achbari & Davidov, 2019; Benítez & Padilla, 2014; Latcheva, 2011; Meitinger, 2017; 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. 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 modeling as an alternative to regression analysis or latent profile analysis instead of cluster analysis (Brown, 2015). Also, alternatives to the MGCFA-based MI assessment and latent mean-value comparison might be applied (Kotzur et al., 2019). Other works that determined 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. The application of a broader selection of methods and their combination might lead to multifaceted, meaningful, and reliable models of social perception and a comprehensive confirmatory evaluation of the theoretical assumptions of the SCM.

Rethinking SCM theory. The fact that more than 65% of the re-analyzed measurement models showed unacceptable model fit does not only question the SCM's structural validity, but also its proposed universality. From the previous statements about the SCM's universality, we would have expected that warmth and competence are measured validly and meaningfully as separate dimensions for every target, which was certainly not the case. Therefore, we propose a re-specification of the SCM's proposed universality accompanied with empirical tests of this assumption (e.g., see Boer et al., 2018).

Critical Reflections

We obtained and analyzed data from 78 datasets from 43 studies that we identified as eligible and that were conducted in English. We chose a language-specific context to ensure that translation issues did not confound our results (Sechrest et al., 1972). Nonetheless, we acquired data from various countries (e.g., Australia, Canada, Great Britain and Northern Ireland, India, New Zealand, Pakistan, United States), which represents a number of different contexts. Moreover, our datasets included a variety of targets, sample characteristics and sizes, data collection modes, and measurement models. Thus, we believe that our findings are generalizable to SCM research conducted in English, the language in which the SCM was initially proposed. We acknowledge that unknown features of the studies at hand might have affected the generalizability of our results (e.g., context sensitivity[7]; Flake et al., 2017) and that due to the broad research landscape, studies might have escaped our notice despite the extensive literature review we conducted. Thus, future studies could build on ours by replicating our analyses in other contexts, languages and samples, with other measures, survey structures, and methods of applications.

We used MGCFA, which is frequently applied for testing MI. This is a procedure requiring numerous individual decisions, e.g., which parameters to free when establishing PMI, which might lead to non-reproducible or disputable MI solutions (Sass, 2011). Moreover, the model fit criteria and cut-off criteria we chose directly affected our results and equally allowed for a certain liberty as there is a variety of proposed indices and cut-off criteria with individual strengths and weaknesses (e.g., Chen, 2007; Hu & Bentler, 1999; Sass, 2011; Schermelleh-Engel et al., 2003; West et al., 2012). Because of potentially diverging outcomes between researchers, we used a four-eye-principle in all analyses and standardized procedures, which we made transparent and reproducible through detailed pre-registration.

Asparouhov and Muthén (2014) argued that the criteria and methods used in MGCFA for testing MI may be too strict in case of smaller deviations from the equivalence assumption (see also Muthén & Asparpuhov, 2013; van de Schoot et al., 2013). As a result, several new approaches, all of them less strict than MGCFA, have been proposed: *Approximate measurement invariance* (Muthén & Asparouhov, 2013), *exploratory structural equation modelling* (ESEM; Asparouhov & Muthén, 2009; Marsh et al., 2009), or *alignment optimization (*Asparouhov & Muthén, 2014). Although these new approaches appear promising, we chose the MGCFA approach as it has been the first and consequently the most commonly used approach to testing MI (van de Schoot et al., 2015). Thus, it has the largest literature base to draw upon regarding aspects of methodology and application. In comparison, most alternative approaches are comparatively recent (all were proposed during the last ten years) and therefore might lack application, guidelines and comparability with other approaches. What is more, unlike some other approaches, MGCFA solutions can be transferred with relative ease to further analyses, such as structural equation modelling, which results in a broader applicability for researchers.

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 analyzed presented the necessary sample size for this approach, which is why we chose MGCFA instead. Methodically, 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 MI assessment by increasing the chi-square value and reducing the estimated standard errors (B. Muthén, personal communication, March 6, 2020). But given the fact that a high chi-square value on its own was no criterion in our analysis, and that standard errors were not considered at all, we feel this bias is passable.

Conclusion

Despite these limitations, we are convinced of the relevance and critical impact of our findings on SCM research. Our results question whether valid and meaningful interpretations about warmth and competence can be made using current SCM operationalizations (Flake & Fried, 2020; Flake et al., 2017). Although we demonstrated these hidden invalidities, we believe that, in line with Popper's (1959) ideas on the scientific process, the response to this issue can only be a collective one. In line with Ellemers (2021), we hope that our work has stimulated respectful, animated, and fruitful discussions striving to collaboratively and constructively revising and improving research on the fundamental dimensions of social perception.

Footnotes

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

[2] For instance, an alternative factorial structure would be a one-factor-model of global evaluation.

[3] Flake and colleagues (2017) outline three phases of construct validation: *Substantive validation*, which includes, beyond other aspects, literature review, construct conceptualization, and item development; *structural validation*, which includes item and factor analysis, reliability, and measurement invariance assessment; and *external validatio*n, which refers to predictive, convergent and discriminant validity analysis and the examination of subgroup differences. In terms of substantive and external validation, the SCM's construct validity is well established: A literature review reveals that the theoretical underpinnings of the SCM are manifold and have been continuously advanced (e.g., Abele et al., 2021; Cuddy et al., 2008; Fiske, 2018; Fiske et al., 2002) and the scale development has been described (Fiske et al., 2002) Additionally, we have already given numerous examples of SCM applications relating to the external validation.

[4] Horn and McArdle (1992) conducted a "step-down" approach by implementing a fully-restricted model first and subsequently relaxing restrictions. However, model adaptations are more easily introduced using the "step-up" approach which successively imposes model constraints (e.g., Brown, 2015). We selected the "step-up" approach based on the broad consensus of its superiority (Vandenberg & Lance, 2000).

[5] Literature offers different terminologies for the different steps of MI testing (Brown, 2015; Horn & McArdle, 1992; Meredith, 1993): (I) configural, structural, or equal form invariance, (II) metric, weak factorial, equal factor loadings, or measurement unit invariance, (III) scalar, strong factorial, equal intercepts, or full score invariance.

[6] Kenny et al. (2015) argue that the RMSEA performs poorly in models with small degrees of freedom (e.g., df ≤ 5) or low sample size (e.g., N ≤ 50). As a consequence, we examined the effect of the RMSEA on the model acceptance rate. In total, 46 targets (7.85% of all targets) were discarded

based only on the RMSEA value with all other model fit indices being acceptable. Of these, 28 targets showed small degrees of freedom and/or low sample size according to Kenny et al. (2015).

[7] Context sensitivity was an aspect we did not examine as the majority of datasets stemmed from the United States and thus sample sizes for other contexts were too small for robust comparisons.

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. https://doi.org/10.1037/rev0000262

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. https://doi.org/10.1037/0022-3514.93.5.751

Achbari, W., & Davidov, E. (2019, July). *Re-assessing the radius of generalized trust: Measurement invariance, think aloud protocols, and the role of education* [Conference presentation]. Conference of the European Survey Research Association (ESRA), Zagreb, Croatia.

*Amaral, A. A., Powell, D. M., & Ho, J. L. (2019). Why does impression management positively influence interview ratings? The mediating role of competence and warmth. *International Journal of Selection and Assessment*, *27*(4), 315-327. https://doi.org/10.1111/ijsa.12260

*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. https://doi.org/10.1111/j.1530-2415.2011.01242.x

Asparouhov, T., & Muthén, B. O. (2009). Exploratory structural equation modelling. *Structural Equation Modeling, 16*(3), 397-438. https://doi.org/10.1080/10705510903008204

Asparouhov, T., & Muthén, B. O. (2014). Multiple-group factor analysis alignment. *Structural Equation Modelling, 21*(4), 1-14. https://doi.org/10.1080/10705511.2014.919210

Benítez, I., & Padilla, J. L. (2014). Analysis of nonequivalent assessments across different linguistic groups using a mixed methods approach: Understanding the causes of differential item functioning by cognitive interviewing. *Journal of Mixed Methods Research, 8*(1), 52–68. https://doi.org/10.1177/1558689813488245

Binggeli, S., Krings, F., & Sczesny, S. (2014). Perceived competition explains regional differences in the stereotype content of immigrant groups. *Social Psychology, 45*(1)*,* 62-70. https://doi.org/10.1027/1864-9335/a000160

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. 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. https://doi.org/10.1177/0022022117749042

*Boysen, G. A. (2017). Exploring the relation between masculinity and mental illness stigma using the stereotype content model and BIAS map. *The Journal of Social Psychology*, *157*(1), 98-113. http://doi.org/10.1080/00224545.2016.1181600

Brambilla, M., Rusconi, P., Sacchi S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology, 41*(2), 135-143. https://doi.org/10.1002/ejsp.744

Brambilla, M., Sacchi, S., Rusconi, P., & Goodwin, G. P. (2021). The primacy of morality in impression development: Theory, research, and future directions. *Advances in Experimental Social Psychology.* Advance online publication.

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

*Bufquin, D., DiPietro, R., Orlowski, M., & Partlow, C. (2018). Social evaluations of restaurant managers: The effects on frontline employees’ job attitudes and turnover intentions. *International Journal of Contemporary Hospitality Management*, *30*(3), 1827-1844. https://doi.org/10.1108/IJCHM-11-2016-0617

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structure: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456-466. https://doi.org/10.1037/0033-2909.105.3.456

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process*. Psicothema*, *20*, 872-882.

*Caprariello, P. A., Cuddy, A. J., & Fiske, S. T. (2009). Social structure shapes cultural stereotypes and emotions: A causal test of the stereotype content model. *Group Processes & Intergroup Relations*, *12*(2), 147-155. https://doi.org/10.1177/1368430208101053

*Carew, M. T., Noor, M., & Burns, J. (2019). The impact of exposure to media coverage of the 2012 Paralympic Games on mixed physical ability interactions. *Journal of Community & Applied Social Psychology*, *29*(2), 104-120. https://doi.org/10.1002/casp.2387

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464-504. https://doi.org/10.1080/10705510701301834

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

*Clow, K. A., & Leach, A. M. (2015). Stigma and wrongful conviction: All exonerees are not perceived equal. *Psychology, Crime & Law*, *21*(2), 172-185. https://doi.org/10.1080/1068316X.2014.951645

*Cornwell, J. F., Bajger, A. T., & Higgins, E. T. (2015). Judging political hearts and minds: How political dynamics drive social judgments. *Personality and Social Psychology Bulletin*, *41*(8), 1053-1068. https://doi.org/10.1177/0146167215589720

*Crandall, C. S., Bahns, A. J., Warner, R., & Schaller, M. (2011). Stereotypes as justifications of prejudice. *Personality and Social Psychology Bulletin*, *37*(11), 1488-1498. https://doi.org/10.1177/0146167211411723

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 40, pp. 61-149). Elsevier Academic Press. https:// doi.org/10.1016/S0065-2601(07)00002-0

Cuddy, A. J. C., Fiske, S. T., Kwan, V. S. Y., Glick, P., Demoulin, S., Leyens, J.-P., Bond, M. H., Croizet, J. C., Ellemers, N., Sleebos, E., Htun, T. T., Kim, H.-J., Maio, G., Perry, J., Petkova, K., Todorov, V., Rodrígues-Bailón, R., Morales, E., Moya, M., . . . Ziegler, R. (2009). Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social* Psychology, *48*(1), 1-33. https://doi.org/10.1348/014466608X314935

Davenport, E. C. Jr., 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. https://doi.org/10.1111/emip.12095

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

*Davis, W. E., Abney, S., Perekslis, S., Eshun, S. L., & Dunn, R. (2018). Multidimensional perfectionism and perceptions of potential relationship partners. *Personality and Individual Differences*, *127*, 31-38. https://doi.org/10.1016/j.paid.2018.01.039

*Diekfuss, J. A., De Larwelle, J., & McFadden, S. H. (2018). Diagnosis makes a difference: Perceptions of older persons with dementia symptoms. *Experimental Aging Research*, *44*(2), 148-161. https://doi.org/10.1080/0361073X.2017.1422475

*Durante, F., Fiske, S. T., Kervyn, N., Cuddy, A. J., 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., Iatridis, T., Contreras, J. M., . . . 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. https://doi.org/10.1111/bjso.12005

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. 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 of the United States of America, 114*(4), 669-674. https://doi.org/10.1073/pnas.1611874114

Eckes, T. (2002). Paternalistic and envious gender stereotypes: Testing predictions from the stereotype content model. *Sex Roles*, *47*(3-4), 99-114. https://doi.org/10.1023/A:1021020920715

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

*Erhart, R. S., & Hall, D. L. (2019). A descriptive and comparative analysis of the content of stereotypes about native Americans. *Race and Social Problems*, *11*(3), 225-242. https://doi.org/10.1007/s12552-019-09264-1

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

Fiske, S.T. (2017b). Prejudice in cultural contexts: Shared stereotypes (gender, age) versus variable stereotypes (race, ethnicity, religion). *Perspectives on Psychological Science, 12*(5)*,* 791-799. https://doi.org/10.1177/1745691617708204

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in* Psychological Science, *27*(2), 67-73. https://doi.org/10.1177/0963721417738825

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

Fiske, S. T., Cuddy, A. J., 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. https://doi.org/10.1037//0022-3514.82.6.878

Fiske, S. T., & North, M. S. (2015). Social psychological measures of stereotyping and prejudice. In J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of Personality and Social Psychological Constructs.* Oxford, UK: Academic Press.

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. https://doi.org/10.1111/0022-4537.00128

Flake, J., & 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. https://doi.org/10.1177/2515245920952393

Flake, J., 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. 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. 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. (2019). Stereotypes as historical accidents: Images of social class in postcommunist versus capitalist societies. *Personality and Social Psychology* Bulletin, *46*(6), 927-943. 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. https://doi.org/10.3389/fpsyg.2019.01643

*Gul, P., & Uskul, A. K. (2019). Men's perceptions and emotional responses to becoming a caregiver father: The role of Individual differences in masculine honor ideals and reputation concerns. *Frontiers in Psychology, 10*, 1442. https://doi.org/10.3389/fpsyg.2019.01442

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*, 116-132.

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. 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. https://doi.org/10.1080/19312458.2020.1718629

*He, J. C., Kang, S. K., Tse, K., & Toh, S. M. (2019). Stereotypes at work: Occupational stereotypes predict race and gender segregation in the workforce. *Journal of Vocational Behavior*, *115*(14), 103318. https://doi.org/10.1016/j.jvb.2019.103318

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

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

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science, 3*(2), 166- 184. https://doi.org/10.1177/2515245919882903

IBM Corporation. (2019). IBM SPSS Statistics (Version 26). [Computer software]. Armonk, NY: IBM Corp.

Janssens, H., Verkuyten, M., & Khan, A. (2014). Perceived social structural relations and group stereotypes: A test of the Stereotype Content Model in Malaysia. *Asian Journal of Social Psychology, 18*(1), 52-61. 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. https://doi.org/10.1177/0049124114543236

*Kervyn, N., Chan, E., Malone, C., Korpusik, A., & Ybarra, O. (2014). Not all disasters are equal in the public's eye: The negativity effect on warmth in brand perception. *Social Cognition*, *32*(3), 256- 275. https://doi.org/10.1521/soco.2014.32.3.256

*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). *European Journal of Social Psychology, 43*(7), 673-681. https://doi.org/10.1002/ejsp.1978

Kervyn, N., Fiske, S., & Yzerbyt, V. (2015). Forecasting the primary dimension of social perception*.* *Social Psychology, 46*(1), 36-45. https://doi.org/10.1027/1864-9335/a000219

*Khan, S. S., & Liu, J. H. (2008). Intergroup attributions and ethnocentrism in the Indian subcontinent: The ultimate attribution error revisited. *Journal of Cross-Cultural Psychology*, *39*(1), 16-36. https://doi.org/10.1177/0022022107311843

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. https://doi.org/10.1037/pspa0000046

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 B. Grawronski (Ed.), *Advances in Experimental Social Psychology*, (Vol. 63, pp. 1-68). https://doi.org/10.1016/bs.aesp.2020.11.001

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

Kotzur, P. F., Veit, S., Namyslo, A., Holthausen, M. A., 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. https://doi.org/10.1111/bjso.12375

Lai, L., & Babcock, L. C. (2013). Asian Americans and workplace discrimination: The interplay between sex of evaluators and the perception of social skills. *Journal of Organizational Behavior*, *34*(4), 310-326. https://doi.org/10.1002/job.1799

Latcheva, R. (2011). Cognitive interviewing and factor- analytic techniques: A mixed method approach to validity of survey items measuring national identity. *Quality* & *Quantity*, *45*(6), 1175–1199. https://doi.org/ 10.1007/s11135-009-9285-0

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. https://doi.org/10.1037/0022-3514.93.2.234

Lee, T. L., & Fiske, S. T. (2006). Not an outgroup, not yet an ingroup: Immigrants in the stereotype content model. *International Journal of Intercultural Relations, 30*(6), 751-768. https://doi.org/10.1016/j.ijintrel.2006.06.005

*Lebowitz, M. S., Ahn, W. K., & Oltman, K. (2015). Sometimes more competent, but always less warm: Perceptions of biologically oriented mental-health clinicians. *International Journal of Social Psychiatry, 61*(7), 668-676. https://doi.org/10.1177/0020764015573086

*Levine, E. E., & Schweitzer, M. E. (2015). The affective and interpersonal consequences of obesity. *Organizational Behavior and Human Decision Processes*, *127*, 66-84. https://doi.org/10.1016/j.obhdp.2015.01.002

*Marcus, J., Fritzsche, B. A., Le, H., & Reeves, M. D. (2016). Validation of the work-related age-based stereotypes (WAS) scale. *Journal of Managerial Psychology, 31*(5), 989-1004. https://doi.org/10.1108/JMP-11-2014-0320

Marsh, H. W., Muthén, B. O., Asparouhov, T., Lüdtke, O., Robitrzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling, 16*(3), 439-476. https://doi.org/10.1080/10705510903008220

*Meagher, B. R. (2017). Judging the gender of the inanimate: Benevolent sexism and gender stereotypes guide impressions of physical objects. *British Journal of Social Psychology*, *56*(3), 537-560. https://doi.org/10.1111/bjso.12198

*Meijs, M. H., Ratliff, K. A., & Lammers, J. (2017). The discrepancy between how women see themselves and feminists predicts identification with feminism. *Sex Roles*, *77*(5-6), 293-308. https://doi.org/10.1007/s11199-016-0733-8

*Meijs, M., Ratliff, K. A., & Lammers, J. (2019). Perceptions of feminist beliefs influence ratings of warmth and competence. *Group Processes & Intergroup Relations*, *22*(2), 253-270. https://doi.org/10.1177/1368430217733115

Meitinger, K. (2017). Necessary but insufficient. why measurement invariance tests need online probing as a complementary tool. *Public Opinion Quarterly*, *81*(2), 447–472. https://doi.org/10.1093/poq/nfx009

Meitinger, K., Davidov, E., Schmidt, P., & Braun, M. (2020). Measurement invariance: Testing for it and explaining why it is absent. *Survey Research Methods, 14*(4), 345-349. https://doi.org/10.18148/srm/2020.v14i4.7655

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525-543. https://doi.org/10.1007/BF02294825

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

*Mills, K. E., Han, Z., Robbins, J., & Weary, D. M. (2018). Institutional transparency improves public perception of lab animal technicians and support for animal research. *PloS one*, *13*(2), e0193262. https://doi.org/10.1371/journal.pone.0193262

*Morton, T. A., Rabinovich, A., & Postmes, T. (2012). Who we were and who we will be: The temporal context of women's in-group stereotype content. *British Journal of Social Psychology*, *51*(2), 346-362. https://doi.org/10.1111/j.2044-8309.2010.02013.x

*Motsi, T., & Park, J. E. (2020). National stereotypes as antecedents of country-of-origin image: The role of the stereotype content model. *Journal of International Consumer Marketing*, *32*(2), 115- 127. https://doi.org/10.1080/08961530.2019.1653241

*Mroz, J. E., Yoerger, M., & Allen, J. A. (2018). Leadership in workplace meetings: The intersection of leadership styles and follower gender. *Journal of Leadership & Organizational Studies*, *25*(3), 309-322. https://doi.org/10.1177/1548051817750542

Muthén, B., & Asparouhov, T. (2013, January 11). *BSEM measurement invariance analysis*. Mplus web notes no. 17. https://www.statmodel.com/examples/webnotes/webnote17.pdf

Muthén, L. K., & Muthén, B. O. (2009). 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. https://doi.org/10.1207.S15328007SEM0904_8

Muthén, L. K., & Muthén, B. O. (1998-2019). *Mplus User's Guide* (8th ed.). Los Angeles, CA: Muthén & Muthén. https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf

*Oldmeadow, J. A. (2018). Stereotype content and morality: How competence and warmth arise from morally significant interactions. *British Journal of Social Psychology*, *57*(4), 834-854. https://doi.org/10.1111/bjso.12262

*Oldmeadow, J. A., & Fiske, S. T. (2010). Social status and the pursuit of positive social identity: Systematic domains of intergroup differentiation and discrimination for high-and low-status groups. *Group Processes & Intergroup Relations*, *13*(4), 425-444. https://doi.org/10.1177/1368430209355650

Oldmeadow, J. A., & Fiske, S. T. (2012). Contentment to resentment: Variation in stereotype content across status systems. Analyses of Social Issues and Public Policy*, *12*(1), 324-339. https://doi.org/10.1111/j.1530-2415.2011.01277.x

Popper, K. (1959). *The logic of scientific discovery.* Abington-on-Thames: Routledge.

*Rogers, K. B., Schröder, T., & Scholl, W. (2013). The affective structure of stereotype content: Behavior and emotion in intergroup context. *Social Psychology Quarterly*, *76*(2), 125-150. https://doi.org/10.1177/0190272513480191

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*(4), 507-514. https://doi.org/10.1007/BF02296192

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.

*Schlehofer, M. M., Casad, B. J., Bligh, M. C., & Grotto, A. R. (2011). Navigating public prejudices: The impact of media and attitudes on high-profile female political leaders. *Sex Roles*, *65*(1-2), 69-82. https://doi.org/10.1007/s11199-011-9965-9

Sechrest, L., Fay, T. L., & Zaidi, S. H. (1972). Problems of translation in cross-cultural research. *Journal of Cross-Cultural Psychology, 3*(1), 41-56. https://doi.org/10.1177/002202217200300103

*Sevillano, V., & Fiske, S. T. (2016a). Fantasia: Being emotionally involved with a stereotyped target changes stereotype warmth. *International Journal of Intercultural Relations*, *54*, 1-14. https://doi.org/10.1016/j.ijintrel.2016.06.001

*Sevillano, V., & Fiske, S. T. (2016b). Warmth and competence in animals. *Journal of Applied Social* Psychology, 46, 276–293. https://doi.org/10.1111/jasp.12361

*Shea, C. T., & Hawn, O. V. (2019). Microfoundations of corporate social responsibility and irresponsibility. *Academy of Management Journal*, *62*(5), 1609-1642. https://doi.org/10.5465/amj.2014.0795

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

Stanciu, A., Cohrs, J. C., Hanke, K., & Gavreliuk, A. (2017). Within-country 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. https://doi.org/10.1080/00224545.2016.1262812

Steenkamp, J-B. E. M., & 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

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory analysis framework. *Journal of Psychoeducational Assessment, 29*(4), 347.363. https://doi.org/10.1177%2F0734282911406661

Sayans-Jímenez, 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. https://doi.org/10.3389/fpsyg.2017.01692

*Swencionis, J. K., & Fiske, S. T. (2016). Promote up, ingratiate down: Status comparisons drive warmth-competence tradeoffs in impression management. *Journal of Experimental Social Psychology, 64*, 27-34. https://doi.org/10.1016/j.jesp.2016.01.004

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Measurement invariance. *Frontiers in Psychology*, *6*, 1064. https://doi.org/10.3389/fpsyg.2015.01064

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

*Van de Ven, N., Meijs, M. H. J., & Vingerhoets, A. (2017). What emotional tears convey: Tearful individuals are seen as warmer, but also as less competent. *British Journal of Social Psychology, 56*(1), 146–160. https://doi.org/10.1111/bjso.12162

Vauclair, C. M., Hanke, K., Huang, L. L., & Abrams, D. (2016). Are Asian cultures really less ageist than Western ones? It depends on the questions asked. *International Journal of Psychology*, *52*(2), 136-144. https://doi.org/10.1002/ijop.12292

Visintin, E. P., Birtel, M. D., & Crisp, R. J. (2017). The role of multicultural and colorblind ideologies and typicality in imagined contact interventions. International Journal of Intercultural Relations, 59*, 1-8. https://doi.org/10.1016/j.ijintrel.2017.04.010

*Wang, D., Oppewal, H., & Thomas, D. (2014). Exploring attitudes and affiliation intentions toward consumers who engage in socially shared superstitious behaviors: A study of students in the east and the west. *Psychology & Marketing*, *31*(3)*,* 203-213. https://doi.org/10.1002/mar.20687

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in Structural Equation Modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 209-231). The Guilford Press*.*

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*, 505-508. https://doi.org/10.1177/2515245919882903

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*, 913-934. https://doi.org/10.1177%2F0013164413495237

*Wolf, L. J., von Hecker, U., & Maio, G. R. (2017). Affective and cognitive orientations in intergroup perception. *Personality and Social Psychology Bulletin*, *43*(6), 828-844. https://doi.org/10.1177/0146167217699582

Yzerbyt, V. Y., 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. https://doi.org/10.1177/1368430205053944

*Zickfeld, J. H., & Schubert, T. W. (2018). Warm and touching tears: Tearful individuals are perceived as warmer because we assume they feel moved and touched. *Cognition and Emotion*, *32*(8), 1691–1699. https://doi.org/10.1080/02699931.2018.1430556

*Zickfeld, J. H., van de Ven, N., Schubert, T. W., & Vingerhoets, A. (2018). Are tearful individuals perceived as less competent? Probably not. *Comprehensive Results in Social Psychology*, *3*(2), 119-139. https://doi.org/10.1080/23743603.2018.1514254

Tables and Figures

Table 1

An Overview of MI Examination in the Current SCM Literature

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 behaviors	CFA and MACS[1]	1/1 targets in both samples	partial scalar MI

Note. [1] Methods and/or model fit criteria applied deviated from the ones chosen in this paper. [2] MI level not specified.

Table 2
Dataset Descriptions

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	*N*	Description	M_{Age} (*SD*)	Sex	Target(s)
1	Amaral et al. (2019)	good-natured, sincere, friendly, well-intentioned, trustworthy, warm, communal	capable, intelligent, efficient, skillful, confident, competent, agentic	5-point Likert	123	Canadian job applicants	21	101 female, 22 male	Interviewer-rated, Self-rated, Video coder-rated
2	Asbrock et al. (2011) - Pilot study NZ sample	warm, friendly	competent, capable	9-point Likert (*not at all-extremely*)	98	New Zealand citizens	30.54 (11.61)	61 female, 36 male, 1 unreported	Asians, Business women, Christians, Disabled people, Drug addicts, Elderly people, Feminists, Homeless people, Maori, NZ Europeans, Pacific Nations people, Poor people, Pregnant women, Property developers, Rich investors, Stay-at-home mothers, Teachers, The middle class, Welfare beneficiaries

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
3	Boysen (2017) Study 1b	warm, friendly, good-natured, honest	competent, intelligent, skilled, capable	5-point Likert (*not at all-extremely*)	458	US-citizens	37 (14)	-	Alcohol, Anorexia, ASPD, BDD, Bipolar, Dependent PD, Insomnia, OCD, Orgasm, Pedophilia, Pyromania, Sexual arousal
4	Boysen (2017) Study 2	warm, friendly, good-natured, honest	competent, intelligent, skilled, capable	5-point Likert (*not at all-extremely*)	162	US-citizens	36 (12)	78 female, 84 male	Male – warm, Male – cold, Female – warm, Female – cold
5	Boysen (2017) Study 3	warm, friendly, good-natured, honest	competent, intelligent, skilled, capable	5-point Likert (*not at all-extremely*)	396	US-citizens	36 (13)	198 female, 198 male	Male – warm – arousal, Male – warm – orgasmic, Male – warm – voyeur, Male – cold – arousal, Male – cold – orgasmic, Male – cold – voyeur, Female – warm – arousal, Female – warm – orgasmic, Female – warm – voyeur, Female – cold – arousal, Female – cold – orgasmic, Female – cold – voyeur

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
6	Bufquin et al. (2018)	warm, good-natured, sincere, tolerant	competent, confident, intelligent, competitive, independent	5-point (*Strongly disagree-strongly agree*)	781	US restaurant employees	-	485 female	Restaurant managers, Co-workers
7	Caprariello et al. (2009)	good-natured, warm	competent, capable	7-point (*extremely unlikely-extremely likely*)	120	US-students	20.1 (1.6)	89 female, 31 male	Envied group, Hated group, Pitied group, Pride group
8	Carew et al. (2019)	warm, good-natured, well-intentioned, friendly, likeable	competent, confident, intelligent, capable, efficient, skillful	9-point Likert (*not at all-very much so*)	166	British citizens	41.2	125 female, 41 male	Disabled – time 1, Disabled – time 2, Non-disabled – time 1, Non-disabled – time 2
9	Clow & Leach (2015)	friendly, warm, respected, sincere, trustworthy, liked	intelligent, confident, weak, lazy, mentally ill, competent	5-point (*not at all-extremely*)	125	Canadian psychology students	20.35 (4.72)	77 female	Average person, Wrongfully convicted because of false confession, Wrongfully convicted because of jailhouse snitch, Wrongfully convicted because of mistaken eyewitness, Wrongfully convicted person

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
10	Cornwell et al. (2015) Study 1a	friendly, well-intentioned, trustworthy, warm, good-natured, sincere, tolerant	competent, confident, capable, efficient, intelligent, skillful	7-point (*not at all – very much so*)	125	US-mTurk workers	-	-	Pro-Obama – warmth, Pro-Obama – competence, Anti-Obama – warmth, Anti-Obama – competence
11	Cornwell et al. (2015) Study 1b	friendly, well-intentioned, trustworthy, warm, good-natured, sincere, tolerant	competent, confident, capable, efficient, intelligent, skillful	7-point (*not at all – very much so*)	183	US-mTurk workers	-	86 female, 97 male	Pro-Obama – warmth, Pro-Obama – competence, Anti-Obama – warmth, Anti-Obama – competence, Pro-Romney – warmth, Pro-Romney – competence, Anti-Romney – warmth, Anti-Romney – competence
12	Cornwell et al. (2015) Study 2	friendly, well-intentioned, trustworthy, warm, good-natured, sincere, tolerant	competent, confident, capable, efficient, intelligent, skillful	7-point (*not at all – very much so*)	233	US-mTurk workers	>26	133 female, 100 male	Pro-Kerry – warmth, Pro-Kerry – competence, Anti-Kerry – warmth, Anti-Kerry – competence, Pro-Bush – warmth, Pro-Bush – competence, Anti-Bush – warmth, Anti-Bush – competence

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
13	Cornwell et al. (2015) Study 3	friendly, well-intentioned, trustworthy, warm, good-natured, sincere, tolerant	competent, confident, capable, efficient, intelligent, skillful	7-point (*not at all – very much so*)	356	US-mTurk workers	-	187 female, 168 male	Anti-Cuomo/Huck – Republican – warmth, Anti-Cuomo/Huck – Republican – competence, Anti-Cuomo/Huck – Democrat – warmth, Anti-Cuomo/Huck – Democrat – competence, Anti-Biden/Christie – Republican – warmth, Anti-Biden/Christie – Republican – competence, Anti-Biden/Christie – Democrat – warmth, Anti-Biden/Christie – Democrat – competence, Pro-Cuomo/Huck – Republican – warmth, Pro-Cuomo/Huck – Republican – competence, Pro-Cuomo/Huck – Democrat – warmth, Pro-Cuomo/Huck – Democrat – competence, Pro-Biden/Christie – Republican – warmth, Pro-Biden/Christie – Republican – competence, Pro-Biden/Christie – Democrat – warmth, Pro-Biden/Christie – Democrat – competence

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
14	Crandall et al. (2011) Study 3	(un)friendly, (in)sincere, (not)warm, selfish-generous	lazy-hardworking, messy-neat, (in)capable), (un)confident, (in)competent	7-point semantic differential	130	US-students	-	84 female	Azerbaijan, Eritrea, Mauritania
15	Davis et al. (2018)	warm, friendly	competent, capable	7-point Likert (*strongly disagree-strongly agree*)	381	US-citizens	35.1	170 female, 210 male, 1 diverse	Baseline, Non-perfectionist, Self-oriented perfectionist, Socially-prescribed perfectionist
16	Diekfuss et al. (2018)	friendly, well-intentioned, trustworthy, warm, good-natured, sincere	competent, confident, capable, efficient, intelligent, skillful	5-point Likert (*not at all-extremely*)	120	US-students	20.5	82 female, 38 male	Alzheimer's disease, Chronic traumatic encephalopathy, No diagnosis, Wernicke-Korsakoff syndrome

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
17	Durante et al. (2013) Sample 1	warm, well-intentioned	competent, capable	5-point Likert	57	Asian-Australian citizens	20.2	25 female, 32 male	Aboriginal Australians, Asians, Asian Australian, Australians, Blacks, Blue collar workers, Buddhists, Catholics, Children, Christians, Elderly people, Fat people, Gay people, Handicapped people, Immigrants, Indians, Men, Middle class, Muslims, Poor people, Refugees, Rich people, Teenagers, Tradies, Unemployed people, University students, Whites, White Australians, White collar workers, Women

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
18	Durante et al. (2013) Sample 2	warm, well-intentioned	competent, capable	5-point Likert	48	White-Australian citizens	23	29 female, 19 male	Aboriginal Australians, Asian, Asian Australian, Australians, Black, Blue collar, Buddhists, Catholics, Children, Christians, Elderly, Fat people, Gay, Handicapped, Immigrants, Indians, Men, Middle class, Muslims, Poor, Refugees, Rich, Teenagers, Tradies, Unemployed, University students, White, White Australians, White collar, Women

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
19	Durante et al. (2013)	warm, well-intentioned	competent, capable	5-point Likert	125	Northern Irish citizens	20.4	102 female, 23 male	Asians, Black people, British, Catholic, Chavs, Chinese, Disabled, Eastern, Gays, Homeless, Immigrants, Irish, Muslims, Polish, Protestants, Rich people, Students, Traveler, Unemployed, Western, White collar, Whites
20	Erhart & Hall (2019) Study 1	tolerant, warm, good-natured, sincere	competent, confident, independent, competitive, intelligent	5-point Likert (*not at all - extremely*)	58	US-students	23.07 (5.07)	76% female	Asian Americans, Native Americans

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
21	Erhart & Hall (2019) Study 2	tolerant, warm, good-natured, sincere	competent, confident, independent, competitive, intelligent	5-point Likert (*not at all - extremely*)	59	US-students	23.78 (8.14)	64% female	African Americans, Native Americans
22	Gul & Uskul (2019) Study 1	friendly, well-intentioned, trustworthy, warm, good-natured, sincere, moral, loyal, fair, helpful	competent, confident, capable, efficient, intelligent, skillful	9-point Likert (*extremely disagree-extremely agree*)	155	UK students and working adults	33.59	84 female, 71 male	Breadwinner dad, Caregiver dad
23	Gul & Uskul (2019) Study 2	warm, friendly, sociable, moral, loyal, fair	competent, capable, efficient, skillful	7-point Likert (*not at all-very much*)	119	UK students and working adults	36.55	119 male	Breadwinner dad, Caregiver dad

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
24	He et al. (2019)	warm, good-natured, sincere, friendly, well-intentioned, trustworthy	competent, capable, intelligent, efficient, skillful, confident	7-point (*not at all-extremely*)	1046	US-mTurk workers	33.8 (11.26)	607 female	Accountants, Actors, Architects, Artists, Banktellers, Bartenders, Bus drivers, Cashiers, CEOs, Chefs, Childcare workers, Computer programmers, Construction workers, Custodians, Custom service representatives, Dentists, Directors, Doctors, Electricians, Engineers, Factory workers, Farmers, Financial advisors, Firefighters, Fishermen, Garbage collectors, Graphic designers, Lab technicians, Landscapers, Lawyers, Librarians, Maids, Managers, Mechanics, Medical assistants, Military, Musicians, New anchors, Nurses, Paramedics, Pilots, Plumbers, Policemen, Politicians, Postal workers, Principals, Professors, Psychiatrists, Salespersons, Scientists, Secretaries, Security guards, Taxi drivers,

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
									Teachers, Tech-support workers, Truck drivers, Unemployed people, Vets, Waiters, Welders, Writers
25	Kervyn et al. (2014)	friendly, warm	competent, capable	5-point Likert (*does not describe at all-describes extremely well*)	1000	US-citizens	46.1	514 female, 486 male	Advil, AIG, Bank of America, Goldman Sachs, Honda, JPMorgan, Morgan Stanley, Toyota, Travelers Insurance, Tylenol
26	Kervyn et al. (2013) Study 1	warm-cold, friendly-unfriendly	competent-incompetent, capable-incapable	7-point Likert	61	US-citizens	33.8	36 female, 25 male	Americans, Arabs, Asians, Black professionals, British, Christians, Elderly, Feminists, Homeless, Housewives, Irish, Jews, Mental disabilities, Middle-class, Physical disabilities, Poor, Rich Turks, Welfare recipients, Whites

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
27	Kervyn et al. (2013) Study 2	warm-cold, friendly-unfriendly	competent-incompetent, capable-incapable	7-point Likert	73	US-citizens	35.8	34 female	Asians, Atheists, Blacks, Blue-collar workers, Catholics, Children, Christians, Conservatives, Elderly people, Gays, Hispanics, Jews, Liberals, Men, Middle-class people, Muslims, Poor people, Rich people, Teenagers, White-collar workers, Whites, Women, Young people
28	Kervyn et al. (Study 3)	warm-cold, friendly-unfriendly	competent-incompetent, capable-incapable	7-point Likert	90	US-citizens	37.3	54 female, 36 male	Impotent – negative, Impotent – positive, Potent – negative, Potent – positive
29	Khan & Liu (2008) Study 1	good-natured, helpful, generous, honest	intelligent, competent	7-point Likert (*not at all true-very true*)	154	Indian students	22.9/23.6 (4.8/7.6)	97 female	Hindus, Muslims
30	Khan & Liu (2008) Study 2	good-natured, helpful, generous, honest	intelligent, competent	7-point Likert (*not at all true-very true*)	145	Pakistani citizens	29.1/29.6 (11.5/8.8)	52 female, 1 diverse	Hindus, Muslims

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
31	Lebowitz et al. (2015) Study 1	tolerant, good-natured, compassionate, warm, flexible, interested in others, open-minded, respectful	confident, competent, intelligent, capable, independent, competitive, skilled, educated	9-point semantic differential	606	US-citizens	30.57 (10.08)	61% male	Biologically-oriented clinician – bipolar disorder, Biologically-oriented clinician – major depression, Biologically-oriented clinician – narcissistic personality disorder, Biologically-oriented clinician – schizophrenia, Biologically-oriented clinician – social phobia, Psychosocially-oriented clinician – bipolar disorder, Psychosocially-oriented clinician – major depression, Psychosocially-oriented clinician – narcissistic personality disorder, Psychosocially-oriented clinician – schizophrenia, Psychosocially-oriented clinician – social phobia

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
32	Lebowitz et al. (2015) Study 2	tolerant, good-natured, compassionate, warm, flexible, interested in others, open-minded, respectful	confident, competent, intelligent, capable, independent, competitive, skilled, educated	9-point semantic differential	586	US-citizens	31.42 (11.05)	52,9% male, 44,4% female, 2,7% unknown	Biologically-oriented clinician – bipolar disorder, Biologically-oriented clinician – major depression, Biologically-oriented clinician – narcissistic personality disorder, Biologically-oriented clinician – schizophrenia, Biologically-oriented clinician – social phobia, Psychosocially-oriented clinician – bipolar disorder, Psychosocially-oriented clinician – major depression, Psychosocially-oriented clinician – narcissistic personality disorder, Psychosocially-oriented clinician – schizophrenia, Psychosocially-oriented clinician – social phobia

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
33	Lebowitz et al. (2015) Study 3	tolerant, good-natured, compassionate, warm, flexible, interested in others, open-minded, respectful	confident, competent, intelligent, capable, independent, competitive, skilled, educated	9-point semantic differential	98	US-citizens	30.99 (8.80)	48% male	Biologically-oriented, Psychosocially-oriented
34	Levine & Schweitzer (2015) Pilot Study	warm, sincere, good-natured, tolerant	competent, independent, confident, intelligent, competitive	7-point Likert	152	US-citizens	32.6	77 female, 75 male	Male – obese, Male – healthy, Female – obese, Female – healthy

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
35	Levine & Schweitzer (2015) Study 4	warm, sincere, good-natured, tolerant	competent, independent, confident, intelligent, competitive	7-point Likert	604	US-citizens	30.7	305 female, 299 male	Cold – female – stimulus 1 – weight loss, Cold – female – stimulus 1 – no weight loss, Cold – female – stimulus 2 – weight loss, Cold – female – stimulus 2 – no weight loss, Cold – male – stimulus 1 – weight loss, Cold – male – stimulus 1 – no weight loss, Cold – male – stimulus 2 – weight loss, Cold – male – stimulus 2 – no weight loss, Warm – female – stimulus 1 – weight loss, Warm – female – stimulus 1 – no weight loss, Warm – female – stimulus 2 – weight loss, Warm – female – stimulus 2 – no weight loss, Warm – male – stimulus 1 – weight loss, Warm – male – stimulus 1 – no weight loss, Warm – male – stimulus 2 – weight loss, Warm – male – stimulus 2 – no weight loss

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
36	Levine & Schweitzer (2015) Study 5	warm, sincere, good-natured, tolerant	competent, independent, confident, intelligent, competitive	7-point Likert	600	US-citizens	30	181 female, 419 male	Control – female, Control – male, Cold – female, Cold – male, Warm – female, Warm – male
37	Marcus et al. (2016) Sample 1	warm-hearted, warm personality, likeable, cold (r), kind, friendly	competent, high achiever, capable, top performer, enhances organizational productivity, skilled	6-point (*very much disagree-very much agree*)	454	US-students	19.36 (1.72)	67% female	Experience salient – with career-transition – young adult, Experience salient – with career-transition – old adult, Experience salient – between career-transition – young adults, Experience salient – between career-transition – old adults, Experience not salient – with career-transition – young adult, Experience not salient – with career-transition – old adult, Experience not salient – between career-transition – young adults, Experience not salient – between career-transition – old adults

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
38	Marcus et al. (2016) Sample 2	warm, friendly, cold manner (r), likeable, warm-hearted, kind	competent, productive, top performer, high achiever, skilful, capable	6-point (*very much disagree-very much agree*)	709	US-students	18.73 (1.64)	61% female	No intervention – with career-transition – young applicant, No intervention – with career-transition – old applicant, No intervention – between career-transition – young applicant, No intervention – between career-transition – old applicant, presence of intervention – with career-transition – young applicant, presence of intervention – with career-transition – old applicant, presence of intervention – between career-transition – young applicant, presence of intervention – between career-transition – old applicant
39	Meagher (2017) Study 2	friendly, honest, likeable, sincere, trustworthy, warm	affluent, competent, intelligent, prestigious, skilful, sophisticated, successful	6-point Likert (*not at all-extremely*)	181	US and Canadian mTurk workers	34.86 (12.04)	91 male	Men, Women

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
40	Meijs et al. (2017) Study 1	concerned with appearance, attractive, fun, likeable, nurturing, open-minded	ambitious, independent, intelligent, opinionated, career-oriented	5-point (*not at all - extremely*)	387	Female US-citizens	32.1 (13.5)	All female	Feminists, Participants themselves
41	Meijs et al. (2017) Study 2	concerned with appearance, attractive, fun, likeable, nurturing, open-minded	ambitious, independent, intelligent, opinionated, career-oriented	5-point (*not at all - extremely*)	288	Female US mTurk workers	32.7 (11.3)	All female	Feminists, Participants themselves
42	Meijs et al. (2017) Study 3	concerned with appearance, attractive, fun, likeable, nurturing, open-minded	ambitious, independent, intelligent, opinionated, career-oriented	5-point (*not at all - extremely*)	116	Female US mTurk workers	35.3 (10.6)	All female	Feminists, Participants themselves

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
43	Meijs et al. (2019) Study 1b	concerned with appearance, attractive, fun, likeable, nurturing, open-minded	ambitious, independent, intelligent, opinionated, career-oriented	7-point (*strongly disagree - strongly agree*)	610	US mTurk workers	32.4	250 female, 360 male	Control, Feminists, Women with feminist beliefs
44	Meijs et al. (2019) Study 2	concerned with appearance, attractive, fun, likeable, nurturing, open-minded	ambitious, independent, intelligent, opinionated, career-oriented	7-point (*strongly disagree - strongly agree*)	302	US mTurk workers	33.7 (12)	136 female, 166 male	Control – introduced by others, Control – self-introduced, Feminists – introduced by others, Feminists – self-introduced, Women with feminist beliefs – introduced by others, Women with feminist beliefs – self-introduced
45	Meijs et al. (2019) Study 3	concerned with appearance, attractive, fun, likeable, nurturing, open-minded	ambitious, independent, intelligent, opinionated, career-oriented	7-point (*strongly disagree - strongly agree*)	403	US mTurk workers	33.38	159 female, 244 male	Control, Feminist - only-label Feminist - with-label, Feminist - without-label

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
46	Meijs et al. (2019) Study 4	concerned with appearance, attractive, fun, likeable, nurturing, open-minded	ambitious, independent, intelligent, opinionated, career-oriented	7-point (*strongly disagree - strongly agree*)	631	US mTurk workers	31.39	212 female, 419 male	Feminist – label, Feminist – no-label, Feminist – reject-label, Non-Feminist – label, Non-Feminist – no-label, Non-Feminist – reject-label
47	Meijs et al. (2019) Study 5	concerned with appearance, attractive, fun, likeable, nurturing, open-minded	ambitious, independent, intelligent, opinionated, career-oriented	7-point (*strongly disagree - strongly agree*)	214	US mTurk workers	32.42	73 female, 141 male	Strong feminist beliefs, weak feminist beliefs
48	Meyer & Asbrock (2018) Study 1	likeable, warm, good-natured	competent, competitive, independent	5-point Likert	314	European and US Prolific Academic users	37.9 (12.44)	189 female, 118 male	Able-bodied people, Able-bodied people who choose to implant technology into their bodies to enhance their capabilities, Cyborgs, Homeless people, Old people, People with mental disabilities, People with physical disabilities, People with physical disabilities wo wear bionic protheses, Physicians, Rich people

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
49	Meyer & Asbrock (2018) Study 2	likeable, warm, good-natured	competent, competitive, independent	5-point Likert	87	European and US Prolific Academic users	33.45 (11.88)	55 female, 2 other	Arm disability – low technicality prostheses, Arm disability – medium technicality prostheses, Arm disability – high technicality prostheses, Leg disability – low technicality prostheses, Leg disability – medium technicality prostheses, Leg disability – high technicality prostheses, paraplegic – low technicality prostheses, paraplegic – medium technicality prostheses, paraplegic – high technicality prostheses
50	Mills et al. (2018)	warm, trustworthy, honest	intelligent, competent, confident	7-point Likert	550	US and Canadian mTurk workers	36	332 female, 218 male	Cow – in lab, Cow – on farm, Dog – in lab, Dog – on farm, Mouse – in lab, Mouse – on farm

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
51	Morton et al. (2012) Study 1	warm, good-natured, friendly	competent, independent, competitive, intelligent, skilled	7-point Likert (*strongly disagree-strongly agree*)	43	British students	19.3	43 female	Women – future primed, Women – past primed
52	Morton et al. (2012) Study 2	decent, warm, good-natured, friendly, moral	strong, active, skilled, competent, decent	7-point Likert (*strongly disagree-strongly agree*)	93	British students	20.4	46 female, 47 male	Future primed – female, Future primed – male, Past primed – female, Past primed – male
53	Motsi & Park (2020)	friendly, warm, good-natured, sincere	confident, competent, capable, skillful	7-point Likert	382	US students	-	226 female, 156 male	China, India

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
54	Mroz et al. (2018) Study 1	knows how to comfort others, enjoys bringing people together, feels others' emotions, takes an interest in other peoples' lives, cheers people up, makes people feel at ease, takes time out for others, doesn't like to get involved in other people's problems(r), isn't really interested in others(r), tries not to think about the needy(r)	learns quickly, uses their brain, excels in what they do, looks at the facts, meets the challenges, seeks explanation of things, needs things explained only once, knows how to apply their knowledge	5-point Likert (*not at all-very much*)	125	US-students	39.7	84 female, 39 male, 2 other	Men, Women

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
55	Mroz et al. (2018) Study 2	knows how to comfort others, enjoys bringing people together, feels others' emotions, takes an interest in other peoples' lives, cheers people up, makes people feel at ease, takes time out for others, doesn't like to get involved in other people's problems(r), isn't really interested in others(r), tries not to think about the needy(r)	learns quickly, uses their brain, excels in what they do, looks at the facts, meets the challenges, seeks explanation of things, needs things explained only once, knows how to apply their knowledge	5-point Likert (*not at all-very much*)	331	US mTurk workers	37	164 female, 163 male, 4 other	Female rater – directive leadership style, Female rater – participative leadership style, Male rater – directive leadership style, Male rater – participative leadership style

Table 2 (continued)

		Measures		Sample					
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
56	Oldmeadow & Fiske (2010) Study 1 Sample 1	sincere, friendly, trustworthy, likeable	smart, capable, intelligent, efficient	7-point Likert	69	UK-students (high status group)	17	53 female	High status college, Low status college
56	Oldmeadow & Fiske (2010) Study 1 Sample 2	sincere, friendly, trustworthy, likeable	smart, capable, intelligent, efficient	7-point Likert	100	UK-students (low status group)	17	68 female	High status college, Low status college
57	Oldmeadow & Fiske (2010) Study 2 Sample 1	warm, friendly, sincere, trustworthy	Competent, capable, intelligent, efficient	7-point Likert	64	UK-students (high status group)	25.1	51 female	High status college, Low status college
57	Oldmeadow & Fiske (2010) Study 2 Sample 2	warm, friendly, sincere, trustworthy	Competent, capable, intelligent, efficient	7-point Likert	47	UK-students (low status group)	19.7	38 female	High status college, Low status college

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
58	Oldmeadow & Fiske (2012) Study 1	friendly, warm, sincere	competent, capable, intelligent	7-point Likert	345	UK-students	20.6 (3.9)	207 female, 136 male	High-status group, Low-status group
59	Oldmeadow & Fiske (2012) Study 2	friendly, sociable, trustworthy, sincere	skilful, organized, teamwork, coordination	7-point Likert	132	UK-students	19	111 female, 21 male	High-status group, Low-status group
60	Oldmeadow (2018) Study 1	warm, friendly, trustworthy	competent, capable, intelligent	9-point Likert	60	UK-students	25 (6.4)	36 female, 24 male	Circles, Squares
61	Oldmeadow (2018) Study 2	honest, sincere, trustworthy, likeable, warm, friendly	competent, clever, skilled, assertive, confident, determined	7-point Likert	108	Australian students	30.45 (10.71)	94 female	Circles, Squares

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
62	Rogers et al. (2013)	warm, tolerant, good-natured, sincere	competent, confident, independent, competitive, intelligent	5-point (*1 = not at all to 5 = extremely*)	92	US-Students	-	52% female, 48% male	Americans, Arabs, Asians, Black professionals, Blacks, Blue-collar workers, Businesswomen, Christians, Feminists, Gay men, Housewives, Immigrants, Italians, Jews, Men, Migrant workers, Myself as I really am, Politicians, Students, The British, The disabled, The elderly, The homeless, The Irish, The mentally retarded, The middle class, The poor, The rich, The unemployed, The well-educated, The young, Turks, Welfare recipients, Whites, Women
63	Schlehofer et al. (2011)	tolerant, warm, good-natured, sincere, friendly, well-intentional, trustworthy	competent, confident, independent, competitive, intelligent, skilful, capable, efficient	7-point Likert (*not at all-extremely*)	341	US-students	23.5	174 female, 167 male	Positive article about female politician – female rater, Positive article about female politician – male rater, Negative article about female politician – female rater, Negative article about female politician – male rater

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
64	Sevillano & Fiske (2016a) Study 1	sincere, friendly, well-intentioned, trustworthy, good-natured, warm	competent, confident, capable, intelligent, efficient, skilful	9-point Likert (*not at all-extremely*)	176	US-citizens	27.2	123 female	Objective rater, Perspective taking
65	Sevillano & Fiske (2016a) Study 2	sincere, friendly, well-intentioned, trustworthy, good-natured, warm	competent, confident, capable, intelligent, efficient, skilful	9-point Likert (*not at all-extremely*)	73	US-students	19.5	35 female	Cooperative framework, Competitive framework
66	Sevillano & Fiske (2016b)	friendly, well-intentioned, warm	competent, intelligent, skilful	9-point Likert (*not at all-extremely), 10* (*does not apply*)	135	US-citizens	36.4	81 female, 54 male	Bear, Bird, Cat, Chicken, Cow, Dog, Duck, Elephant, Fish, Giraffe, Hamster, Hippopotamus, Horse, Leopard, Lion, Lizard, Monkey, Mouse, Pig, Rabbit, Rat, Snake, Tiger, Whale, Zebra

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
67	Shea & Hawn (2019)	tolerant, warm, good-natured, sincere, sociable, caring, unfriendly, insensitive, friendly	competent, confident, independent, competitive, intelligent, capable, skilled, lazy, disorganized	5-point Likert	109	US-citizens	31.1	70 male, 39 female	Control, Corporate social irresponsibility, Corporate social responsibility
68	Swencionis & Fiske (2016) Study 1	considerate, cooperative, courteous, forgiving, generous, kind, patient, sincere, trustworthy, understanding	ambitious, capable, clever, creative, independent, intelligent, logical, responsible, self-reliant, talented	7-point	151	US mTurk workers	/	83 female	Equally-ranked colleague, Higher-ranked colleague, Lower-ranked colleague
69	Van de Ven et al. (2017) Study 2	warm, nice, friendly, sincere	competent, self-assured, capable, skilled	7-point Likert	653	US mTurk workers	34.3	312 female, 339 male, 2 diverse	Female – no tears, Female – tears, Male – no tears, Male – tears

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
70	Visintin et al. (2017)	friendly, likable, helpful	intelligent, competent, capable	7-point Likert (*not at all–very much*)	76	British students	20	63 female, 13 male	Asians, British Muslims, Moroccans, People with schizophrenia, Physically disabled people
71	Wang et al. (2014) Study 2	gentle, helpful to others, kind, aware of feelings of others, understanding, warm in relation to others, friendly	independent, active, competitive, decisive, never gives up easily, self-confident, feels superior, stands up under pressure, competent	7-point scale	165	Australian students	-	65 male	Eastern superstitious behavior, Eastern superstitious palm-reading, No superstitious behavior, No superstitious talking, Western superstitious behavior, Western superstitious tarot-card-reading
72	Wolf et al. (2017) Study 1	helpful, warm, good-natured	ambitious, skillful, competent	7-point Likert (*very uncharacteristic-very characteristic*)	200	US-citizens	38.1	94 female, 106 male	American people, Elderly, German people, Homeless people, Middle-class people, Welfare recipients

Table 2 (continued)

		Measures			Sample				
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
73	Wolf et al. (2017) Study 2	affectionate, sociable, helpful, happy, warm, good-natured	determined, skillful, persistent, intelligent, competent, ambitious	5-point Likert (*very uncharacteristic-very characteristic*)	125	US-citizens	27.5	52 female, 73 male	Asian people, Elderly, German people, Housewives, Rich people, South American people
74	Wolf et al. (2017) Study 3 Sample 1	empathetic, helpful, sentimental, humorous, happy, popular, sociable, good-natured, warm	scientific, determined, persistent, skillful, industrious, intelligent, independent, ambitious	5-point Likert (*very uncharacteristic-very characteristic*)	120	US-citizens	36.13	48 female, 68 male, 4 other	Children, Elderly people, Housewives, Irish people, Italian people, South American people
75	Wolf et al. (2017) Study 3 Sample 2	affectionate, sociable, helpful, happy, warm, good-natured	determined, skillful, persistent, intelligent, competent, ambitious	5-point Likert (*very uncharacteristic-very characteristic*)	123	US-citizens	28.2	59 female, 61 mal, 3 other	Asian people, Feminists, German people, Jewish people, Professionals, Rich people
76	Zickfeld et al. (2018) Study 1	warm, nice, friendly, sincere	competent, self-assured, capable, skilled	7-point Likert	518	US-citizens	35.9	279 female, 239 male	Female – no tears, Female – tears, Male – no tears, Male – tears

Table 2 (continued)

		Measures		Sample					
Dataset No.	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (*SD*)	Sex	Target(s)
77	Zickfeld et al. (2018) Study 2	warm, nice, friendly, sincere	competent, self-assured, capable, skilled	7-point Likert	471	US-citizens	37.1	202 female, 254 male, 1 diverse	Female – no tears, Female – tears, Male – no tears, Male – tears
78	Zickfeld & Schubert (2018)	warm, nice, friendly, sincere	competent, self-assured, capable, skilled	7-point Likert	350	US-citizens	37.9	149 female, 198 male, 3 diverse	Female – no tears, Female – tears, Male – no tears, Male – tears

Table 3

CFA and MI results for all datasets

				Measurement Invariance				
Dataset	Targets	Total CFA Fit	% CFA Fit	Configural	Metric (Groups)	% Targets[1]	Scalar (Groups)	% Targets[1]
1	3	0	00.0	/	/	/	/	/
2	20	9	45.0	Full	Full (6)	66.7	Full (4)	44.4
3	12	1	8.3	/	/	/	/	/
4	4	1	25.0	/	/	/	/	/
5	12	3	25.0	Full	/	/	/	/
6	2	2	100.0	Full	Partial (2)	100.0	Partial (2)	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	Full	Full (3)	100.0	Full (3)	100.0
15	5	4	80.0	Full	Full (2)	50.0	Full (2)	50.0
16	4	0	00.0	/	/	/	/	/
17	30	20	66.7	Full	Full (8)	40.0	Full (2)	10.0
18	30	23	76.7	Full	Full (10)	43.5	Full (2)	8.6
19	22	14	63.6	Full	Full (6)	42.9	/	/
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	Full	Partial (13)	100.0	Partial (8)	61.5
25	10	10	100.0	Full	Full (10)	100.0	Full (10)	100.0

Table 3 (continued)

				Measurement Invariance				
Dataset	Targets	Total CFA Fit	% CFA Fit	Configural	Metric (Groups)	% Targets[1]	Scalar (Groups)	% Targets[1]
26	20	9	45.0	Full	Full (2)	22.2	/	/
27	23	11	47.8	Full	Full (7)	63.6	/	/
28	4	0	00.0	/	/	/	/	/
29	2	2	100.0	Full	Full (2)	100.0	Full (2)	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	Full	/	/	/	/
36	6	3	50.0	Full	Partial (3)	100.0	Partial (3)	100.0
37	8	2	25.0	Full	Full (2)	100.0	Full (2)	100.0
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	Full	Partial (8)	100.0	Partial (2)	25.0
49	3	3	100.0	Full	Full (2)	66.7	Partial (2)	66.7
50	6	3	50.0	Full	Partial (2)	66.7	/	/
51	2	0	00.0	/	/	/	/	/
52	4	0	00.0	/	/	/	/	/
53	2	2	100.0	Full	Full (2)	100.0	/	/
54	2	0	00.0	/	/	/	/	/

Table 3 (continued)

				Measurement Invariance				
Dataset	Targets	Total CFA Fit	% CFA Fit	Configural	Metric (Groups)	% Targets[1]	Scalar (Groups)	% Targets[1]
55	4	0	00.0	/	/	/	/	/
56	2	0	00.0	/	/	/	/	/
57	2	0	00.0	/	/	/	/	/
58	2	2	100.0	Full	Full (2)	100.0	Partial (2)	100.0
59	2	1	50.0	/	/	/	/	/
60	2	1	50.0	/	/	/	/	/
61	2	0	00.0	/	/	/	/	/
62	35	8	22.8	Full	Partial (6)	75.0	Partial (2)	25.0
63	4	1	25.0	/	/	/	/	/
64	2	0	00.0	/	/	/	/	/
65	2	0	00.0	/	/	/	/	/
66	25	15	60.0	Full	Partial (13)	86.7	Partial (4)	26.7
67	3	1	33.3	/	/	/	/	/
68	3	0	00.0	/	/	/	/	/
69	4	2	50.0	Full	Full (2)	100.0	Partial (2)	100.0
70	5	0	00.0	/	/	/	/	/
71	3	0	00.0	/	/	/	/	/
72	6	4	66.7	Full	Partial (4)	100.0	Full (2)	50.0
73	6	1	16.7	/	/	/	/	/
74	6	0	00.0	/	/	/	/	/
75	6	1	16.7	/	/	/	/	/
76	4	4	50.0	Full	Full (4)	100.0	Partial (4)	100.0
77	4	3	75.0	Full	Full (3)	100.0	Partial (3)	100.0
78	4	4	100.0	Full	Full (4)	100.0	Partial (4)	100.0
Total	586	204	34.8	189	77 Full/51 Partial	62.7	29 Full/38 Partial	32.8

Note. [1] of accepted CFA models (column Total CFA Fit).

Table 4

Cases of Inadequate Parameter Performance in All CFA Models Separated by Accepted and Rejected Models

	Number of CFA Models		Unsuccess-ful Model Estimation	Problematic Factor Loadings[1, 2]		High Correlation[3]		Implausible Estimates[2]	
Dataset	Acc.	Rej.	Rej.	Acc.	Rej.	Acc.	Rej.	Acc.	Rej.
1	0	3	0	0	0	0	0	0	0
2	9	11	0	0	2	3	2	1	4
3	1	11	0	0	0	0	3	0	1
4	1	3	0	0	1	1	0	0	0
5	3	9	3	0	0	0	3	0	0
6	2	0	0	0	0	2	0	0	0
7	1	3	1	0	0	0	0	1	2
8	0	4	1	0	0	0	0	0	0
9	0	5	4	0	2	0	0	0	0
10	0	4	4	0	0	0	0	0	0
11	0	8	8	0	0	0	0	0	0
12	0	8	8	0	0	0	0	0	0
13	0	16	16	0	0	0	0	0	0
14	3	0	0	1	0	3	0	0	0
15	4	1	0	0	0	0	0	0	0
16	0	4	4	0	0	0	0	0	0
17	20	10	2	1	0	12	8	7	2
18	23	7	2	25	6	8	1	6	3
19	14	8	0	8	2	11	6	6	6
20	0	2	0	0	6	0	0	0	0
21	0	2	0	0	8	0	0	0	0
22	0	2	0	0	0	0	1	0	0
23	1	1	0	0	0	0	0	0	0
24	13	48	0	1	5	6	15	0	0
25	10	0	0	0	0	10	0	0	0
26	9	11	5	22	3	3	1	8	4
27	11	12	8	23	5	0	1	11	4
28	0	4	3	0	4	0	0	0	1
29	2	0	0	0	0	2	0	0	0
30	1	1	0	0	0	0	1	0	0
31	1	9	0	0	5	0	3	0	0
32	0	10	0	0	8	0	4	0	0
33	0	2	0	0	1	0	1	0	0
34	1	3	0	0	0	0	0	0	1
35	3	13	0	3	9	2	0	0	0
36	3	3	0	0	0	0	0	0	0
37	2	6	0	0	1	0	0	0	0
38	1	7	0	0	1	0	1	0	0
39	0	2	0	0	0	0	1	0	0
40	0	2	0	0	5	0	0	0	0

Table 4 (continued)

	Number of CFA Models		Unsuccessful Model Estimation	Problematic Factor Loadings[1, 2]		High Correlation[3]		Implausible Estimates[2, 4]	
Dataset	Acc.	Rej.	Rej.	Acc.	Rej.	Acc.	Rej.	Acc.	Rej.
41	0	2	0	0	4	0	0	0	0
42	0	2	0	0	4	0	0	0	0
43	0	3	0	0	8	0	0	0	0
44	0	6	0	0	15	0	3	0	0
45	0	4	0	0	15	0	1	0	0
46	1	5	0	3	15	1	0	0	0
47	0	2	0	0	5	0	1	0	0
48	8	2	0	0	0	1	0	0	0
49	3	0	0	1	0	0	0	0	0
50	3	3	0	0	0	0	0	1	0
51	0	2	0	0	6	0	0	0	3
52	0	4	0	0	11	0	1	0	4
53	2	0	0	0	0	0	0	0	0
54	0	2	0	0	8	0	0	0	1
55	0	4	0	0	2	0	0	0	0
56	0	2	0	0	0	0	0	0	0
57	0	2	0	0	0	0	0	0	0
58	2	0	0	0	0	0	0	0	0
59	1	1	0	0	0	0	0	0	0
60	1	1	0	0	0	0	0	1	0
61	0	2	0	0	2	0	0	0	0
62	8	27	0	1	45	3	3	0	2
63	1	3	0	1	5	0	0	0	0
64	0	2	0	0	2	0	1	0	0
65	0	2	0	0	5	0	0	0	1
66	15	10	0	1	0	2	3	0	1
67	1	2	0	0	2	1	0	0	0
68	0	3	0	0	0	0	1	0	3
69	2	2	0	0	0	0	0	0	0
70	0	5	0	0	0	0	5	0	1
71	0	3	0	0	1	0	0	0	2
72	4	2	0	0	1	1	0	0	0
73	1	5	0	0	2	0	0	0	0
74	0	6	0	0	16	0	0	0	0
75	1	5	0	0	7	0	0	0	0
76	4	0	0	0	0	0	0	0	0
77	3	1	0	0	0	0	0	0	0
78	4	0	0	0	0	0	0	0	0
Totals	204	382	69	91	255	72	71	42	46

Note. [1] Meaning low (|standardized factor loading| < .40) or non-significant factor loadings. [2] As several cases of problematic parameter performance might have occurred per target, single CFA models can be counted more than once. [3] High factor correlations occurred maximum once per target. [4] Implausible estimation values (i.e., Heywood cases) include e.g., |standardized factor loadings| > 1, factor correlations |r| > 1 and/or negative residual variances.

Table 5

Performance of Selected Items Separated by Poor (Above) and Good (Below) Performance

		|Factor Loading| < 0.40 (%)			Non-sig. Factor Loading (%)		
Item	% used[1]	Accepted	Rejected	Total	Accepted	Rejected	Total
Competitive[2]	17.0	2.04	27.55	29.59	1.02	13.27	14.29
Concerned with appearance	5.0	3.70	81.48	85.19	0.00	44.44	44.44
Confident[2]	38.0	0.44	9.33	9.78	0.89	8.44	9.33
Opinionated	5.0	3.70	77.78	81.48	3.70	33.33	37.04
Persistent	3.0	0.00	16.67	16.67	0.00	11.11	11.11
Educated	4.0	0.00	0.00	0.00	0.00	0.00	0.00
Good-natured[2]	41.0	0.00	0.00	0.00	0.00	0.41	0.41
Likeable	10.0	0.00	0.00	0.00	0.00	0.00	0.00
Skillful[2]	31.0	0.00	0.55	0.55	0.00	0.55	0.55
Trustworthy[2]	24.0	0.00	0.00	0.00	0.00	0.00	0.00

Note. [1] of 586 tested CFA Models. [2] Items used in Fiske et al. (2002).

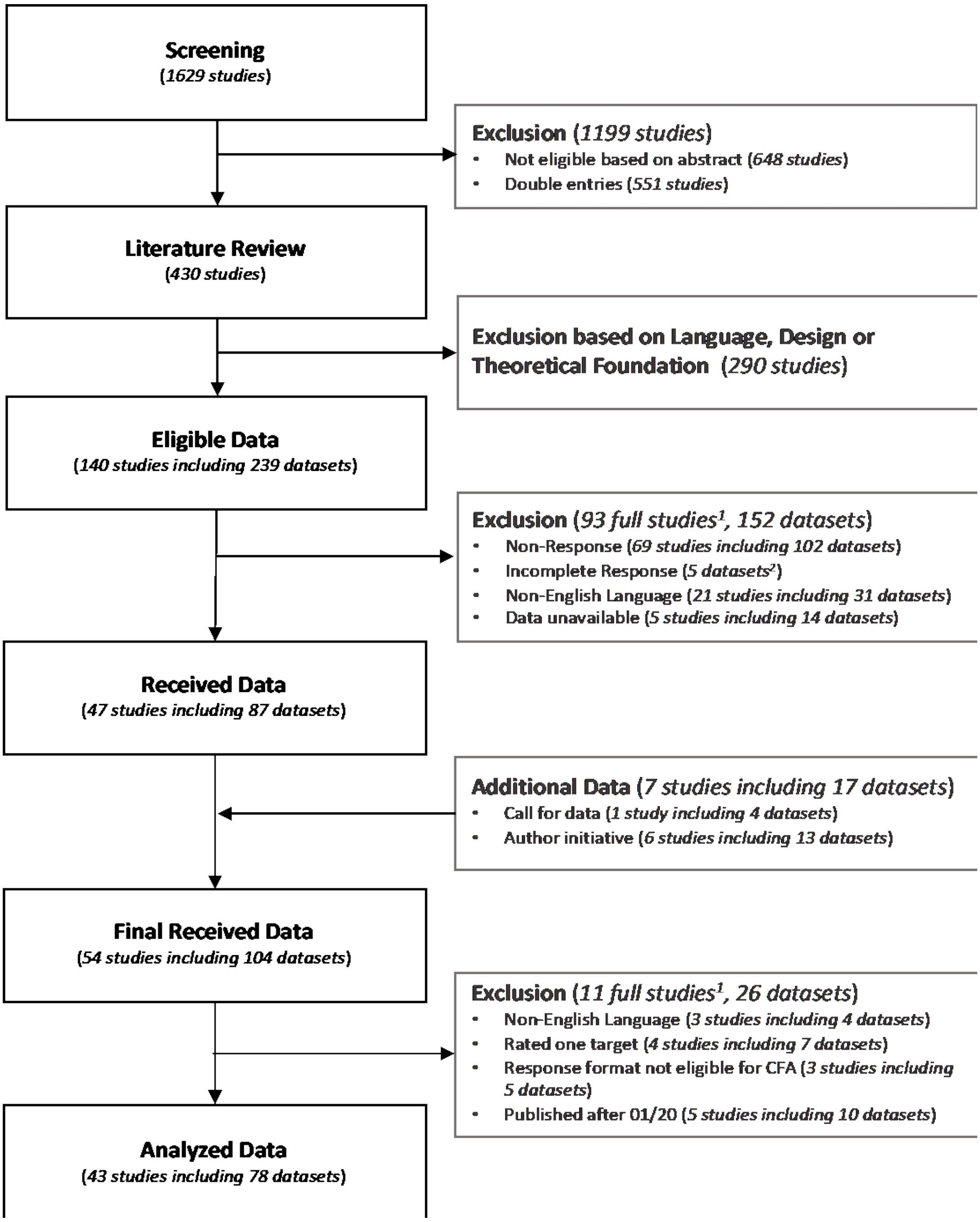

Screening
(1629 studies)
Exclusion (1199 studies)
• Not eligible based on abstract (648 studies)
• Double entries (551 studies)
Literature Review
(430 studies)
Exclusion based on Language, Design or Theoretical Foundation (290 studies)
Eligible Data
(140 studies including 239 datasets)
Exclusion (93 full studies¹, 152 datasets)
• Non-Response (69 studies including 102 datasets)
• Incomplete Response (5 datasets²)
• Non-English Language (21 studies including 31 datasets)
• Data unavailable (5 studies including 14 datasets)
Received Data
(47 studies including 87 datasets)
Additional Data (7 studies including 17 datasets)
• Call for data (1 study including 4 datasets)
• Author initiative (6 studies including 13 datasets)
Final Received Data
(54 studies including 104 datasets)
Exclusion (11 full studies¹, 26 datasets)
• Non-English Language (3 studies including 4 datasets)
• Rated one target (4 studies including 7 datasets)
• Response format not eligible for CFA (3 studies including 5 datasets)
• Published after 01/20 (5 studies including 10 datasets)
Analyzed Data
(43 studies including 78 datasets)

Figure 1. Depiction of the Data Collection Process

Note. [1] Different datasets pertaining to the same study were in some cases excluded for different reason; thus, the studies numbered for the different reasons do not add up exactly to the overall number of excluded studies. [2]This refers to cases in which we did not receive all datasets from one study. We re-analyzed the datasets we received, and excluded those that we did not receive.

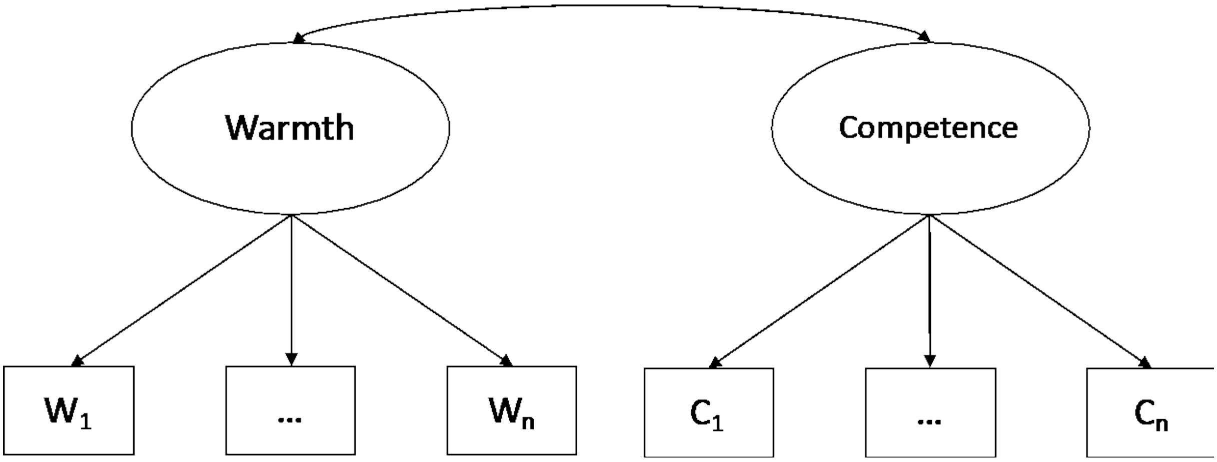


Figure 2. Exemplary Factor-Analytical Model of Stereotype Content Data

Note. Rectangles depict observed items, ellipses depict latent factors. W = Warmth, C = Competence.

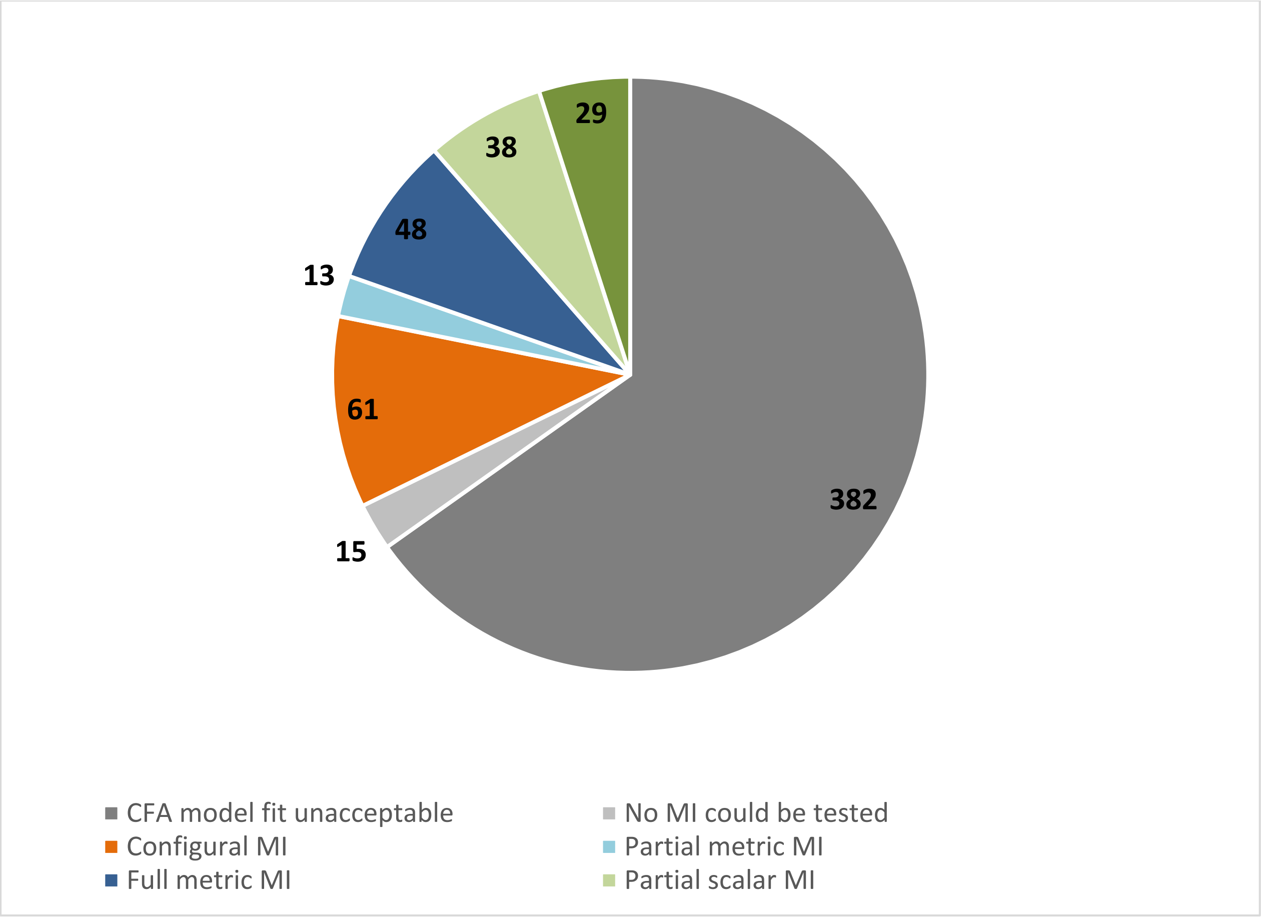


Figure 3. Highest level of established measurement invariance based on CFA models.

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 is 586.

Supplementary Materials for Manuscript # 1

The supplementary materials for Manuscript # 1 are stored in the Open Science Framework, see https://osf.io/srh36/.

They include:

- Example Mplus syntax for the CFA and measurement invariance assessment, as referred to in the preregistration;
- The document OSM – 1, containing the parameter performance results of the (partial) scalar measurement invariant models;
- The document OSM – 2, containing the parameter performance results for each warmth and competence indicator used in our re-analysis.

Preregistration for Manuscript # 2

The preregistration was competed using an AsPredicted template on May 21, 2019.

OSF preregistration link: https://osf.io/486h7/?view_only=e1b25da1084f4e248a621be36b31a153

Have any data been collected for this study already?

- **Yes, at least some data have been collected for this study already**
- No

The described project is a re-analysis of existing and published data. All data have already been collected.

What's the main question being asked or hypothesis being tested in this study?

Do measures of warmth and competence (based upon the Stereotype Content Model (SCM) by Fiske, Cuddy, Glick and Xu (2002) of different social groups in Germany allow for latent mean value comparisons by holding up to at least partial scalar measurement invariance?

Describe the key dependent variable(s) specifying how they will be measured.

For each social group, the perceived warmth and competence measures will be used. Warmth is represented by items such as good-natured, warm and likeable, and competence by items such as competent, independent and competitive. As we are conducting a reanalysis of existing data, there might be variation in the number of items per scale and their formulation. We will consider the differences in our analysis.

How many and which conditions will participants be assigned to?

None.

Specify exactly which analyses you will conduct to examine the main question/hypothesis.

In order to pursue our research question, we will obtain as many German datasets as possible that contain measures of the SCM dimensions warmth and competence by contacting corresponding authors of SCM studies conducted in Germany using German items with German samples and published until April 2019. To be included in our study, comparisons of different social groups within data sets, and/or the same social group across data sets or conditions should be possible. We deem this to be given when the same indicators to measure warmth and competence have been used. Since we want to conduct a test for measurement invariance, we will only include studies that measure warmth and competence using at least two items per scale. We will conduct a reanalysis of these data under the new aspect of the scales' measurement properties. Whereas most of the original studies focused on the content of stereotypes or their relationship to other constructs, we aim to investigate the methodical question of whether the scales to assess warmth and competence for different social groups allow for latent mean value comparison by showing at least partial scalar measurement invariance (a) within and (b) across data sets/conditions. For each data set, we will conduct a step-up approach of testing measurement invariance (Brown, 2015). The measurement model is specified by a latent warmth and competence factor predicting the corresponding warmth and competence items, respectively. No cross-loadings or covarying indicator residuals are allowed. The factors warmth and competence may correlate with each other. The robust maximum likelihood estimator will be used. The procedure for testing both research questions is as following:

1. For all social groups/conditions assessed, the model fit will be individually evaluated. A model fit of $\chi 2/df \leq 3$, RMSEA ≤ 0.08, SRMR ≤ 0.10 and CFI ≥ 0.95 is deemed satisfactory (Schermelleh-Engel, Moosbrugger, & Müller, 2003). For groups/conditions that do not fulfill one or more of these criteria, the model will be rejected and they will be excluded from subsequent analyses.
2. Equal form (i.e., configural invariance) will be evaluated by testing the model fit for all groups/conditions simultaneously. If the model fit is not satisfactory according to the above-mentioned criteria, groups/conditions with the highest contribution to the Chi-Square index at the configural model will be stepwise excluded from the analyses until the model fit is acceptable.
3. The remaining groups will then be tested for equal factor loadings (i.e., metric invariance). This is achieved if the model fit remains satisfactory and additionally does

not differ significantly from the more freely-estimated equal form model, i.e. the Santorra-Bentler corrected χ2-difference test is not significant ($p > 0.05$) and based upon Chen (2007) for N ≤ 300 changes are ≥ -0.005 for CFI, ≤ 0.010 for RMSEA, ≤ 0.025 for SRMR and for N > 300 changes are ≥ -0.010 for CFI, ≤ 0.015 for RMSEA, ≤ 0.030 for SRMR.

4. If these criteria apply, the groups will be tested for equal intercepts (i.e., scalar invariance), applying the same criteria as above except for SRMR, which will be satisfactory if changes are ≤ 0.005 for N ≤ 300 and ≤ 0.010 for N > 300 (Chen, 2007). If the model does not hold up to either the equal factor loadings or intercepts presumption, it will be tested for partial measurement invariance. At least two indicators per latent factor have to stay constrained to equality. Moreover, the freed parameters must not vary between groups/conditions. For models that don't hold up to (partial) metric measurement invariance (i.e. equal loadings), equal intercepts cannot be tested for this model. If partial measurement invariance is not reached, step-wise exclusion of groups/conditions as outlined above will be performed to test if there is a subset of groups that does hold up to partial measurement invariance. In one of the datasets (Kotzur, Friehs, Asbrock, & van Zalk, 2019), measurement invariance has already been investigated, but using a different approach. All other data sets have not been analyzed with regard to factorial structure and measurement invariance (unless reported in the original publications).

Any secondary analyses?

n/a

How many observations will be collected or what will determine sample size?

No need to justify decision, but be precise about exactly how the number will be determined.

We will reanalyze existing data sets in Germany.

Anything else you would like to pre-register?

(e.g., data exclusions, variables collected for exploratory purposes, unusual analyses planned?)

Literature

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (2nd Edition). The Guilford Press. New York.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

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. https://doi.org/10.4324/9781315187280

Kotzur, P. F., Friehs, M., Asbrock, F., & van Zalk, M. H. W. (2019). Stereotype Content of Refugee Subgroups in Germany. *European Journal of Social Psychology*. https://doi.org/10.1002/ejsp.2585

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* (Vol. 8). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.509.4258&rep=rep1&type=pdf

Manuscript # 2

A Preregistered Examination of Scale Properties of Stereotype Content Measures: The German Case

Maria-Therese Friehs[1]*, Patrick F. Kotzur[2]*, Ann-Kristin C. Zöller[3], Ulrich Wagner[4], and Frank Asbrock[5]

[1] FernUniversität in Hagen, Germany; ORCID: https://orcid.org/0000-0002-5897-8226

[2] Durham University, United Kingdom; ORCID: https://orcid.org/0000-0002-5193-3359

[3] Hannover Medical School, Germany

[4] Philipps-University Marburg, Germany; ORCID: https://orcid.org/0000-0001-6716-9212

[5] Chemnitz University of Technology, Germany; ORCID: https://orcid.org/0000-0002-6348-2946

* First authorship: Maria-Therese Friehs and Patrick F. Kotzur share the first authorship.

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

The authors declare no conflict of interests with respect to the authorship or the publication of this article. The authors thank Johanna Böttcher, Rebecca Cramer, Annemarie Juchler, Vera Kaiser, Adelina Kletke, Corinne Lamadé, Julia Nolting, and Tobias Schmidt for their assistance in preparing this manuscript.

Data availability statement: The used data stem from a reanalysis of published data and should be requested from the authors of the original publications. Detailed information on all analyses conducted in this article are available upon request from the corresponding author of this article.

Corresponding author: Maria-Therese Friehs, Department of Psychology, FernUniversität in Hagen, Universitätsstraße 33, 58097 Hagen, Germany, maria-therese.friehs@fernuni-hagen.de

Preregistration:

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

Manuscript:

Total word count: 5,654

Abstract: 237

Tables and Figures-Document:

Tables: 2

Figures: 2

Online Supplementary Materials:

3 documents, labelled “OSM-A to OSM-C”, “OSM-D1”, “OSM-D2”

Abstract

The stereotype content model (SCM), which defines warmth and competence as fundamental dimensions of social perception, plays a prominent role in contemporary research. Recently, researchers suggested that the SCM scales currently utilised in English contexts might perform less well than previously assumed (Friehs et al., 2021). This was particularly the case when it came to meeting prerequisites for mean-value comparisons, which are the kinds of analyses that SCM scales are mostly submitted to. We build on this research by investigating the scale properties of SCM measures in the German language context. Thus, we investigated the reliability, dimensionality and cross-target group measurement equivalence of German SCM scales in 29 published data sets (*N* = 10,854) using a preregistered analysis protocol. Confirmatory factor analyses of 507 SCM measurement models showed that the reliability of the used scales was on average good and that they showed adequate dimensionality in 35.10 % of all cases. We additionally assessed (partial) scalar measurement equivalence as a prerequisite for meaningful mean-value comparisons and found evidence for it in 11.44% of all cases. Our findings echo those from the English context and indicate that the currently utilised German scales perform less well than we would have hoped. Moreover, our findings contribute to a debate about how to measure stereotype content, and we call on all researchers to invest in scale development efforts to ensure highly reliable and valid social perception research in Germany and elsewhere.

Keywords: Stereotype Content Model, *Construct Validity, Confirmatory Factor Analysis, Measurement Equivalence, Germany*

A Preregistered Examination of Scale Properties of Stereotype Content Measures: The German case

The Stereotype Content Model (SCM; Fiske, Cuddy, Glick, & Xu, 2002) proposes that perceptions of warmth (i.e., benign vs. hostile intent) and competence (i.e., (in-)ability to enact intent) constitute fundamental dimensions for the evaluation of individuals, social groups, and cultures (Cuddy, Fiske, & Glick, 2008). According to the model, perceived warmth and competence are predicted by competition/threat and status, respectively (Fiske et al., 2002; Kervyn, Fiske, & Yzerbyt, 2015). The interplay of warmth and competence is theorised to evoke distinct emotional and behavioural reactions (Cuddy, Fiske, & Glick, 2007).

The SCM enjoys high popularity among social perception researchers*,* impressively demonstrated by about 3000 citations of the seminal study by Fiske et al. (2002) on Web of Science as of May 2021. In line with this high interest, there are currently a number of ongoing methodological and conceptual debates on different aspects of the model, including the adequate number of fundamental dimensions of social perception (e.g., Abele, Ellemers, Fiske, Koch, & Yzerbyt, 2021; Brambilla, Sacchi, Rusconi, & Goodwin, 2021; Koch, Imhoff, Dotsch, Unkelbach, & Alves, 2016; Koch, Yzerbyt, Abele, Ellemers, & Fiske, 2021; Leach, Ellemers, & Barreto, 2007; Stanciu, 2015), and how to most effectively measure stereotype content (e.g., Friehs et al., 2021; Halkias & Diamantopoulos, 2020; Kotzur et al., 2020). We aim to contribute to these debates by focusing on the German context.

In psychological research, scales are typically developed following established procedures to make sure that, among other things, researchers measure the construct they intend to measure (also referred to as construct validity; e.g., Brown, 2015) and to ensure that the scales fulfil the preconditions for submitting them to the kinds of analyses they intend to (Flake & Fried, 2020; Flake, Pek, & Hehman, 2017; Vandenberg & Lance, 2000). In the case of the SCM, Fiske, Xu, Cuddy, and Glick (1999) developed an initial scale by asking participants to rate 17 social groups on 27 trait adjectives that they selected based on a comprehensive literature review. Next, they performed a series of oblique exploratory factor analyses for each of the groups separately. Based on these analyses, they selected ten traits that they deemed to capture warmth and competence most consistently across target groups[1]. The results of the exploratory factor analysis supported this selection most of the time; in some instances, however, more than two dimensions emerged. In the seminal follow-up study most subsequent SCM works have cited, Fiske et al. (2002) used a subset of these items, whereby the specific subset varied from study to study. Research building on Fiske et al. (1999, 2002) continued to use these items, while also flexibly amending new or excluding existing items, oftentimes without providing a rationale for these decisions or formally assessing scale properties.

The limitations associated with the scale development procedure Fiske and colleagues (1999, 2002) used and with the common practice of subsequent SCM researchers to add and discard items are thoroughly documented elsewhere (Friehs et al., 2021; Halkias & Diamantopoulos, 2020). In brief, these authors argue that the SCM measures have not been developed according to widely recommended and available statistical procedures (e.g., Bandalos, 2018; Brown, 2015). What is more, they argue that a formal test of the performance of the scale would make sure that authors measure what they intend to measure (construct validity), that the underlying construct is indeed two-dimensional (one warmth factor, one competence factor; dimensionality), and that the prerequisites for the analytical procedures authors submit the scales to in order to produce valid and reliable results are met (Friehs et al., 2021).

A formal examination of such features requires analyses in a confirmatory analysis framework and includes the investigation of the factor structure (i.e., dimensionality) and measurement equivalence (Vandenberg & Lance, 2000). In confirmatory factor analysis, the researcher defines a theoretical model, for which they pre-specify the relations between model parameters, including latent factors (i.e., warmth, competence) and observed indicators (i.e., warmth and competence items). In these so-called measurement models, it is then tested to what extent the theoretical model matches empirical reality, which allows for conclusions regarding the scales' dimensionality and construct validity. In measurement equivalence tests, the researcher formally compares model parameters of confirmatory factor analysis models across multiple target groups. Measurement equivalence tests whether the used scale indeed measures the two theoretically expected SCM dimensions warmth and competence consistently across investigated target groups (configural measurement equivalence; speaks to identical dimensionality across groups), whether the scale can be used for comparative correlational analyses between target groups (metric measurement equivalence), and finally, whether warmth and competence means of different target groups can be meaningfully compared with one another without bias (scalar measurement equivalence; cf. Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014; Vandenberg & Lance, 2000). One type of measurement equivalence builds on the other; if configural measurement equivalence is not found, metric and scalar measurement equivalence cannot be established either. Similarly, metric equivalence needs to be established before scalar equivalence can be tested. Most SCM applications make use of warmth and competence mean values to determine whether a particular target group or cluster of target groups is perceived different in terms of warmth/competence levels compared to another one; an operation that requires scalar measurement equivalence.

Friehs and colleagues (2021) recently delivered a formal test of scale performance for English SCM scales. In their reanalysis of 78 data sets from 43 published articles over the last two decades, the authors examined various characteristics of scales SCM researchers have used to date, including the scales' reliability, dimensionality and measurement equivalence. Overall, the measurement seemed to be reliable when inspecting McDonald's omegas as internal consistency indicator. This means that the measurement tended to be consistent across the board. However, the authors found evidence for the theoretically expected two-dimensional structure for sobering 34.81% of the 586 target group-specific SCM measurement models. Additionally, when checking whether mean values of warmth and competence could be validly compared across target groups within each of these data sets (scalar measurement equivalence), this was the case for 11.43% of all target groups that the studies reported on. Thus, the scales currently utilised in English contexts appear to perform less well than previously thought, particularly when it comes to meeting dimensionality expectations and the prerequisites for the kinds of analyses that these scales were typically used for (scalar measurement equivalence).

These results have many important implications for SCM research. To date, it remains unclear to what extent existing studies succeeded to measure what they intended to measure (construct validity), what the adequate number of fundamental dimensions of social perceptions might be (dimensionality; see also Brambilla et al., 2021; Koch et al., 2016; Leach et al., 2007; Stanciu, 2015), and how to most effectively measure stereotype content in general (see also Halkias & Diamantopoulos, 2020; Kotzur et al., 2020). This is because “the verity of results about a psychological construct hinges on the validity of its measurement” (Flake et al., 2017, p. 370), for which we – as of today – might not have sufficient evidence when it comes to SCM scales.

Recent meta-scientific discussions have pointed out both a general need for replication of (social) psychological results (e.g., Earl & Trafimow, 2015; Schimmack, 2020) and a need to consider context as an important influence on research findings (e.g., Pettigrew, 2018, 2021). Thus, Friehs and colleagues (2021)’s findings based on English scales require replication for SCM scales used in other country and language contexts. One context in which the SCM has gained popularity among researchers (us included) is the German context. The first SCM publication by Eckes (2002) based the analyses on a German translation of the items proposed by Fiske et al. (1999). A subset of these items has been used in later work (e.g., Asbrock, 2010; Hansen, Rakic, & Steffens, 2017, 2018; Hellmann, Berthold, Rees, & Hellmann, 2015; Kemme, Essien, & Stelter, 2020; Kotzur, Friehs, Asbrock, & van Zalk, 2019; Kotzur, Forsbach, & Wagner, 2017; Kotzur, Schäfer, & Wagner, 2019; Kotzur et al., 2020). Whereas the majority of these studies examined scale performance based on exploratory approaches (e.g., principal component analysis; Asbrock, 2010) or reliability estimates only (e.g., Eckes, 2002; Hansen et al., 2017, 2018; Hellmann et al., 2015; Kotzur et al., 2017; Kotzur, Schäfer et al., 2019), less than a handful of German studies reported results from confirmatory approaches suited to fully assess the relevant aspects of scale performance previously discussed (Hackbart, Rapior, & Thies, 2020; Kotzur, Friehs et al., 2019; Kotzur et al., 2020). Based on their scale performance assessment, the authors of these few studies mostly concluded that the warmth and competence items did not perform as well as they would have liked for some of the target groups they investigated, which limited their main analyses (Kotzur, Friehs et al., 2019; Kotzur et al., 2020). Given these initial indications, it is imperative to gain a better understanding of the status quo of the performance of German-language SCM scales.

Building on Friehs et al.'s (2021) research, our research questions (RQ) therefore are:

> RQ1: Do German warmth and competence indicators that have been used in previous research form valid and reliable SCM scales?
>
> RQ2: Do German SCM scales that have been used in previous research allow for valid interpretation of warmth and competence mean-value comparisons as presented in the original publications?

To answer these questions, we conducted a systematic re-analysis of 29 data sets collected in German(y) stemming from 23 SCM papers.

Methods

Our analyses were preregistered on OSF (https://osf.io/486h7/?view_only=870dbff75b004b29aeffae7d27c62518). Our eligibility criteria were similar to those of Friehs et al. (2021): (I) To investigate measurement equivalence within the data sets of original publications, data sets needed to include at least two target groups, be it within the same sample, or across two sub-samples (e.g., experimental groups); (II) Since any assessment of reliability and validity requires multiple items, warmth and measure needed to be assessed with more than one item each; (III) Since we used a latent variable approach that can be sensitive to sample size (Kenny, Kaniskan, & McCoach, 2015), sample sizes of the data sets needed to be $N \geq 50$[2].

We scanned academic search engines (PsycInfo, PSYNDEX, Google Scholar) for eligible data sets published until Mid-August 2020[2] and sent out calls for data via the mailing list of the German Psychological Society (DGPs). We identified 32 eligible publications and finally analysed 23 publications including 29 data sets (for details regarding the data inclusion, see Figure 1 and OSM-B). We are very thankful for this great resonance and support from the SCM community. The publications included in the re-analyses are listed in Table 1 and marked with an asterisk in the references.

- Figure 1 about here -

The modelling process followed the procedure suggested by Friehs et al. (2021) and is explained in detail in OSM-C. First, we conducted confirmatory factor analyses for each target group or sample using the SCM scale information described in the original publications. As such, we specified one warmth factor predicting the mentioned warmth indicators, and one competence factor predicting mentioned competence indicators. The two factors were allowed to correlate. If scale development and applications in German contexts were successful, and if the items indeed measured the postulated two dimensions of stereotype content, such a model should fit the data well. We accepted models that fulfilled the model fit criteria by Schermelleh-Engel, Moosbrugger and Müller (2003, see also OSM-C). Confirmatory factor analysis models with non-acceptable model fit did not support the claim that the used items form valid two-dimensional SCM scales (RQ1) and were therefore discarded from further analyses. We did, however, compute McDonald's ω_{total} for warmth and competence for all successfully estimated confirmatory factor analysis models to be able to speak to the reliability for both of these SCM dimensions for all target groups.

Second, for each data set separately, we tested measurement equivalence including all confirmatory factor analysis models that had acceptable model fit using multiple-group confirmatory factor analysis (Byrne, Shavelson, & Muthén, 1989) with increasingly restrictive, nested models (Steenkamp & Baumgartner, 1998). We first tested all accepted confirmatory factor analysis models per data set for equal form (i.e., configural measurement equivalence), that is, whether the number of factors and the factor loading patterns are comparable across confirmatory factor analysis models. If model fit was acceptable, we introduced equality restrictions for factor loadings of identical indicators across target group-specific confirmatory factor analysis models (i.e., metric measurement equivalence). Metric measurement equivalence implies equal warmth and competence measurement units across confirmatory factor analysis models and is a precondition for (latent) correlational/regression-based analysis (Steenkamp & Baumgartner, 1998), such as predicting emotions and behavioural tendencies for different target groups from warmth and competence (like in Kotzur, Schäfer, et al., 2019; see also Cuddy et al., 2007). Metric measurement equivalence 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 measurement equivalence models, we added equality restrictions to indicator intercepts of identical indicators across confirmatory factor analysis models (i.e., scalar measurement equivalence). Scalar measurement equivalence implies equal points-of-zero (i.e., equal item difficulty) of similar SCM indicators across target group-specific confirmatory factor analysis models and forms the precondition for warmth and competence mean-value comparisons (Steenkamp & Baumgartner, 1998) across target groups (like in Asbrock, 2010) or samples (like in Kotzur et al., 2017). Scalar measurement equivalence 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 measurement equivalence was not achieved, we aimed at improving model fit by introducing partial measurement equivalence (Byrne et al., 1989). Partial measurement equivalence allows some exceptions from the equality constraints of measurement properties across confirmatory factor analysis models, and thus the equality assumption is somewhat limited, but still generally accepted (Davidov et al., 2014). We identified eligible constraints using modification indices and introduced partial measurement equivalence by releasing equality constraints of factor loadings and/or indicator intercepts on the preconditions that for at least two indicators per factor, all parameters remained equal across confirmatory factor analysis models (Davidov et al., 2014). If introducing metric or scalar partial measurement equivalence still resulted in unacceptable model fit, we excluded the confirmatory factor analysis model with the highest χ^2-value contribution in the fully constrained measurement equivalence model from analysis, and repeated the entire process, always aiming to reach (partial) scalar measurement equivalence as a requirement for RQ2.

Results and Discussion

For details concerning the included data sets (e.g., target groups, scale wording, sample information), see Table 1. Summary information concerning the share of acceptable CFA models and measurement equivalence performance per data set is presented in Table 2. We provide detailed information concerning the data set- and target group-/sample-specific model fits, reliability and measurement equivalence assessment processes in OSM-D1 and OSM-D2.

- Tables 1 and 2 about here –

The first research question addressed the extent to which German warmth and competence indicators that have been used in previous research form valid and reliable scales (RQ1). Therefore, we applied confirmatory factor analyses to the SCM items of *K* = 507 target groups to test the measurement models as proposed by the original publications. Overall, *k* = 178 CFA models from 20 original publications achieved acceptable model fit, indicating that we found evidence for the theoretically expected two-dimensional structure for 35.10% of all CFA models. This finding also implies that the measurement of warmth and competence was conflicting with theoretical expectations in more than 64% of all cases, which were distributed across 21 out of 23 original publications. Scale reliability across *k* = 497 models (excluding 10 models with implausible parameter estimates or which did not converge) was $M\omega_{\text{total}}$ = .849, $SD\omega_{\text{total}}$ = .068 (min = .553, max = .969) for warmth and $M\omega_{\text{total}}$ = .809, $SD\omega_{\text{total}}$ = .078 for competence (min = .474, max = .969).

We also assessed measurement equivalence to determine the extent to which the used SCM scales allowed for meaningful warmth and competence comparisons as presented in the original publications (RQ2). We checked this only in those cases in which the two-dimensional structure was found (see RQ1). We inspected measurement equivalence for *k* = 160 target groups from 17 data sets (given that measurement equivalence assessments require the comparison of at least two acceptable confirmatory factor analysis models per comparison, *k* = 18 models from 12 data sets had to be excluded due to lack of possible comparisons). The highest level of measurement equivalence we found was full scalar measurement equivalence for *k* = 19 target groups from seven data sets, partial scalar measurement equivalence for *k* = 39 target groups from eight data sets, full metric measurement equivalence for *k* = 35 target groups from nine data sets, and partial metric measurement equivalence for *k* = 50 target groups from seven data sets. Consequently, a valid comparison of warmth and competence scores (i.e., partial or full scalar equivalence) was given for 11.44% of all target groups. A summary of our findings is depicted in Figure 2.

- Figure 2 about here –

These findings demonstrate that the German scales performed less well than we would have hoped. Our test of the extent to which German warmth and competence indicators that have been used in previous research form valid and reliable scales (RQ1) indicates that the German scales were, similarly to Friehs et al. (2021)'s study on English scales, on average highly internally consistent – a feature that has oftentimes been checked by authors (e.g., Kotzur et al., 2017; Kotzur, Schäfer et al., 2019). However, we found limited evidence for the expected two-factorial dimensionality of the scales, which also replicates Friehs et al. (2021)'s findings with a surprisingly high level of congruence. This indicates that the assumption that the underlying construct is two dimensional (one warmth factor, measured by suggested warmth indicators; one competence factor, measured by suggested competence indicators) was not often met. Ironically, our results are thus also compatible with earliest SCM work (Fiske et al., 1999), which also did not consistently find the two theorised stereotype content dimensions using exploratory methods. Moreover, we found that German SCM scales that have been used in previous research oftentimes do not allow for valid interpretation of warmth and competence mean-value comparisons (RQ2). This implies that most mean-value comparisons of target groups on the SCM scales in the re-analysed data would result in "comparing apples with oranges" (Davidov et al., 2014).

This means we might be in trouble, because our results suggest that the validity issues that are associated with the findings and that have been raised with regard to the English-speaking scales (Friehs et al., 2021) extend to the German case. Specifically, to say it with Flake et al. (2017), the inability to establish acceptable warmth and competence scales in accordance with the theory means that “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 into the primary results“ (Flake et al., 2017, p. 370). In other words, the results of our investigation of measurement properties (i.e., reliability, dimensionality and measurement equivalence) indicate that we ultimately cannot be sure if what we measured with SCM scales is what we intended to measure (i.e., one warmth and one competence factor). This issue raises questions about construct validity of the scales, which in turn raises questions about the validity of the conclusions we can draw based on these measures (i.e., internal validity). Indeed, even when a two-dimensional structure of SCM scales could be found, the measurement equivalence results indicate that this did not automatically mean that the scales could be used for the kinds of analyses that these scales are typically used for (i.e., mean-value comparisons). For further discussion of the potential reasons and implications, we refer readers to Friehs et al. (2021), who present an extensive discussion of their strikingly similar findings in the English language context.

Although some of these issues have been criticised before (Friehs et al., 2021; Halkias & Diamantopoulos, 2020; Kotzur, Friehs et al., 2019; Kotzur et al., 2020), this is the first systematic examination of this topic that illustrates the full extent in the German language context. A further strength of our research is that we used a pre-registered analytical procedure to re-analyse about 70% of all existing German publications, presenting a thorough and comprehensive assessment of measurement properties. This induces us with confidence that our results are robust and generalizable within the context under scrutiny.

One takeaway from this is that checking scales’ reliability is a good start, yet not enough to fully evaluate measurement quality (Flake et al., 2017). Moreover, our results add fuel to the debate of what the most adequate number of fundamental dimensions of social perceptions might be. Whereas the SCM argues for two (Fiske et al., 2002) and sometimes empirically finds more (e.g., Fiske et al., 1999), others argued for three (Brambilla et al., 2021; Koch et al., 2016; Leach et al., 2007), and some even for more dimensions (Abele et al., 2021; Stanciu, 2015). Additionally, our findings contribute to the debate of how to most effectively measure stereotype content (Friehs et al., 2021; Halkias & Diamantopoulos, 2020; Kotzur et al., 2020).

Nonetheless, we need to acknowledge limitations. For instance, we employed the most commonly used multigroup confirmatory factor analysis approach to assess measurement equivalence. Despite its popularity, some researchers have argued that this exact method might yield less favourable results compared to other more liberal approaches that allow for small differences in parameters (i.e., approximate measurement equivalence; Davidov et al., 2014). Although these alternative methods have their own limitations (e.g., required knowledge about priors), future studies could explore these applications as one future avenue.

Moreover, although our models showed relatively low model complexity and we excluded data sets with $N < 50$, other researchers recommend sample sizes larger than our minimum cut-off criterion for latent variable modelling (e.g., Boomsma, 1985). Indeed, it has been reported that low sample sizes negatively impact model fit (e.g., Kenny et al., 2015). Introducing stricter inclusion criteria based on sample size may, however, also have biased results, in a way that a smaller share of all available data and used scales could have been re-analysed. To investigate this issue, we computed Pearson's correlations to relate sample-size to rates of CFA models with acceptable model fit per sample. Though we identified a descriptive pattern that studies with higher sample sizes showed somewhat higher rates of acceptably fitting CFA model, this trend was non-significant, $r(28) = .162$, $p = .409$. So even among the CFA models with high sample-size, model-fits were often unacceptable (but see Friehs et al., 2021, who did find evidence for a positive trend).

Despite these limitations, we are convinced of the relevance and critical impact of our findings on SCM research. If we were to follow other scholars' advice (e.g., Flake et al., 2017), the limited evidence for construct validity – or that we measure what we actually intended to measure – questions whether valid claims about warmth and competence can currently be made in the German context. Even if we were to support less drastic interpretations, at the very least, our results question the usability of currently utilised German SCM scales.

A critical question we as the scientific community need to ask ourselves is where we go from here, now that we have demonstrated that the existing SCM scales do not perform as well as we would have hoped in more than one language context. We strongly believe that, in line with Popper's (1959) ideas on the scientific process, the response to the raised issues can only be a collective one. Friehs et al. (2021) have formulated some concrete suggestions that we all can try to implement the next time we plan a study or collect and analyse data. These steps include joining forces to develop new measures with the aim to validly and reliably capture warmth and competence (for a diligent example in the German context, see e.g., Halkias & Diamantopoulos., 2020), routinely testing the SCM's measurement properties before main analyses using CFA, and, if the research question dictates it, assessing measurement equivalence across target groups or samples to be compared. We strongly believe that these measures could help us to move the field forward. Thus, we emphatically appeal to all SCM researchers to consider these suggestions and make the measurement of SCM and discussions about these issues a priority.

Footnotes

[1] By target groups, we refer to all kinds of stimuli that have been evaluated and compared on warmth and competence dimensions (e.g., social groups, cultures).

[2] This aspect slightly deviates from the original preregistration. For a detailed explanation, see OSM-A.

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. https://doi.org/10.1037/rev0000262

*Asbrock, F. (2010). Stereotypes of social groups in Germany in terms of warmth and competence. *Social Psychology, 41*(2), 76-81. https://doi.org/10-1027/1864-9335/a000011

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

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika, 50*, 229-242. https://doi.org/ 10.1007/BF02294248

Brambilla, M., Sacchi, S., Rusconi, P., & Goodwin, G. P. (2021). The primacy of morality in impression development: Theory, research, and future directions. *Advances in Experimental Social Psychology.* Advance online publication.

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

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466. 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, 14*(3), 464-504. https://doi.org/10.1080/10705510701301834

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology, 92*(4), 631-648. https://doi.org/10.1037/0022-3514.92.4.631

Cuddy, A. j. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology, 40,* 61-149. https://doi.org/10.1016/S0065-2601(07)00002-0

*Cuddy, A. J. C., Fiske, S. T., Kwan, V. S. Y., Glick, P., Demoulin, S., Leyens, J.-P., . . . Ziegler, R. (2009). Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology, 48*(1), 1-33. https://doi.org/10.1348/014466608X314935

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

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology, 6*, 621. https://doi.org/10.3389/fpsyg.2015.00621

Eckes, T. (2002). Paternalistic and envious gender stereotypes: Testing predictions from the stereotype content model. *Sex Roles*, *47*(3-4), 99-114. https://doi.org/10.1023/A:1021020920715

*Ehrke, F., Bruckmüller, S., & Steffens, M. C. (2020). A double-edged sword: How social diversity affects trust in representatives via perceived competence and warmth. *European Journal of Social Psychology, 50*(7), 1540-1554. https://doi.org/10.1002/ejsp.2709

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. 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. https://doi.org/10.1111/0022-4537.00128

Flake, J., & 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. https://doi.org/10.1177/2515245920952393

Flake, J., 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. https://doi.org/10.1177/1948550617693063

Friehs, M.-T., Böttcher, J., Kotzur, P. F., Lüttmer, T., Wagner, U., Asbrock, F., & Van Zalk, M. H. W. (2021). *Examining the structural validity of Stereotype Content Measures – A preregistered re-analysis of published data and discussion of possible future directions.* PsyArXiv. https://doi.org/10.31234/osf.io/dej4m

*Fröhlich, L. & Schulte, I. (2019). Warmth and competence stereotypes of immigrant groups in Germany. *PLoS ONE, 14*(9), e0223103. https://doi.org/10.1371/journal.pone.0223103

*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*, 116-132. 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. https://doi.org/10.1016/j.ijresmar.2020.02.004

*Hansen, K., Rakic, T., & Steffens, M. C. (2017). Competent and warm? How mismatching appearance and accent influence first impressions. *Experimental Psychology, 64*(1), 27-36. https://doi.org/10.1027/1618-3169/a000348

*Hansen, K., Rakic, T., & Steffens, M. C. (2018). Foreign-looking native-accented people: More competent when first seen rather than heard. *Social Psychological and Personality Science, 9*(8), 1001-1009. https://doi.org/10.1177/1948550617732389

*Hellmann, J. H., Berthold, A., Rees, J. H., & Hellmann, D. F. (2015). "A letter for Dr. Outgroup": On the effects of an indicator of competence and changes of altruism toward a member of a stigmatized out-group. *Frontiers in Psychology, 6*, 1422. https://doi.org/10.3389/fpsyg.2015.01422

*Ihme, T. A., & Möller, J. (2015). „He who can, does; he who cannot, teaches?": Stereotype threat and preservice teachers. *Journal of Educational Psychology, 107*(1), 300-308. https://doi.org/10.1037/a0037373

*Imhoff, R., Woelki, J., Hanke, S., & Dotsch, R. (2013). Warmth and competence in your face! Visual encoding of stereotype content. *Frontiers in Psychology, 4*, 386. https://doi.org/10.3389/fpsyg.2013.00386

*Janda, C., Asbrock, F., Herget, M., Kues, J. N., & Weise, C. (2019). Changing the perception of premenstrual dysphoric disorder – An online-experiment using the Stereotype Content Model. *Women & Health, 59*(9), 967-984. https://doi.org/10.1080/03630242.2019.1584599

*Kemme, S., Essien, I., & Stelter, M. (2020). Antimuslimische Einstellungen in der Polizei? Der Zusammenhang von Kontakthäufigkeit und -qualität mit Vorurteilen gegenüber Muslimen [Anti-Muslim attitudes in the police force? The relationship of frequency and quality of contact with prejudice towards Muslims]. *Monatsschrift für Kriminologie und Strafrechtsreform, 103*(2), 129-149. https://doi.org/10.1515/mks-2020-2048

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. https://doi.org/10.1177/0049124114543236

Kervyn, N., Fiske, S., & Yzerbyt, V. (2015). Forecasting the primary dimension of social perception*.* *Social Psychology, 46*(1), 36-45. 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. https://doi.org/10.1037/pspa0000046

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. *Advances in Experimental Social Psychology, 63*, 1-68. https://doi.org/10.1016/bs.aesp.2020.11.001

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

*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. https://doi.org/10.1027/1864-9335/a000312

Kotzur, P. F., Schäfer, S., & Wagner, U. (2019). Meeting a nice asylum seeker: Intergroup contact changes stereotype content perceptions and associated emotional prejudice, and encourages solidarity-based collective action. *British Journal of Social Psychology, 58*(3), 668-690. https://doi.org/10.1111/bjso.12304

*Kotzur, P. F., Veit, S., Namyslo, A., Holthausen, M.-A., 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. 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. https://doi.org/10.1037/0022-3514.93.2.234

*Nett, T., Dorrough, A., Jekel, M., & Glöckner, A. (2020). Perceived biological and social characteristics of a representative set of German first names. *Social Psychology, 51*(1), 17-34. https://doi.org/10.1027/1864-9335/a000383

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

Pettigrew, T. F. (2021). *Contextual social psychology: Reanalyzing prejudice, voting, and intergroup contact.* Washington: American Psychological Association.

Popper, K. (1959). *The logic of scientific discovery.* Abington-on-Thames: Routledge.

*Renner, A.-M. (2019). Should female politicians avoid appearing emotional? Gender-specific effects of politicians' emotions on the attribution of competence and warmth. *Politische Psychologie, 7*(1), 112-132.

*Rennung, M., Blum, J., & Göritz, A. S. (2016). To strike a pose: No stereotype backlash for posing women. *Frontiers in Psychology, 7*, 1463. https://doi.org/10.3389/fpsyg.2016.01463

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. https://doi.org/10.1037/cap0000246

*Schwind, V., Deierlein, N., Poguntke, R., & Henze, N. (2019). Understanding the social acceptability of mobile devices using the SCM. *CHI 2019 Paper*. https://doi.org/10.1145/3290605.3300591

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

*Stanciu, A., Vauclair, C.-M., & Rodda, N. (2019). Evidence for stereotype accommodation as an expression of immigrants' socio-cognitive adaptation. *International Journal of Intercultural Relations, 72*, 76-86. https://doi.org/10.1016/j.ijintrel.2019.07.003

Steenkamp, J-B. E. M., & 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

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

*Winter, K., Scholl, A., & Sassenberg, K. (2020). A matter of flexibility: Changing outgroup attitudes through messages with negations. *Journal of Personality and Social Psychology*, *120*(4), 956-976. https://doi.org/10.1037/pspi0000305

*Wyszynski, M. C., Guerra, R., & Bierwiaczonek, K. (2020). Good refugees, bad migrants? Intergroup helping orientations towards refugees, migrants, and economic migrants in Germany. *Journal of Applied Social Psychology*, *50*, 607-618. https://doi.org/10.1111/jasp.12699

Tables

Table 1

Detailed Information on all Analysed Studies

		Measures			Samples				
#	Reference	Warmth	Competence	Scale	*N*	Description	M_{Age} (*SD*)	Sex	Targets[a]
1	Asbrock (2010)	likable, warm, good-natured	competent, competitive, independent	5-point scale (*not at all* to *completely*)	82	students of educational science	23.00 (2.77)	40.2% male, 4 unknown	Asians, Athletes, Blue-collar workers, Career women, Eco-freaks, Feminists, Foreigners, Gay men, Germans, Housewives, Jews, Lesbian women, Married people, Men, Muslims, Musicians, People with mental disabilities, People with physical disabilities, Physicians, Rich people, Senior citizens, Single people, Students, The homeless, The unemployed, Turks, White-collar workers, Welfare recipients, Women
2	Cuddy et al. (2009)	friendly, warm, good-natured, sincere	competent, confident, capable, skillful	5-point scale (*not at all* to *very much*)	98	students	23.44 (3.19)	52.0% male, 1 unknown	Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, The Netherlands, Portugal, Spain, Sweden, United Kingdom

		Measures			Samples				
#	Reference	Warmth	Competence	Scale	*N*	Description	M_{Age} (*SD*)	Sex	Targets[a]
3	Ehrke et al. (2020) – Study 2	altruistic, considerate, honest, supportive, understanding, selfish [reverse coded]	consistent, rational, assertive, energetic, determined, insecure [reverse coded]	7-point scale (*not at all* to *entirely)*	248	heterogeneous adult sample	46.81 (14.22)	30.6% male, 3 unknown	Diverse political party, Heterogeneous political party
4	Fröhlich & Schulte (2019)	warm, likable, good-natured, friendly	competent, independent, competitive, capable	5-point scale (*not at all* to *completely*)	200	mainly students	35.33 (11.69)	24.0% male, 1 other, 1 unknown	Germans Migrants from: Afghanistan, African countries, Albania, Arab countries, Bulgaria, China, Egypt, Italy, Greece, Pakistan, Romania, Russia, Syria, Tunisia, Turkey
5	Hackbart et al. (2020)	warm, good-natured, well-intentioned, likeable	competent, efficient, capable, intelligent	7-point scale (*not at all* to *completely*)	247	mainly students	29.40 (10.00)	23.9% male, 5 unknown	Educational counsellors with the following characteristics: German origin-Female, German origin-Male, Turkish origin-Female, Turkish origin-Male

		Measures			Samples				
#	Reference	Warmth	Compe-tence	Scale	*N*	Description	M_{Age} (*SD*)	Sex	Targets[a]
6	Hansen et al. (2017) – Job interview context	likeable, warm, good-natured	competent, compe-titive, inde-pendent	7-point scale (*not at all* to *very much*)	110	under-graduate students	22.33 (3.24)	33.5% male	Applicants with the following characteristics: German face – German accent, German face – Turkish accent, Turkish face- German accent, Turkish face – Turkish accent
7	Hansen et al. (2017) – Roommate search context				105				
8	Hansen et al. (2018) – Study 1a	likeable, warm, good-natured	competent, compe-titive, inde-pendent	7-point scale (*not at all* to *very much*)	60	under-graduate students	23.32 (4.50)	31.7% male	T1: German face, Turkish face T2: German face – German accent, German face – Turkish accent, Turkish face- German accent, Turkish face – Turkish accent
9	Hansen et al. (2018) – Study 1b				54		22.69 (3.76)	37% male	

		Measures			Samples				
#	Reference	Warmth	Competence	Scale	*N*	Description	M_{Age} (*SD*)	Sex	Targets[a]
10	Hellmann et al. (2015)	likeable, warm, good-natured	competent, competitive, independent	5-point scale (*not at all* to *completely*)	72	students	20.68 (3.56)	15.3% male, 5 unknown	Germans, The French, Turks
11	Ihme & Möller (2015)	likeable, helpful, sincere, warm, kind	competent, industrious, intelligent, determined	5-point scale (*not at all* to *extremely*)	120	mainly students	29.00 (9.92)	26.7% male	Students from the subjects Computer Science, Law, Psychology, Teacher training
12	Imhoff et al. (2013) - Pretest	benevolence, trustworthiness, heartiness	capacity, efficiency, competitiveness	10-point scale	96	not specified	29.86 (10.51)	35.4% male, 5 unknown	Professions (Artist, Attorney, Broker, Cab driver, Elementary school teacher, Engineer, Entrepreneur, Estate agent, Geriatric aide, Homemaker, Letter carrier, Manager, Meter maid, Nursery teacher, Physician, Politician) each combined with gender (Female, Male) (in total 32 targets)

		Measures			Samples				
#	Reference	Warmth	Competence	Scale	*N*	Description	M_{Age} (*SD*)	Sex	Targets[a]
13	Imhoff et al. (2013) – Main study			7-point scale	92		36.22 (14.38)	42.4% male	Male manager, Male nursery teacher
14	Janda et al. (2019)	friendly, warm-hearted, trustworthy, outgoing, empathetic, honest, likeable, good-natured, sociable, endearing, popular amongst her peers	productive, capable, intelligent, ambitious, assertive, determined, competent, successful, single-minded, independent, efficient	5-point scale (*strongly disagree* to *strongly agree*)	216	students	23.00 (3.26)	50.0% male	A Woman with PMDD in the following experimental conditions: Control, Psychoeducation, Stereotypes

		Measures			Samples				
#	Reference	Warmth	Competence	Scale	*N*	Description	M_{Age} (*SD*)	Sex	Targets[a]
15	Kemme et al. (2020)	likable, warm, good-natured	competent, competitive, independent	scale from 0% to 100%	364	police officer students	26.48 (5.45)	54.3% male	German men, Muslim men
16	Kotzur et al. (2017)	warm, friendly, well intentioned	competent, capable, independent	5-point scale (*not at all* to *very*)	371	university students without migration background	26.90 (9.24)	32.1% male, 7 other	Asylum seekers, Economic refugees, Refugees, War refugees
17	Kotzur et al. (2019)	likable, warm, good-natured	competent, competitive, independent	5-point scale (*not at all* to *very much*)	264	mainly students	24.21 (4.65)	26.6% male, 1.1% other	Afghan refugees, Christian refugees, Economic refugees, Elderly people, Germans, Homeless people, Iraqi refugees, Muslim refugees, Rich people, Refugees, Refugees from Eritrea, Refugees from North Africa, Refugees from the Balkans, Syrian refugees, Turkish migrants, War refugees

		Measures			Samples				
#	Reference	Warmth	Competence	Scale	*N*	Description	M_{Age} (*SD*)	Sex	Targets[a]
18	Kotzur et al. (2020) – Study 1	warm, friendly, well intentioned	competent, competitive, independent	5-point scale (*not at all* to *completely*)	301	mainly students	23.44 (6.29)	27.2% male	Athletes, Elderly, Homeless people, Muslims, Rich people, Students Each target was rated either from an individual or a societal perspective
19	Kotzur et al. (2020) – Study 2	warm, friendly, well intentioned, sincere, tolerant	competent, capable, independent, confident, intelligent	5-point scale (*not at all* to *completely*)	126	mainly students	27.51 (11.87)	27.8% male, 0.8% other	Elderly, Feminists, Germans, Homeless people, Housewives, Jobless, Physicians, Rich people, Turks Each target was rated either from an individual or a societal perspective
20	Kotzur et al. (2020) – Study 3	likable, trustworthy, warm, benevolent	competent, laborious, reliable, highly educated	7-point Semantic differential scale	1221	clickworkers	40.74 (10.68)	45% male	Germans, Immigrants from Albania, Bulgaria, China, Egypt, Italy, Poland, Romania, Russia, Turkey Each target was rated either from an individual or a societal perspective

		Measures			Samples				
#	Reference	Warmth	Competence	Scale	N	Description	M_{Age} (SD)	Sex	Targets[a]
21	Lepthien et al. (2017)	warm, friendly	competent, capable	7-point scale	281	heterogeneous adult sample	45.49 (12.53)	51.6% female	No demarketing condition, Demarketing condition
22	Nett et al. (2020)	likeable, warm, good-natured	competent, competitive, independent	7-point scale (*not* to *very*)	973	not specified	34.24 (10.69)	27% male, 3 other	Female names: Alexa, Alexandra, Alica, Alina, Angelika, Anita, Annette, Antonia, Beate, Bella, Bettina, Brigitte, Britta, Caroline, Celine, Chiara, Christine, Cindy, Claudia, Cornelia, Daniela, Doris, Elfriede, Elisabeth, Elke, Erika, Erna, Eva, Franziska, Gabi, Gertrud, Gina, Gisela, Hannelore, Heidi, Heike, Inge, Ines, Ingrid, Irmtraud, Jacqueline, Janine, Jessica, Jessie, Johanna, Judith, Jutta, Karin, Katharina, Kathleen, Kerstin, Kimberley, Larissa, Laura, Lea, Lena, Leonie, Lilly, Lisa, Lola, Mandy, Manuela, Maria, Martina, Melanie, Melissa, Merle, Mia, Miriam, Monika, Nicole, Nina, Petra, Regina, Renate, Sabine, Sandra, Sarah, Sarina, Selina, Silke, Silvia, Sonja, Sophia, Stefanie, Steffi, Stella,

		Measures			Samples				
#	Reference	Warmth	Competence	Scale	*N*	Description	M_{Age} (*SD*)	Sex	Targets[a]
									Susanne, Svea, Tabea, Tanja, Tess, Ulrike, Ursula, Uschi, Ute, Veronika, Yvonne, Zoe Male names: Alex, Alexander, Alfred, Andreas, Bastian, Bela, Bernd, Ceyhan, Christian, Christoph, Daniel, Dave, David, Dennis, Dieter, Dirk, Dylan, Fabian, Felix, Finn, Florian, Flynn, Frank, Franz, Friedrich, Gerd, Günter, Hans, Harald, Heiko, Heinz, Helmut, Herbert, Hermann, Holger, Horst, Jan, Jason, Jens, Joel, Johannes, Jörg, Joris, Julian, Jürgen, Justin, Karl, Karlheinz, Kevin, Klaus, Lars, Leon, Levi, Liam, Lionel, Luca, Lukas, Luke, Malik, Manfred, Manuel, Mario, Mark, Markus, Martin, Mats, Matt, Matthias, Max, Maximilian, Michael, Mike, Milo, Moritz, Niclas, Nils, Olaf, Oliver, Otto, Paul, Peter, Phil, Philipp,

		Measures			Samples				
#	Reference	Warmth	Competence	Scale	*N*	Description	M_{Age} (*SD*)	Sex	Targets[a]
									Rainer, Ralf, Rick, Ritchie, Robert, Sebastian, Simon, Stefan, Sven, Thomas, Thorsten, Tim, Tobi, Tobias, Tom, Tyson, Ulrich, Uwe, Volker, Walter, Wolfgang, Yannick
23	Renner (2019)	trustworthy, likeable	strong leaders, able to solve problems	5-point scale (*not at all* to *completely*)	1718	representative German internet user sample aged 18-68	44.30 (14.60)	54.4% male	Female politicians showing the following behaviors: Negative dominance, Negative submissiveness, Neutral, Positive dominance
24	Rennung et al. (2016)	warm, good-natured, likeable	competent, capable, confident	6-point scale (*not at all* to *a lot*)	2473	Heterogeneous adult sample	48.60 (14.50)	42.9% male	Pictures of the following positions: High power in females, High power in males, Low power in females, Low power in males

		Measures			Samples				
#	Reference	Warmth	Competence	Scale	*N*	Description	M_{Age} (*SD*)	Sex	Targets[a]
25	Schwind et al. (2019) – Study 1	tolerant, warm, good-natured, sincere	competent, confident, independent, competitive, intelligent	5-point scale (*not at all* to *extremely*)	71	computer science students	23.81 (5.16)	78.9% male	Person (Career people, Environmentalists, Homeless people, Rich people, Physicians, Senior citizens, Singles, Welfare recipients) each combined with all devices (Blood glucose sensors, Blood pressure monitors, EEG headsets, LED glasses, Narrative clips, Quadcopters, Tablets, VR headsets) (in total 64 targets)
26	Schwind et al. (2019) – Study 2				77	not specified	25.81 (8.16)	50.6% male	Blood glucose sensors, Blood pressure monitors, E-reader, EEG headsets, Fitness trackers, Gesture trackers, Head-mounted action camera, Hearing aid, LED glasses, LED tie, Narrative clips, Quadcopters, Smart glasses, Smartphone, Tablets, VR headsets

		Measures			Samples				
#	Reference	Warmth	Competence	Scale	*N*	Description	M_{Age} (*SD*)	Sex	Targets[a]
27	Stanciu et al. (2019)	warm, amusing, good-natured, well-intended, honest	conscientious, organized, diligent, competent, efficient, independent	5-point scale (*strongly disagree* to *strongly agree*)	209	mainly students	24.22 (5.17)	35.0% male	Homosexuals, Politicians, Rich people, Unemployed people, Women

		Measures			Samples				
#	Reference	Warmth	Competence	Scale	*N*	Description	M_{Age} (*SD*)	Sex	Targets[a]
28	Winter et al. (2020)	dishonest - sincere, egoistic - altruistic, threatening – benevolent, repellent - likable, untrustworthy – trustworthy, cold - warm	unassertive – competitive, unconfident - -confident, powerless - powerful, low status - high status, poor - wealthy, dominated - dominating	Sliding bar from 0 to 100	301	students	23.22 (3.43)	28.2% male	Doctors showing the following behaviors: Affirmations, Control condition, Negations
29	Wyszynski et al. (2020)	tolerant, warm, good-natured, sincere	competent, confident, independent, competitive, intelligent	5-point scale (*not at all* to *extremely*)	304	German-nationality convenience sample	36.34 (16.54)	42.1% male	Economic migrants, Migrants, Refugees

Note. N = Number of participants, *M* = mean value, *SD* = standard deviation. [a] Here, we only considered and reported targets that were rated by *n* ≥ 50 participants.

Table 2

CFA and ME Results Overview

		CFA		Configural ME	Metric ME	Scalar ME		
Sample	Total # Target Groups	# Acceptable Target Groups	% of Total Target Groups	# Target Groups	# Target Groups (ME Level)	# Target Groups (ME Level)	% of Acceptable Target Groups	% of Total Target Groups
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	/	/	/	/	/

		CFA		Configural ME	Metric ME	Scalar ME		
Sample	Total # Target Groups	# Acceptable Target Groups	% of Total Target Groups	# Target Groups	# Target Groups (ME Level)	# Target Groups (ME Level)	% of Acceptable Target Groups	% of Total Target Groups
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**

Note. ME = Measurement equivalence. # = 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. The references of the sample numbers can be obtained in Table 1. Note that for Samples 18-20, measurement equivalence was tested for each target group separately across experimental conditions if prerequisites were met (see Table 1 for details).

Figures

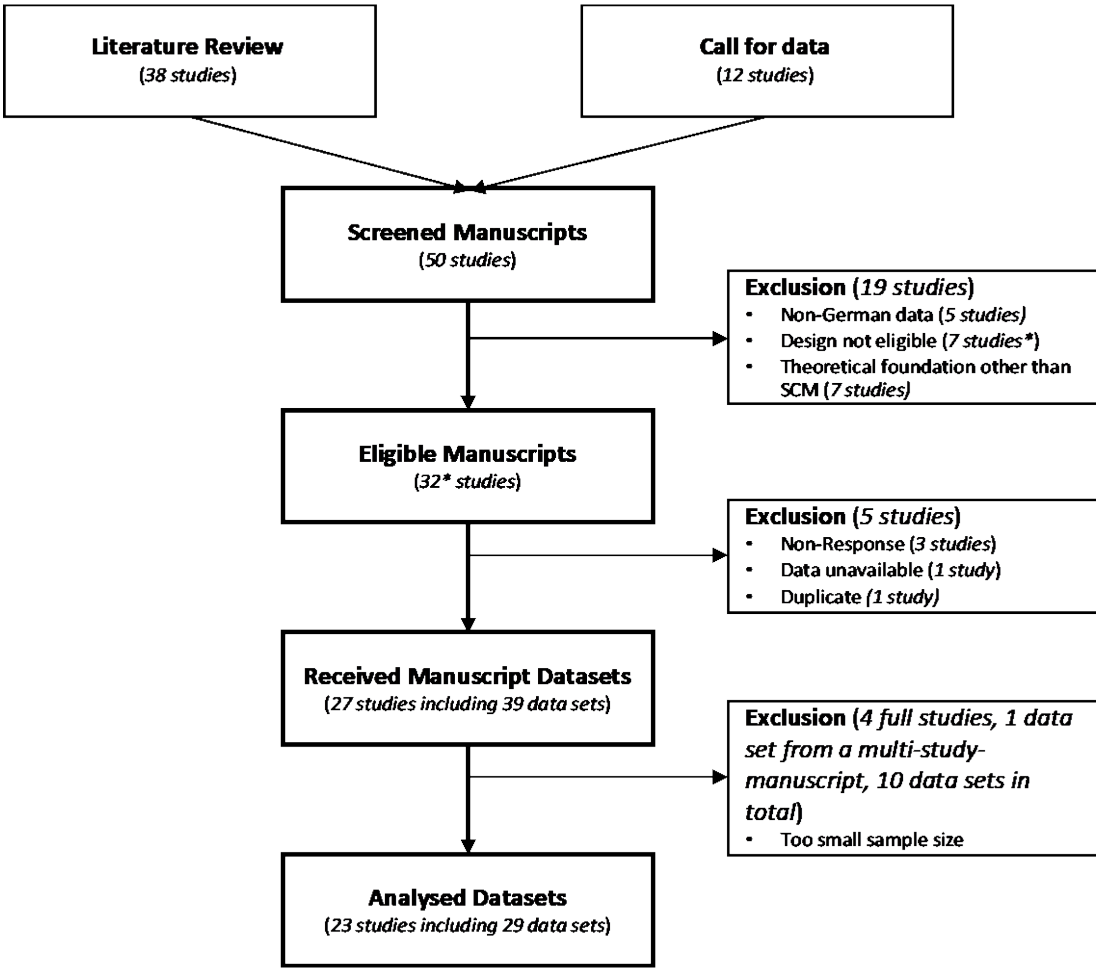

Literature Review
(38 studies)
Call for data
(12 studies)
Screened Manuscripts
(50 studies)
Exclusion (19 studies)
• Non-German data (5 studies)
• Design not eligible (7 studies*)
• Theoretical foundation other than SCM (7 studies)
Eligible Manuscripts
(32* studies)
Exclusion (5 studies)
• Non-Response (3 studies)
• Data unavailable (1 study)
• Duplicate (1 study)
Received Manuscript Datasets
(27 studies including 39 data sets)
Exclusion (4 full studies, 1 data set from a multi-study-manuscript, 10 data sets in total)
• Too small sample size
Analysed Datasets
(23 studies including 29 data sets)

Figure 1. Flow chart of the data selection process. *One manuscript included three different data sets, which were excluded for different reasons (design not eligible and too small sample size). Therefore, this study was counted double in this flow chart.

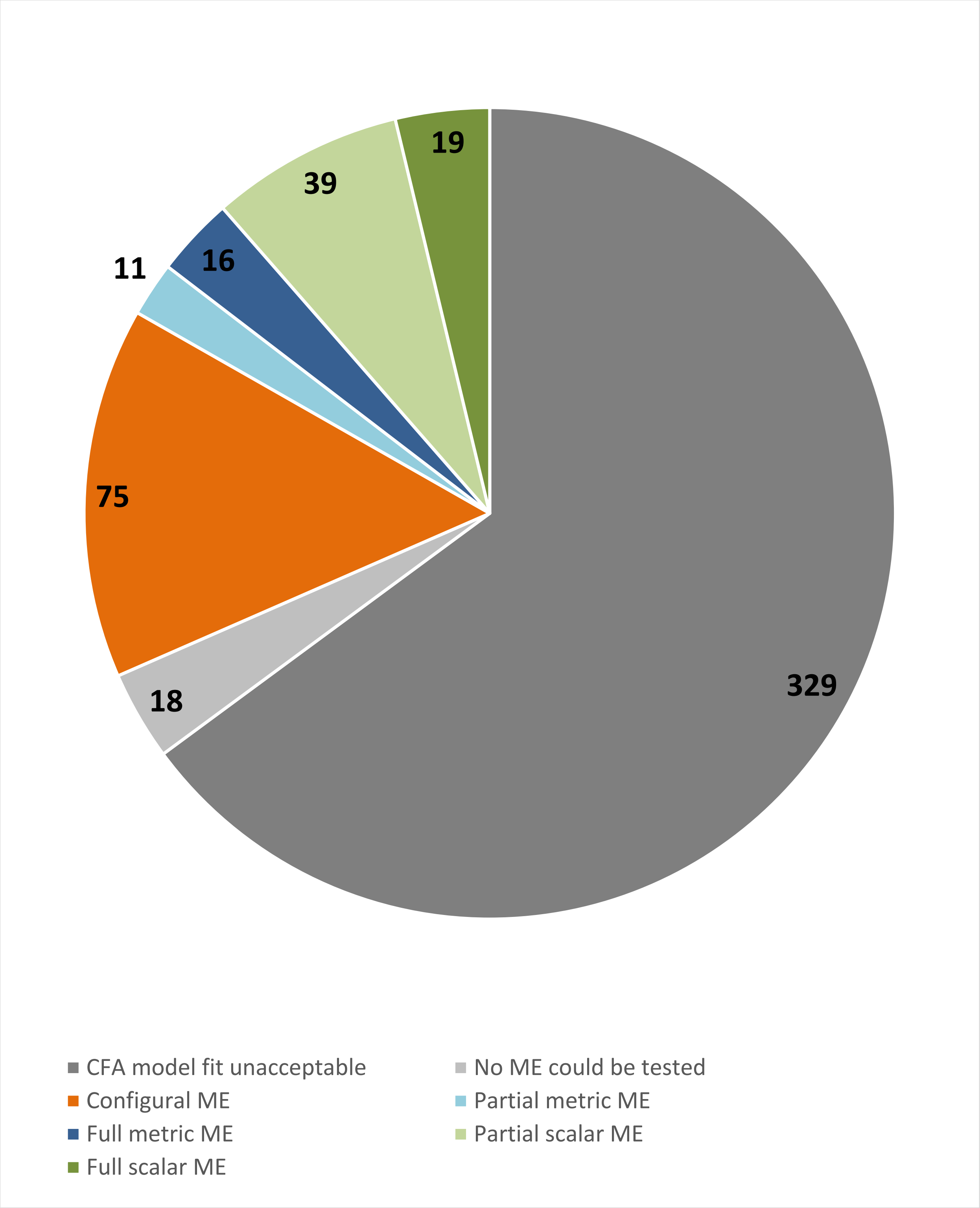


Figure 2. Highest level of established measurement equivalence per target group. Please note that the figure shows the highest level of measurement equivalence in which the target group dropped out from analyses, thus the numbers vary compared to the descriptions in the text. ME = Measurement equivalence. Total number of measurement model *K* = 507.

Supplementary Materials for Manuscript # 2

The supplementary materials for Manuscript # 3 are stored in the Open Science Framework, see https://osf.io/jqzet/.

They include:

- The deviations of the manuscript from the preregistration (OSM – A);
- Detailed data inclusion criteria and an overview of the data inclusion process (OSM – B);
- A detailed description of the analytic procedure (OSM – C);
- Detailed model-fit information for the Confirmatory Factor Analysis for all CFA models (OSM – D1);
- Detailed model-fit information for the measurement invariance assessment for all samples (OSM – D2).

Manuscript # 3

Stereotype Content of Refugee Subgroups in Germany

Patrick F. Kotzur

University of Osnabrück, Philipps-University of Marburg

Maria-Therese Friehs

University of Osnabrück

Frank Asbrock

Chemnitz University of Technology

Maarten H. W. van Zalk

University of Osnabrück

First authorship: Patrick F. Kotzur and Maria-Therese Friehs share the first authorship.

Corresponding author: Patrick F. Kotzur, Department of Psychology, University of Osnabrück, Seminarstraße 20, 49074 Osnabrück, Germany (Phone: +49163 7201303; E-mail: pakotzur@uos.de)

Conflict of Interest: The authors declare no conflict of interest. Acknowledgments: We thank Peter Schmidt for his advice on conducting our analyses, as well as Lea Müller and Leonie Schmidt for their support preparing this manuscript.

Stereotype Content of Refugee Subgroups in Germany

Word Count:

Abstract: 136

Main text, Abstract and Footnotes: 9370

Graphics: 3 Tables, 1 Figure

Online Supplementary Material (OSF): 10 Tables

Abstract

Stereotypes of refugee subgroups are still understudied. We contribute to this body of research by investigating differences in stereotype content, meaning warmth and competence ratings, of refugee subgroups in Germany (*N* = 264). Most extant Stereotype Content Model research is based on observed warmth and competence means values. We applied latent variable modelling using the alignment optimisation to ensure meaningful and reliable mean value comparisons. Generic refugees were rated as lacking warmth and competence. Warmth assessments of refugee subgroups varied depending on flight motives, geographical origin, and religious affiliation, implying that perceptions of threat and competition differed between these subgroups. Less differences emerged in competence assessments, indicating that refugee groups are generally regarded as lacking status. Our results enhance knowledge of the stereotype content of refugee subgroups and make a methodological contribution to stereotype content research.

Keywords: stereotype content, refugees, subgroups, Germany, alignment optimization

Stereotype Content of Refugee Subgroups in Germany

March 2016: A German news platform describes the situation of migrants in Athens, explaining that for some time, only *War refugees*[1] were allowed to continue on the so-called "Balkan route", while *Economic refugees*[2] were forced to remain behind (n-tv, 2016). June 2016: When analysing the 2015 crime statistics of the federal criminal agency, the widely disseminated German newspaper DIE WELT explains why refugees of certain origins, especially from North Africa or the Balkan states, tend to become more delinquent than immigrants of other origins (Hackensberger, Kalnoky, & Smirnova, 2016). September 2016: The Christian-conservative Bavarian political party CSU demands that Christian migrants should be preferred over Muslim ones in the German migration system ("CSU will christliche Zuwanderer bevorzugen", 2016). These instances exemplify various public discourses in Germany since the beginning of the so-called "refugee crisis" in 2015, when about 890,000 people sought refuge in Germany (Bundesministerium des Innern, für Bau und Heimat [BMI], 2016). Notably, these narratives suggest that refugees have not been represented as one homogenous social group in Germany, but that there are different subtypes of refugees to be distinguished along several dimensions, including the perceived flight motives, religious affiliation, and geographic origin ("CSU will christliche Zuwanderer bevorzugen", 2016; Hackensberger et al., 2016; n-tv, 2016). Recently, empirical studies have begun to investigate whether this differentiation results in different social perceptions of flight- and migration-related subgroups (Bansak, Hainmueller, & Hangartner, 2016; Ditlmann, Koopmans, Michalowski, Rink, & Veit, 2016; Kotzur, Forsbach, & Wagner, 2017).

Further addressing this issue is of high scientific and social relevance: European countries are, and most likely will continue to be, important destinations for people seeking refuge (Eurostat, 2016, 2018), making refugees a relevant social group in these countries. This is especially true for Germany, which received more than a third of all first-time applications within the European Union in 2015 (Eurostat, 2016). However, insights regarding factors shaping receiving society members' perception of the newcomers, such as perceived characteristics of refugees, are scarce. The social perception of refugees has profound consequences, as it is likely to govern refugee-receiving community relations, as well as the broader context of reception of immigrant groups in general (Kotzur, Tropp, & Wagner, 2018). Additionally, the social perception of groups influences whether they are supported or harmed (Cuddy, Fiske, & Glick, 2007). Identifying subgroups running elevated risks of becoming targets of hostility and aggression allows for target group-specific social interventions to improve receiving society-refugee relations. Therefore, understanding factors that impact the social perception of refugees is an important research goal.

While pursuing these goals, we acknowledge that vital statistical preconditions for conducting substantive group comparisons in the social perception literature have only scarcely been tested (i.e., measurement invariance; but see Janssens, Verkuyten, & Khan, 2015; Stanciu, Cohrs, Hanke, & Gavreliuk, 2017). We do so by analysing social perception on a latent level, which allows controlling for reliability differences and assuring valid comparison of constructs across social groups (Kline, 2010). Specifically, we apply the alignment optimisation method (Asparouhov & Muthén, 2014) to establish approximate measurement invariance in our research, a recent, researcher-friendly statistical technique that both tests for and achieves the necessary preconditions for latent mean comparisons.

Stereotype Content of Refugee Subgroups in Germany

An influential theoretical framework to study social perception of groups, such as subgroups of refugees, is the Stereotype Content Model (SCM; Fiske, Cuddy, Glick, & Xu, 2002). The SCM proposes that culturally shared stereotypes towards social groups in a given society are based on two fundamental dimensions of social perception: Warmth, the “potential harm or benefit of the target group’s goals” (Cuddy et al., 2007, p. 632) in relation to one’s ingroup’s goals; and competence, the “degree to which the group can effectively enact those goals” (Cuddy et al., 2007, p. 632). Stereotype content research focusses on culturally shared stereotypes in a given society by asking participants to indicate what they assume most society members think about a specific social group (Fiske et al., 2002). While this procedure is assumed to reduce social desirability, it implies that stereotype content predominantly taps into the perceived majority's perspective on stereotypes, which can be different from individually endorsed stereotypes (Ashmore & Del Boca, 1981). Perceptions of threat and competition serve as predictors of the social group's perceived intentions (i.e., warmth perceptions), whereas perceived status engenders competence perceptions (Binggeli, Krings, & Sczesny, 2014a; Fiske et al., 2002; Kervyn, Fiske, & Yzerbyt, 2015). The SCM has been frequently applied to describe cultural stereotypes of social groups in different national contexts (e.g., Asbrock, 2010; Binggeli et al., 2014a; Burkley, Durante, Fiske, Burkley, & Andrade, 2017; Bye, Herrebrøden, Hietland, Røyset, & Westby, 2014; Clausell & Fiske, 2005; Cuddy et al., 2009; Durante et al., 2013, 2017; Eckes, 2002; Janssens et al., 2015; Sadler, Meagor, & Kaye, 2012; Stanciu et al., 2017). The combination of both dimensions is theorised to predict contemptuous (low warmth/low competence; e.g., homeless people), envious (low warmth/high competence; e.g., rich people), and paternalistic (high warmth/low competence; e.g., elderly people) outgroup perceptions, as well as positive perceptions of allied and ingroups (high warmth/high competence; e.g., one's own national group; Cuddy et al., 2009; Fiske, 2018). These, in turn, result in facilitative (for high warmth and/or high competence groups) and harmful (for low warmth and/or low competence groups) action tendencies towards these groups (Cuddy et al., 2007).

Despite these implications, there are no comprehensive studies investigating the stereotype content of (subgroups of) refugees. Research has shown in many country contexts – including Germany – that immigrant groups are generally rated low on warmth and competence (e.g., Asbrock, 2010; Binggeli, Krings, & Sczensny, 2014b; Eckes, 2002; Lee & Fiske, 2006; "The Fiske lab", n.d.). Intriguingly, studies suggested that stereotype content of generic social groups, meaning social groups without any further describing characteristics, do not need to correspond to specific subgroups, for instance, when additional subgroup information along key dimensions are provided (Binggeli et al., 2014b; Burkley et al., 2017; Bye et al., 2014; Clausell & Fiske, 2005; Eckes, 2002; Lee & Fiske, 2006; Sadler et al., 2012). For instance, in one study, *Women* as a general group were rated as warm and incompetent (Asbrock, 2010). Whereas the stereotype content of the specific subgroup *Housewives* matched this profile, *Career women* were rated as cold and competent (Asbrock, 2010). Recent studies, which we review in the following, provided initial evidence that this mechanism may also apply to present-day subgroups of refugees (Bansak et al., 2016; Binggeli et al., 2014b; Ditlmann et al., 2016; Kotzur et al., 2017; Lee & Fiske, 2006).

Although they did not explicitly examine subgroups of *refugees*, two studies investigated the stereotype content of immigrant subgroups: One in the U.S. (Lee & Fiske, 2006) and one in Switzerland (Binggeli et al., 2014b). Since refugees are an immigrant group, the same stereotype content-organizing principles may apply to refugee subgroups. In both studies, the researchers found that the *region or country of origin* served as an important cue of subgroups' perceived competition and status in the respective society, predicting subgroups' stereotype content. In the U.S., African immigrants were rated as warmer and less competent than immigrants from the Middle East (Lee & Fiske, 2006). In Switzerland, African immigrants were rated as warmer, but less competent, than immigrants from the Balkans[3] (Binggeli et al., 2014b). Recent applicants for asylum in Germany originated mainly from these just-mentioned regions (Middle East: Syria, Afghanistan, Iraq; Balkan: Albania, Kosovo, Serbia; Africa: Eritrea; Juran & Broer, 2017). Indeed, the country of origin mattered as to whether participants were willing to grant asylum to refugees, or send them back to their country of origin (Bansak et al., 2016), arguably a benevolent (helping) or hostile (harming) behavioural intention towards distinct subgroups of origin, that may be reflected in warmth and competence ratings of these groups.

Along with Bansak et al. (2016), recent studies have begun to explicitly focus on the social perception of refugee subgroups in Germany and other European countries (Ditlmann et al., 2016; Kotzur et al., 2017). The only published SCM study investigating subgroups of refugees examined the impact of the *flight motive* on warmth and competence (Kotzur et al., 2017). The authors found that refugees fleeing due to economic reasons (*Economic refugees*) were rated significantly less warm than those that fled due to war (*War refugees*). No significant differences emerged on the competence dimension. The migration motive was an important predictor of the willingness to grant asylum in two further studies (Bansak et al., 2016; Ditlmann et al., 2016). Participants were less willing to welcome refugees fleeing from economic hardship than from war (Bansak et al., 2016; Ditlmann et al., 2016), lending further evidence to the importance of flight motive as a determinant of refugee subgroups' social perception. Consequently, the motive of migration may be an important predictor for the warmth ratings of refugee subgroups.

We have identified further attributes affecting the willingness to grant asylum that arguably may serve as cues for warmth and/or competence ratings (Bansak et al., 2016): For competence, potential cues include age, previous occupation, and language skills (Bansak et al., 2016). For warmth, and its precursors threat and competition, possible cues are religion and vulnerability (Bansak et al., 2016). *Religious affiliation* may be particularly relevant, since present-day refugees often originate from dominantly Muslim countries (Juran & Broer, 2017). Prior SCM studies found that Muslims were rated less warm and competent than Christians in societies where Christians are the majority ("The Fiske lab", n.d.), such as Germany. Consequently, Muslim refugees might be perceived as more threatening and of lesser status than Christian refugees, and thus as less warm and less competent.

Taken together, these findings indicate that generic refugees are rated relatively cold and incompetent. Prior SCM research focusing on other social groups suggested, however, that the stereotype content of subgroups may diverge from this generic view. Researchers have begun to examine characteristics that may lead to shifts in the social perception of subgroups of refugees, mostly by examining the willingness to accept refugees with certain characteristics to one's country (but see Kotzur et al., 2017). We aim to contribute to this body of research using the SCM, a comprehensive social psychological framework that makes a range of predictions regarding antecedents and consequences of such perceptions.

Lastly, we identified two further gaps in this literature that need addressing. Firstly, the refugee subgroups to be studied had been exclusively studied with a top-down approach in prior research on refugee subgroup perception, that is, subgroup characteristics were selected by the researchers (Ditlmann et al., 2016; Bansak et al., 2016; Kotzur et al., 2017). Although there may be a wide range of characteristics of a social group that may be used to derive warmth and competence assessments, the question is still underexplored which of them are typically used by participants to meaningfully distinguish between refugee subgroups. Consequently, using a bottom-up approach might grant valuable insight into the characteristics that participants assume relevant to organise refugee subgroups. Secondly, prior studies have examined the social perception of refugees independently from other social groups. However, including reference groups, whose stereotype content within the SCM model has been reliably depicted within a specific society, is important for a meaningful contextualization of the relative location of refugee subgroups within the two-dimensional warmth by competence space (see, e.g., Lee & Fiske, 2006). Examining how the novel subgroups are rated relative to established social groups scoring particularly high or low on warmth or competence shows insight into the new groups' standing in society (i.e., in relation to important societal benchmarks).

Methodological Advances of SCM Research

In the spirit of the "crisis of confidence" (Kruglanski, Chernikova, & Jasko, 2017, p. 1) that has led researchers to question the appropriateness of the methods used and robustness of social psychological findings, we raise methodological concerns related to the extant SCM research. One major criticism is that most SCM research is based on the analysis of observed values (i.e., computed scale means), confounding the true scale score with measurement error (Kline, 2010). Latent variable modelling is thus more appropriate in most cases, since it accounts for measurement error in the model (for a general readable introduction into the topic, see Cai, 2012). Another criticism is that most extant SCM research refrained from establishing measurement invariance (MI), that is, to test whether an "instrument measures the same concept in the same way across various subgroups of respondents" (Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014, p. 58; but see Janssens et al., 2015; Stanciu, 2015; for a general introduction into the topic, see Vandenberg & Lance, 2000).

SCM research usually aims at comparing (aggregated) mean values of social groups' assessments of warmth and competence in its analyses. This is often done using cluster analysis (e.g., Fiske et al., 2002), or observed mean value comparison (e.g., Kotzur et al., 2017), interpreting results for example as groups x_1,…x_n are rated as warmer/less warm and/or more/less competent than groups y_1, … y_n. Thus, measurement invariance is a key prerequisite to meaningful and valid mean value comparisons. The measures of warmth and competence should show at least (partial) scalar measurement invariance across target groups to avoid the proverbial comparison of "apples and oranges" (Chen, 2008) and to ensure that no systematic bias leads to over- or underestimation of any dimension between social groups (Vandenberg & Lance, 2000). Scalar MI is obtained when factor loadings and intercepts are constrained to be equal across groups, most commonly tested using multiple group confirmatory factor analysis (MGCFA; van de Schoot, Schmidt, De Beuckelaer, Lek, & Zodervan-Zwijnenburg, 2015). Partial measurement invariance refers to releasing the equality constraints, and thus the strict assumptions of the exact measurement invariance model, of highly deviating parameters while establishing equality for at least two other parameters (Byrne, Shavelson, & Muthén, 1989). We are aware of very few studies that have explicitly tested for measurement invariance of SCM constructs, thus ensuring meaningful and valid mean value comparisons of observed and of latent variables between social groups (Janssens et al., 2015; Stanciu, 2015), all using the MGCFA approach.

A recent alternative is the multiple-group factor analysis alignment, or *alignment optimisation* method (Asparouhov & Muthén, 2014). This procedure allows to conduct latent mean value comparisons across groups while establishing a mathematically optimised (partial) scalar measurement invariance pattern (Asparouhov & Muthén, 2014). This approach is recommended for comparisons of large numbers of groups, and combines two advantages compared to other MI approaches: In the alignment optimisation procedure, (a) MI is tested less strictly and arguably enables a researcher to more realistically examine a broader range of group comparisons, without rejecting a group comparison early in the procedure because of MI violation (thus accounting for recent criticism of the MGCFA approach; see Van de Schoot, Kluytmans, Tummers, Lugtig, Hox, & Muthén, 2013); and (b) (partial) MI is established in an automated, easy-to-interpret manner using an algorithm that discovers “a solution where there are a few large noninvariant measurement parameters and many approximately invariant measurement parameters” (Asparouhov & Muthén, 2014, p. 3). This avoids cumbersome manual model improvements using modification indices that are required when establishing partial measurement invariance in the MGCFA framework, which might result in (a) potential errors made by the researcher at each manual modification index assessment, and (b) difficulties in replication of findings (Asparouhov & Muthén, 2014). Applied to the research at hand, alignment optimisation allows us to strengthen the internal and construct validity as well as the interpretability of our findings by generating (partial) measurement invariant latent mean values of warmth and competence assessments across a variety of target groups, while relaxing the oftentimes unrealistic exact measurement invariance assumption (Muthén & Asparouhov, 2013).

The Present Study’s Contributions and Predictions

With the present study, we intend to contribute to the literature by investigating stereotype content of subgroups of refugees. While researchers of extant studies investigating the social perception of refugee subgroups have selected the subgroup attributes themselves (Bansak et al., 2016; Ditlmann et al., 2016; Kotzur et al., 2017), we asked participants to nominate meaningful subgroup categories in a pilot study. This allowed us to investigate the stereotype content of a range of subgroup dimensions that are most likely meaningful to our participants (see, e.g., Binggeli et al., 2014a; Lee & Fiske, 2006). Moreover, we included reference groups that have been identified previously to score particularly high or low on either or both stereotype content dimensions in the present country context (Asbrock, 2010; Eckes, 2002), enabling us to map groups within the SCM space more comprehensively than prior studies (such as Kotzur et al., 2017).

Prior research indicates that generic refugees are rated relatively cold and incompetent (“The Fiske lab”, n.d.). We expected that subgroups of refugees diverge from this generic view:

Specifically, we expected differences in warmth and competence ratings between subgroups, whereby country or region of origin, flight motives, and religious affiliations may serve as meaningful subgroup organisers (Bansak et al., 2016; Binggeli et al., 2014b; Ditlmann et al., 2016; Kotzur et al., 2017; Lee & Fiske, 2006). Based on previously reported findings, we expected that, as for flight motive, subgroups fleeing from war should be rated warmer than subgroups that fled for economic reasons (Bansak et al., 2016; Ditlmann et al., 2016; Kotzur et al., 2017). As for country or region of origin, refugees from African countries and regions are expected to be rated warmer and less competent than refugees from Balkan or Middle Eastern countries (Binggeli et al., 2014b; Lee & Fiske, 2006). As for religious affiliation, we expected Muslim refugees to be rated less warm and less competent than Christian refugees, based on findings in other SCM studies on religious groups (e.g., "The Fiske lab", n.d.).

With regard to the societal reference groups, we selected groups that in prior research reliably scored in the four quadrants of the two-dimensional warmth-competence space (Asbrock, 2010; Fiske, 2018). In accordance with our expectation that refugees, and thus refugee subgroups, are rated relatively cold and incompetent, we expected that *Germans*, a high warmth-high competence reference group, receive higher values than refugee subgroups on both dimensions. *Elderly people*, a high warmth-low competence reference group, should receive higher warmth assessments and *Rich people*, a low warmth-high competence group, higher competence ratings compared to refugee subgroups. Lastly, we included *Homeless people* as a low warmth-low competence group. Based on Asbrock (2010), who found "foreigners" to be rated similar in warmth and higher in competence compared to "the homeless", we assume that *Homeless people* are rated less competent than refugee subgroups[4].

The last intended contribution was to move beyond observed mean analysis and cumbersome MGCFA and apply state-of-the-art methods appropriate to compare latent means of warmth and competence across large numbers of target groups. In order to allow for meaningful and internally as well as construct valid comparisons of the SCM dimensions, we employed the innovative alignment optimisation procedure that automatically establishes the best fitting (partial) scalar invariant solution and estimates as well as compares latent mean values of different target groups accordingly (Asparouhov & Muthén, 2014).

Methods

All data were collected at two German universities in the context of a larger research endeavour on the social perception of social groups. In Germany, the conduction of studies based on anonymous and confidential questionnaires that are not expected to entail any lasting harms or risks for the participants requires no additional permission by an internal review board. Thus, formally obtaining an internal review board approval was not necessary. All procedures were performed in full accordance with the ethical guidelines of the Deutsche Gesellschaft für Psychologie (German Society for Psychology), and adhered to the low-risk study requirements of the universities where the studies have been carried out. Supplementary material, including the surveys, pilot study results, syntaxes, outputs, and results of additional analyses, has been made accessible on the Open Science Framework's website (see https://osf.io/5j7t6/). Raw data is available upon request from the corresponding author; it cannot be made publicly available since participants were informed that data management would be controlled by the study authors at all times.

Pilot Study and Subgroup Generation

We conducted an online pilot study to explore meaningful subcategory dimensions of refugees following the procedures described in Asbrock (2010). Participants (*N* = 80) were simultaneously recruited online through students' mailing lists from two mid-sized German universities (n_1 = 40 and n_2 = 40, respectively). Participants were not incentivised for their participation. All participants gave their informed consent prior to their inclusion in the study. To offer participants some guidance for this task (see Asbrock, 2010), they read the following instruction: "In the text box below, please list all migration-related groups in Germany that come to your mind. Please specifically consider groups with flight experiences and their backgrounds (e.g., relating to geographical, religious, flight cause characteristics). There are no right or wrong answers."

The subsequent text box was unlimited. To increase participants' focus on listing groups, the online survey's continue-button was suppressed for 90 seconds (a full list of all identified groups and further procedural details how the open answers were coded can be found in https://osf.io/rqk96/).

In total, 83 and 108 groups were identified by the two samples of universities 1 and 2, respectively. We deemed groups that were mentioned by more than 20% of one of the samples to be meaningful and commonly used subgroup labels ($n > 8$; Asbrock, 2010), and consequently included them into the main study. As expected, subgroups with reference to flight motive, region or country of origin, and religion emerged: *Syrians, Afghans, Iraqis, Eritreans, Turks, North Africans, People from the Balkans, War refugees, Economic refugees, Muslims, Christians*.

We amended all group descriptions with the label *refugees* (e.g., *Syrian refugees*), except for *Turks*. Historically, *Turks* have been economic migrants to Germany since the 1960s (Bade, 1992), so we labelled this group *Turkish migrants*. We kept the overall generic group *Refugees* in the list to receive reference information. Finally, we included four social groups that have shown to be located in the four extreme quadrants of the SCM as additional reference groups (*Germans* for high warmth/high competence, *Rich people* for low warmth/high competence, *Elderly people* for high warmth/low competence, and *Homeless people* for low warmth/low competence; see also Asbrock, 2010; Binggeli et al., 2014b; Fiske, 2018; Lee & Fiske, 2006).

Main Study

Participants and procedure. Using parallel online surveys, data for the main study were simultaneously collected between February and March 2017 using university-wide mailing lists from the same mid-sized German universities where we conducted our pilot studies. Bentler and Chou (1987) suggested 5 to 10 observations per estimated parameter for latent variable modelling. Asparouhov and Muthén (2014, p. 10) suggested "good recovery for all parameters except the factor variances is found already for $N_g = 100$" when applying alignment optimisation. We thus recruited $N = 264$ ($n_1 = 79$; $n_2 = 185$)[5] German adults (72.3% female, 1.1 % other; $M_{age} = 24.21$, $SD_{age} = 4.65$; 95.5% university students, 4.6% other; 84.8% without migration background, 0.8% missing[6]) to fulfil these prerequisites. The two subsamples from the two universities did not differ significantly on any demographic variables (all *p*s > .05). Thus, we collapsed both subsamples to one joint sample on which we based all subsequent analyses. All participants gave their informed consent prior to their inclusion in the study and were compensated with course credit and the opportunity to donate one Euro to a non-governmental aid organization of their choice.

The survey contained items concerning demographic information, stereotype content, participants' membership to one of the surveyed outgroups, and other variables not relevant to the study at hand (for a complete list of constructs we assessed, see Questionnaire: https://osf.io/e3mqg/). Following the procedure of previous SCM studies (e.g., Cuddy et al., 2009, study 1; Eckes, 2002), we presented stereotype content items one indicator per page, alternating warmth and competence indicators. For each indicator, we asked participants to evaluate the stereotype content of *Refugees* first, followed by a random order of refugee subgroups, *Turkish migrants*, and a random order of the non-migrant anchor groups. To prevent participant fatigue, we implemented a Three-Form-Design by inducing completely random, planned missingness on a subset of refugee groups (Graham, 2009; for an overview over the randomization, please see Online Supplementary Material Table SM1: https://osf.io/y76kq/).

SCM measure. We measured the stereotype content, and thus warmth and competence ratings, with the German-language SCM scales used in Asbrock (2010): "From the perspective of most Germans, how [ITEM] are the following social groups?". A social groups' warmth was assessed using the items "good-natured", "warm", and "likeable", while competence was assessed with "competent", "independent", and "competitive". Answers were given on a scale from 1 (*not at all*) to 5 (*very much*).

Analysis strategy. Following Asparouhov and Muthén (2014), we first assessed the general baseline measurement model fit for all 16 social groups under investigation using confirmatory factor analyses. We determined model fit to be adequate if all criteria formulated by Schermelleh-Engel, Moosbrugger, and Müller (2003) were met: $\chi^2/df < 3$; root mean standard error of approximation (RMSEA) < .08; standardised root mean square residual (SRMR) < .10; comparative fit index (CFI) > .95. Second, as a prerequisite for alignment optimisation, we established a configural model across all groups with acceptable model fit; for these groups, "the number of subscales (i.e., factors), the location of the items (i.e., pattern by which items load onto each factor), and postulated correlations among the subscales (i.e., existence of covariances)" (Byrne, 2008, p. 873) were specified to be equal across groups. In a final step, we used *alignment optimisation*. This analysis strategy allowed us to estimate trustworthy latent means and comparing them for significant group mean differences while at the same time generating an optimised approximate measurement invariance pattern (Cieciuch, Davidov, & Schmidt, 2018).

Results

We conducted all analyses in Mplus Version 8, using robust maximum likelihood estimator (MLR) to account for multivariate non-normality and non-independence of observations (Muthén & Muthén, 1998-2017). Descriptive information on the stereotype content of the surveyed social groups in terms of warmth and competence are outlined in the Online Supplementary Material (see Table SM2: https://osf.io/y76kq/). Correlation tables of all indicators within one social group (Table SM4) and of warmth and competence scales within and across social groups (Table SM5) are also provided in the Online Supplementary Material.

Baseline Model Fit

Following the procedure described above, we first ran 16 single-group confirmatory factor analyses (one per target group). For each group, we specified one warmth factor with the indicators "good-natured", "warm", and "likeable" and one competence factor with the indicators "competent", "independent", and "competitive". Warmth and competence factors were correlated, no cross-loadings or indicator residual covariations were allowed (for the syntax, see the folder Analysis material, 1 – Baseline models, e.g., https://osf.io/y3hwq/). Results are presented in Table 1. Ten out of 16 groups achieved an acceptable model fit: *Refugees, Syrian refugees, Muslim refugees, Afghan refugees, War refugees, Economic refugees, Refugees from Eritrea, Refugees from North*

Africa, Elderly people and *Homeless people.* The six remaining social groups *Christian refugees, Refugees from the Balkans, Iraqi refugees, Turkish migrants, Germans* and *Rich people* were discarded from further analysis due to non-acceptable model fit.

Please insert Table 1 about here

Alignment Optimisation Model

Configural measurement invariance. We entered the ten social groups showing adequate model fit into a simultaneous analysis for the configural measurement model. The model showed good fit, $\chi^2(80) = 114.890$, $p = .006$, $\chi^2/df = 1.436$, RMSEA = .046, SRMR = .034, CFI = .988, allowing us to focus on our research questions using the subsequent alignment optimisation procedure.

Measurement non-invariance*.* The fixed alignment optimisation solution we obtained with *Refugees* as a reference group showed two out of 120 parameters (two indicator intercepts; less than 2% of all parameters) to be non-invariant[7]. This finding indicated that a trustworthy estimation and comparison of latent warmth and competence means was possible, as the share of non-invariant parameters did not exceed 25% (Asparouhov & Muthén, 2014). The result was a latent mean value comparison based on a metric and partial scalar approximate measurement invariant model.

Latent mean values of stereotype content and significance testing. The ranking of the social groups, their latent mean values, and the significant differences to other social groups are outlined in Table 2 for warmth, and in Table 3 for competence. The findings are depicted in Figure 1. Further information on the alignment optimisation model (including factor loadings, indicator intercepts, factor means and variances, and the factor covariation of warmth and competence) are provided in Table SM3 in the Online Supplementary Materials (for more information, see https://osf.io/y76kq/).

Please insert Table 2 about here

Please insert Table 3 about here

Please insert Figure 1 about here

We examined differences in warmth and competence assessments between all refugee subgroups that had achieved acceptable baseline measurement model fit. As expected, both stereotype content dimensions showed statistically significant differences between refugee subgroups. In the following, we focus on selected group differences according to our predictions. For a complete list of significant differences that emerged on both warmth and competence between all groups, see Tables 2 and 3.

Differences regarding implied flight motive. Regarding the implied flight motive, we expected that *War refugees* should be rated warmer than *Economic refugees* (Kotzur et al., 2017). In accordance with this expectation, *War refugees* (latent factor mean α_W = 0.548, rank 2) scored significantly higher on warmth than *Economic refugees* (α_W = -1.105, rank 10). In fact, *War refugees* received the highest warmth ratings of all refugee subgroups. In contrast, *Economic refugees* showed the lowest warmth assessments, significantly lower than any other subgroup included.

We had not formulated any expectations regarding differences in terms of competence based on implied flight motive. In fact, prior research had found none (Kotzur et al., 2017). Unexpectedly, *Economic refugees* received the highest competence ratings of all refugee subgroups (α_C = 0.559, rank 2); significantly higher than *War refugees* (α_C = 0.015, rank 5). In sum, from all refugee subgroups, participants attributed the highest levels of warmth to *War refugees*. In accordance with our prediction, *War refugees* were rated substantially warmer than *Economic refugees*. These received the highest competence ratings of all refugee subgroups, differing significantly from *War refugees*.

Differences regarding origin. Due to non-acceptable baseline model fit, we could not include *Refugees from the Balkans*, *Iraqi Refugees* and *Turkish migrants* into the analysis at hand. For refugees from African countries, acceptable baseline model fit was found for *Refugees from Eritrea* and *Refugees from North Africa*. For refugees from Middle Eastern countries, acceptable baseline model fit was found for *Syrian refugees* and *Afghan refugees*. Therefore, we compared these groups. Based on previous research on the stereotype content of immigrant subgroups of different countries and regions of origins conducted in Switzerland and the U.S. (Binggeli et al., 2014b; Lee & Fiske, 2006), we expected that refugees from African countries should be rated warmer than refugees from Middle Eastern countries.

Our expectations for warmth were not confirmed: *Syrian refugees* (Middle Eastern country) received the highest warmth ratings (α_W = 0.000, rank 4) from all subgroups that indicated a region or country of origin; significantly higher than *Afghan refugees* (α_W = -0.375, rank 7, Middle Eastern country), *Refugees from Eritrea* (α_W = -0.335, rank 6; African country), and *Refugees from North Africa* (α_W = -0.854, rank 9; African region). *Refugees from Eritrea* (African country) and *Afghan refugees* (Middle Eastern region) were rated significantly warmer than *Refugees from North Africa* (African region). Thus, contrary to our assumption, refugees from Middle Eastern countries were rated warmer than or non-significantly different in warmth from refugees of African origin.

In terms of competence, we expected that refugees from Middle Eastern countries should be rated as more competent than refugees from African countries (Binggeli et al., 2014b; Lee & Fiske, 2006). Our expectations were confirmed: *Syrian refugees* (Middle Eastern country) received highest competence ratings (α_C = 0.216, rank 3), non-significantly different to *Afghan refugees* (α_C = -0.018, rank 7; Middle Eastern country), but significantly higher than *Refugees from Eritrea* (α_C = -0.214, rank 8; African country) and *Refugees from North Africa* (α_C = -0.348, rank 9; African region). *Afghan refugees* (Middle Eastern country) were rated significantly more competent than *Refugees from North Africa* (African region), but non-significantly different from *Refugees from Eritrea* (African region). In sum, regarding origin, our predictions for warmth were contradicted, for competence partially confirmed. Refugees from Middle Eastern countries were rated warmer or not significantly different in warmth compared to refugees from African countries, and partially more competent than refugee subgroups of African origin.

Differences regarding religious affiliation. We anticipated that *Muslim refugees* should be rated less benevolently than *Christian refugees*, i.e., receive lower warmth and competence ratings. Unfortunately, we were unable to test this expectation, since we had to exclude *Christian refugees* from our analyses due to non-acceptable baseline model fit. Nonetheless, the results did reveal that *Muslim refugees* were among the groups that received comparatively low warmth ratings (α_W = -0.639, rank 8), providing indirect evidence for a depreciation of *Muslim refugees* relative to other refugee subgroups. *Muslim refugees* were however rated comparatively high in competence (α_C = *0.043, rank 4)*. Thus, we found evidence that *Muslim refugees* were generally depreciated – at least on the warmth dimension – although we were unable to contrast *Muslim refugees* with *Christian refugees.*

Differences to reference groups. We were also interested in identifying the refugee subgroups' locations within the warmth by competence space in relation to societal reference groups. *Elderly people, Homeless people,* and generic *Refugees* were eligible for analysis, while *Germans* and *Rich people* had to be discarded due to non-acceptable model fit. Regarding warmth, we assumed *Elderly people,* a group particularly high on warmth (Fiske, 2018), to score highest of all groups – an assumption that was supported empirically. All groups received significantly lower warmth ratings than *Elderly people* (α_W = 1.931, rank 1). Asbrock (2010) found "foreigners" to be rated similar in warmth to "the homeless". Similarly, *Homeless people* (α_W = -0.108, rank 5), showed non-significantly different warmth ratings from generic *Refugees* (α_W = 0.000, rank 3). Yet, we found significant differences in warmth ratings of *Homeless people* to particular refugee subgroups in both directions: *Homeless people* were rated significantly less warm than *War refugees,* but significantly warmer than *Afghan refugees*, *Muslim refugees*, *Refugees from North Africa* and *Economic refugees*. Regarding generic *Refugees,* subgroups that were rated significantly less warm were *Refugees from Eritrea*, *Afghan refugees, Muslim refugees*, *Refugees from North Africa*, and *Economic refugees*.

For competence, we assumed *Homeless people* to show lowest competence assessments. This expectation was empirically supported, as *Homeless people* (α_C = -0.733, rank 10) indicated significantly lower competence ratings than any other social group. Surprisingly, all refugee subgroups received significantly lower competence ratings than *Elderly people* (α_C = 1.301, rank 1), a group that has also been associated with low competence (Fiske, 2018). Generic *Refugees* (α_C =

0.000, rank 6) were rated significantly less competent than *Economic refugees* and significantly higher in competence than *Refugees from North Africa* (α_C = -0.348, rank 9).[8]

In sum, participants rated refugee subgroups overall less warm than *Elderly people,* a reference group that had been shown to score particularly high on warmth. Most subgroups were rated less warm than generic *Refugees*; only *War Refugees* were rated warmer. *Refugees* were rated non-significantly different in warmth compared to *Homeless people*, the reference group scoring particularly low on warmth. Whereas all subgroups were rated less competent than *Elderly people*, all groups were rated more competent than *Homeless people*, both low competence reference groups.

Discussion

This paper provided multiple insights in the contemporary stereotype content of refugees in Germany. We contributed by investigating the stereotype content of refugee subgroups in the SCM framework and applying state-of-the-art methods appropriate for comparing latent means of warmth and competence. Our results indicated that the stereotype content depended on the refugee subgroup in question. Our research thereby contributed to a growing body of literature that shows that stronger nominal differentiations of groups can lead to distinct stereotype content (Binggeli et al., 2014a, 2014b; Bye et al., 2014; Clausell & Fiske, 2005; Eckes, 2002; Lee & Fiske, 2006).

Stereotype Content of Refugee Subgroups in Germany

Prior research identified many dimensions along which refugees may be categorised into subgroups (Bansak et al., 2016; Ditlmann et al., 2016; Kotzur et al., 2017). Of those, flight motives, country or region of origin, and religious affiliation emerged as meaningful organisers of the social perception of refugee subgroups in our research. By asking participants to freely nominate subgroups of refugees within a society, we complemented previous research that has focused on researcher-generated subgroup dimensions (Bansak et al., 2016; Ditlmann et al., 2016; Kotzur et al., 2017).

As for flight motives, our results regarding warmth were in line with previous research and our corresponding expectations (Kotzur et al., 2017). *War refugees* were rated highly on warmth; in fact, higher than any other refugee group, indicating high levels of benevolence (Cuddy et al., 2007). In contrast, *Economic refugees* were rated less warm (the least warm of all refugee subgroups), indicating perceptions of threat and competition, and thus elevated risk to become recipients of agony and outright rejection (Cuddy et al., 2007). Inconsistently with prior research that found no significant difference in competence ratings between both subgroups (Kotzur et al., 2017), *Economic refugees* received higher competence ratings than *War refugees*. Indeed, *Economic refugees* received the highest competence ratings of all refugee subgroups. These findings indicate that *Economic refugees* are seen as particularly capable of enacting their intentions (Fiske et al., 2002). Thus, in combination with the finding that this subgroup also rated the lowest warmth ratings suggests that *Economic refugees* are perceived as relatively skilled to enact their relatively harmful goals towards German society.

As for the country of origin, we expected that refugees from African countries and regions were rated warmer and less competent than refugees from Balkan or Middle Eastern countries (Binggeli et al., 2014b; Lee & Fiske, 2006). Whereas our expectations for competence were partially supported by the data (*Refugees from Eritrea* and *Refugees from North Africa* were rated less competent than *Syrian refugees*; *Refugees from North Africa* (yet not *Refugees from Eritrea*) were rated less competent than *Afghan refugees*), our expectations for warmth were not. Warmth assessments did either not differ significantly between these refugee groups (*Afghan refugees* and *Refugees from Eritrea*), or the differences were in the opposite direction (*Syrian refugees* were rated more highly on warmth than *Refugees from Eritrea* and *Refugees from North Africa*). These unexpected findings may be related to the general observation that outgroup perceptions, particularly for racial, ethnic, and religious groups, can be context specific (Fiske, 2017). Thus, findings from other country contexts on immigrant groups do not necessarily need to be applicable to Germany. For instance, some Middle Eastern countries are regions of armed conflicts, such as wars and civil wars (Sørli, Gleditsch, & Strand, 2005). When people think about refugees fleeing from war, many Germans first think about refugees from Middle Eastern countries, especially Syria (Kotzur et al., 2017). Since our results showed that refugees fleeing from wars and armed conflicts were rated warmer than refugees that flee for other reasons, it may thus not be surprising that *Syrian refugees* and *Afghan refugees*, fleeing from war-ridden zones, were rated non-significantly different in warmth or even warmer than refugees from African regions. Moreover, the research we based our expectations on used labels referring to larger geographical units (e.g., Africa, Middle East; Binggeli et al., 2014b; Lee & Fiske, 2006), whereas we referred to specific subregions (e.g., North Africa) and countries (e.g., Syria, Afghanistan) within these geographical units. Just like overall perceptions of generic social groups do not necessarily correspond to subgroup perceptions, perceptions of groups from larger geographical units may not correspond to perceptions of specific subgroups within these regions. Two findings supported this conclusion: The finding that two African groups, namely *Refugees from North Africa* and *Refugees from Eritrea,* were rated significantly less warm than the groups from the Middle Eastern countries, and the finding that subgroups stemming from the same geographical region also differed in their stereotype content.

Moreover, note that *Homeless people* were usually perceived as among "the lowest of the low" (Fiske, 2018, p. 68; see also Asbrock, 2010, for a German sample), receiving the lowest warmth and competence scores. Our results showed, however, that *Refugees from North Africa* were rated lower on warmth, while also being rated low on competence. Thus, refugees from this region appeared to be among the most despised subgroups we have investigated in our study. This depreciation may potentially relate to perceptions of particularly high levels of threat emanating from *Refugees from North Africa* after nationwide media reports associating this subgroup with serious criminal offences (e.g., Drüeke, 2016; Hackensberger et al., 2016).

As for religious affiliation, we expected that *Muslim refugees* would be rated less benevolently than *Christian refugees* ("The Fiske lab", n.d.). Although we had to exclude *Christian refugees* from our analyses, we found that *Muslim refugees* had an overall rather low rank in the warmth rating. *Muslim refugees* were also rated significantly less warm than generic *Refugees*. Both findings provide indirect evidence for a depreciation of refugees based on their belief – in line with findings that *Muslim refugees* would be granted asylum less often if participants were to decide (Bansak et al., 2016). Overall, then, our findings are highly compatible with societal discourses on refugees and provide an explanation of differential treatment of subgroups with different flight motives (n-tv, 2016), region or country of origin (Hackensberger et al., 2016), and religion ("CSU will christliche Zuwanderer bevorzugen", 2016).

We found slightly more observable differences between refugee subgroups on the warmth compared to the competence dimension. An explanation may be that whereas different subgroups of refugees were associated with different levels of threat and competition (predictors of warmth; Fiske et al., 2002; Kervyn et al., 2015), refugees may have been perceived as a low status, and thus low competence immigrant group (Fiske et al., 2002); a perception that additionally specified characteristics could hardly change. Indeed, institutionalised barriers and restrictions to access to the labour market ("Access to the labour market", 2018), education, ("Access to education", 2018) and other sources of status may limit the extent in which subgroups might be perceived differently on this dimension. Low status, then, may be a common and defining feature of all refugee groups we investigated.

Indeed, the analysis of the relative location to societal reference groups in the warmth by competence space corroborates this interpretation. We expected that refugee groups were rated relatively low on both warmth and competence. Our results confirmed the assumption that refugee subgroups are rated less warm than *Elderly people*, a high warmth-low competence reference group. Moreover, our results confirmed the assumption that refugee subgroups are rated more competent than *Homeless people*, a low warmth-low competence group. Thus, overall, the cultural stereotypes of refugee subgroups ranged from the low warmth-low competence area (that is, around the low warmth-low competence group *Homeless people*; see also Asbrock, 2010) up to the high warmth-low competence quadrant (that is, close to *Elderly people*, see also Asbrock, 2010).

The relative location within the warmth by competence space does not only hint at differential levels of threat, competition, and status associated with refugee subgroups, but also differential emotional and behavioural intentional consequences (Cuddy et al., 2007). That is, refugee subgroups are likely targets of either elevated contempt and harming intentions, or targets of pity and facilitative intentions, depending on their warmth and competence perceptions (high warmth-low competence or low warmth-low competence; Cuddy et al., 2007). As such, our analyses provide a first step towards identifying subgroups running elevated risks of becoming targets of hostility and aggression that allows for target group-specific social interventions to improve receiving society-refugee relations.

Methodological Advances of SCM Research

A further contribution of our work was that we applied recent and sophisticated methods to the study of stereotype content and social perception in general: By using latent variable modelling instead of analyses based on observed means, which is the dominant approach in previously published SCM literature, we computed warmth and competence scores corrected for measurement error (Cai, 2012), thus increasing reliability of our findings. Moreover, by using the alignment optimisation procedure, we also greatly strengthened the validity of our findings (Asparouhov & Muthén, 2014). Surprisingly, the basic factor structure could not be established in six out of 16 cases, indicating that the two stereotype content dimensions proposed by the SCM could not be replicated empirically in all cases. If we had relied on observed variable analysis, this fact would have remained unnoticed, which might have resulted in comparisons of scale values that would not have validly represented equal warmth and competence constructs in some cases – the well-known "comparing apples with oranges"-problem. Thus, our procedure safeguarded us from erroneously including measures in our final analysis that did not fulfil the basic criteria for mean comparisons, and ultimately protected us from drawing inappropriate conclusions regarding mean differences in warmth and competence ratings across social groups.

Using confirmatory baseline modelling resulted in substantial data reduction, which in turn decreased the information and deductions we were able to present. Thus, our and others' MGCFA findings suggest that measurement invariance is no naturally occurring scale characteristic (Janssen et al., 2015), and, consequently, should be examined carefully in all instances in SCM research. Given the novelty of applying confirmatory latent modelling procedures in SCM research, we can only speculate why some groups did not fit the baseline model. One implication would be that the global claim of the "generality across place, levels, and time" (Fiske, 2018, p. 67) of stereotype content dimensions may not be upheld without (to date unknown) boundary conditions. Therefore, we strongly recommend the increased usage of latent modelling approaches in future SCM research to cast more light on these questions and thus strengthen the empirical foundation of the theoretical framework of the SCM. We want to emphasize that we do not wish to devalue the vast body of prior SCM literature that has contributed to the knowledge about social perception in important ways. To the contrary – we hope that our contribution helps the field to move forward; towards more robust, valid and authentic research demanded for in the light of the "crisis of confidence" in psychology (Kruglanski et al., 2017, p. 1).

Applying the alignment optimisation procedure to generate latent mean values apt for meaningful cross-group comparison appears to be a promising approach: Compared to manually establishing scalar measurement invariant models in an MGCFA, the alignment optimisation is less cumbersome and produces partial measurement invariant results that are not based on the individual decision of a researcher. In line with recent criticism towards the unduly strictness of the MGCFA approach (Muthén & Asparouhov, 2013), the alignment optimisation has less stringent prerequisites, thus producing an optimised approximate measurement invariance pattern when approaches like MGCFA failed (Asparouhov & Muthén, 2014; Cieciuch et al., 2018). Alignment optimisation is especially advantageous when large numbers of groups are to be compared – given that this is usually the case in SCM research, we feel that this approach is very promising for following SCM studies.

Limitations and Future Directions

We were not able to establish a baseline measurement model for more than 35% of groups. This is astounding given that we used warmth and competence scales that had been used in SCM studies in the German context before (e.g., Asbrock, 2010; Eckes, 2002; Kotzur et al., 2017) and that were developed directly from the original SCM scales (see Eckes, 2002). We attribute this loss of data to the more rigorous statistical analyses we applied compared to previously published research (but see Janssens et al., 2015; Stanciu et al., 2017), thus increasing our findings' validity. The baseline model fit of the social groups might be improved through adapted warmth and competence scales: Firstly, including more items that tap into warmth and competence may enhance the reliability of the scales, and may allow ad-hoc adjustments (e.g., excluding certain items that underperform) to include more social groups in our analyses. We are aware that this strategy would put a strain on the overall number of groups that can be studied within one survey. However, since a higher number of groups does not help to produce more insights when their scores cannot be compared, we would like to encourage future research to increase the numbers of items when measuring warmth and competence.

What is more, the established competence items tap mainly into economic and professional competence ("competitive", "competent", "independent"). However, other competence areas may be more associated with competencies of refugee groups, including withstanding threats to their survival, such as very adverse living conditions in their regions of origins, or conditions while migrating. Thus, future research could explore whether our findings replicate with items tapping into other competence areas.

From its beginnings, stereotype content has been measured as a rating of how much participants assume a social group is viewed by most society members (Fiske et al., 2002). This strategy aims both at assessing cultural stereotypes shared by most members of a society, and at reducing social desirability bias in the data (Fiske et al., 2002). Nonetheless, this approach calls for cautious interpretation, as the results do not need to translate directly into participants' personal perceptions of social groups, but rather display what participants believe how most Germans perceive groups. Thus, our results must not be understood as individual expressions of stereotypes. Although cultural and individual stereotypes may differ from each other (Ashmore & Del Boca, 1981), these potential deviations have not yet been systematically evaluated in the SCM framework. Especially in the important context of the social perception of (subgroups of) refugees, such a direct comparison appears a worthy goal of future research.

Further limitations relate to the sample composition: Like in numerous SCM studies and comparative research before us (Asbrock, 2010), we based our research on (typically young and liberal) student samples of two specific universities. Previous research found that student samples, as well as representative samples, support the SCM structure (Fiske, 2015): Both student and representative samples lead to similar conclusions regarding the shared perception of social groups within a given society. Nonetheless, we would welcome studies that base their research on large representative samples to test whether this is also the case stereotype content of refugee subgroups.

Similar to prior research in this domain, we used a cross-sectional design. Although such designs are efficient to investigate the social perceptions of groups at a given time, they do not provide any indication regarding the stability of findings. The stereotype content of groups might change over time for several reasons; one of them being that intergroup contact with refugee groups can help to enhance warmth and competence assessments (Kotzur, Schäfer, & Wagner, 2018) - that is ever more likely to occur in refugee-receiving countries like Germany (Bundesamt für Migration und Flüchtlinge, 2019). Thus, we recommend the repeated assessment of social perception of refugee subgroups in Germany, ideally on a longitudinal basis.

As previously stated, SCM findings from one country context do not necessarily need to generalize to another ("The Fiske lab", n.d.). Moreover, some researchers found within-country regional differences in the endorsement of warmth and competence of social groups (Binggeli et al., 2014a; Stanciu et al., 2017). Although our data stemmed two different research sites within

Germany, we found no such differences between samples (see additional analyses), although power for such comparative analyses was admittedly limited (n_g < 100 for some of the samples; Asparouhov & Muthén, 2014). Therefore, we encourage future studies investigating the social perception of refugee subgroups in other countries that are both destinations and origins of refugees, as well as potential within-country differences of the social perception of refugees. Such studies would further the understanding of the particularities of the social perception of this social group of high social relevance within and across cultures and nations.

Conclusion

In this study, we investigated the stereotype content of subgroups of refugees in Germany. Generic refugees were rated as lacking warmth and competence. Subgroup assessments differed significantly, depending on the insinuated flight motive, region or country of origin, or religious affiliation. Overall, the subgroups' warmth and competence ratings ranged from low warmth/low competence to high warmth/low competence. We produced these insights using alignment optimisation, an appropriate state-of-the-art method to compare multiple latent means. Given its relative user-friendliness and more realistic approach to measurement invariance compared to conventional methods, we hope will be adopted by others in this research field.

Footnotes

[1] In German: “Kriegsflüchtling”, a term commonly used in Germany to refer to people that seek refuge due to war and civil war.

[2] In German: “Wirtschaftsflüchtling”, a term commonly used in Germany to refer to people that seek refuge due to economic hardship.

[3] We calculated differences for these social groups based on the information Binggeli et al., (2014b) provided in Table 1, p. 128.

[4] Although the authors are fully supportive of the aims and strategies for study preregistration of the Center for Open Science and other initiatives, we refrained from pre-registering these predictions as (I) they mostly stem from published research that was not based on the Stereotype Content Model framework (except for Kotzur et al., 2017) or that partially relied on the more general target groups (immigrants; Binggeli et al., 2014b, Lee & Fiske, 2006); and (II) due to the descriptive (i.e., with reference to the relative location of refugee subgroups compared to reference groups in the two-dimensional SCM framework) and exploratory (i.e., with reference to what societally relevant refugee subgroups would be generated) character of the study at hand.

[5] In both subsamples, we excluded participants that provided answers for 50% or less of the variables (n_1 = 6; n_2 = 46), that were multivariate outliers as identified by Mahalanobi’s distance (n_2 = 13), identified with at least one of the surveyed outgroups (n_1 = 3; n_2 = 11), had non-German nationality (n_1 = 1; n_2 = 6), or did not reside mainly in the respective area where the university was located (n_1 = 4; n_2 = 1).

[6] The subsample 1 was 81.0% female, 1.3% other; ; M_{age} = 23.49 years, SD_{age} = 4.76, Min$_{age}$ = 19, Max$_{age}$ = 50; 98.7% University students, 1.3% other; 88.6% without migration background. The subsample 2 was 68.6% female, 1.6% other; M_{age} = 24.52 years, SD_{age} = 4.57, Min$_{age}$ = 19, Max$_{age}$ = 50; 94.1% University students, 6.0% other; 83.2% without migration background, 1.1% missing; 0.5% missing nationality.

[7] For *Homeless people,* the intercepts of the competence item “independent” as well as the warmth item “likeable” were non-invariant.

[8] Although the sample size is very low for such analyses (below n = 100 for subsample 1; Asparouhov & Muthén, 2014), we conducted all presented analyses also in each of the subsamples separately to rule out that the social perception of investigated groups differed systematically between sampling sites. The findings can be found in Tables SM6 and SM7 (for the baseline measurement model fit of all social groups in subsample 1 and 2, respectively), Tables SM8 and SM9 (for the results of the alignment optimisation approach for warmth and competence, respectively), Table SM10 (for further information on the alignment optimisation model) in the online supplementary material: https://osf.io/y76kq/. The analysis output can be found in the folder “Additional analysis: Separated for research location”, e.g., https://osf.io/4h75y/. No substantial differences emerged when comparing warmth and competence ratings of the same subgroups across subsamples. Thus, these additional analyses supported the findings presented above, lending further support for the robustness of our findings.

References

"Access to education". (2018). Retrieved from http://www.asylumineurope.org/reports/country/germany/reception-conditions/employment-education/access-education

"Access to the labour market". (2018). Retrieved from: http://www.asylumineurope.org/reports/country/germany/reception-conditions/employment-education/access-labour-market

Asbrock, F. (2010). Stereotypes of social groups in Germany in terms of warmth and competence. *Social Psychology, 41*(2), 76-81. doi:10-1027/1864-9335/a000011

Ashmore, R. D., & Del Boca, F. K. (1981). Conceptual approaches to stereotypes and stereotyping. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 1-36). Hove, UK: Psychology Press.

Asparouhov, T., & Muthén, B. O. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling, 21*, 1-14. doi: 10.1080/10705511.2014.919210

Bade, K. J. (1992). *Deutsche im Ausland – Fremde in Deutschland [Germans abroad – foreigners in Germany]*. München: Beck.

Bansak, K., Hainmueller, J., & Hangartner, D. (2016). How economic, humanitarian, and religious concerns shape European attitudes toward asylum seekers. *Science, 354*(6309), 217-222. doi:10.1126/science.aag2147

Bentler, P. M., & Chou, C. H. (1987). Practical issues in structural modeling. *Sociological Methods & Research, 16*, 78-117.

Binggeli, S., Krings, F., & Sczesny, S. (2014a). Perceived competition explains regional differences in the stereotype content of immigrant groups. *Social Psychology, 45*(1)*,* 62-70. doi:10.1027/1864-9335/a000160

Binggeli, S., Krings, F., & Sczesny, S. (2014b). Stereotype content associated with immigrant groups in Switzerland. *Swiss Journal of Psychology, 73*(3), 123-133. doi:10.1024.1421-0815/a000133

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: The Guilford Press.

Bundesamt für Migration und Flüchtlinge (2019). *Aktuelle Zahlen zu Asyl.* Retrieved from http://www.bamf.de/SharedDocs/Anlagen/DE/Downloads/Infothek/Statistik/Asyl/aktuelle-zahlen-zu-asyl-dezember-2018.html?nn=7952222

Bundesministerium des Innern, für Bau und Heimat (2016, September 30). *Pressemitteilung – 890.000 Asylsuchenden im Jahr 2015*. Retrieved from https://www.bmi.bund.de/SharedDocs/pressemitteilungen/DE/2016/09/asylsuchende-2015.html

Burkley, E., Durante, F., Fiske, S. T., Burkley, M., & Andrade, A. (2017). Structure and content of Native American stereotypic subgroups: Not just (ig)noble. *Cultural Diversity and Ethnic Minority Psychology, 23*(2), 209-219. doi:10.1037/cdp0000100

Bye, H. H., Herrebrøden, H., Hjetland, G. J., Røyset, G. Ø., & Westby, L. L. (2014). Stereotypes of Norwegian social groups. *Scandinavian Journal of Psychology, 55*(5), 469-476. doi:10.1111/sjop.12141

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema, 20*(4), 872-882.

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.

Cai, L. (2012). Latent variable modeling. *Shanghai Archives of Psychiatry, 24*(2), 118-120. doi:10.3969/j.issn.1002-0829.2012.02.010

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:10.1037/a0013193

Cieciuch, J., Davidov, E., & Schmidt, P. (2018). Alignment optimization: Estimation of the most trustworthy means in cross-cultural studies even in the presence of noninvariance. In E. Davidov, P. Schmidt, & J. Billiet J. (Eds.), *Cross Cultural Analysis: Methods and Applications* (2nd ed., pp. 571-592). NY: Routledge.

Clausell, E., & Fiske, S. T. (2005). When do subgroup parts add up to the stereotypic whole? Mixed stereotype content for gay male subgroups explains overall ratings. *Social Cognition, 23*(2), 161-181. doi:10.1521/soco.23.2.161.65626

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology, 92*(4), 631-648. doi:10.1037/0022-3514.92.4.631

Cuddy, A. J. C., Fiske, S. T., Kwan, V. S. Y., Glick, P., Demoulin, S., Lexens, J.-P., … Ziegler, R. (2009). Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology, 48*(1), 1-33. doi:10.1348/014466608X314935

"CSU will christliche Zuwanderer bevorzugen" [Christian social union wants to privilege Christian immigrants]. (2016, September 08). *ZDF heute.* Retrieved from http://www.heute.de/papier-zur-fluechtlingspolitik-csu-will-christliche-zuwanderer-bevorzugen-45133634.html

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

Ditlmann, R., Koopmans, R., Michalowski, I., Rink, A., & Veit, S. (2016). Verfolgung vor Armut: Ausschlaggebend für die Offenheit der Deutschen ist der Fluchtgrund [Persecution for poverty: Decisive for the openness of Germans is the reason for flight]. *WZB Mitteilungen, 151*, 1–27. Retrieved from https://www.wzb.eu/sites/default/files/publikationen/wzb_mitteilungen/veits24-27151-webpdf-2.pdf

Drüeke, R. (2016). *Die TV-Berichterstattung in ARD und ZDF über die Silvesternacht 2015/2016 in Köln [Media analysis: Public television coverage of the Cologne New Year's Eve 2015/16]*. Gunda-Werner-Institut für Feminismus und Geschlechterdemokratie der Heinrich Böll Stiftung. Retrieved from www.gwi-boell.de/sites/default/files/web_161122_e-paper_gwi_medienanalysekoeln_v100.pdf

Durante, F., Fiske, S. T., Kervyn, N., Cuddy, A. J., Akande, A. D., Adetoun, B. E., ... Barlow, F. K. (2013). Nations' income inequality predicts ambivalence in stereotype content: How societies mind the gap. *British Journal of Social Psychology, 52*(4), 726-746. doi: 10.1111/bjso.12005

Durante, F., Fiske, S. T., Gelfand, M. J., Crippa, F., Suttora, C., Stillwell, A., … Teymoori, A. (2017). Ambivalent stereotypes link to peace, conflict, and inequality across 38 nations. *Proceedings of the National Academy of Sciences of the United States of America, 114*(4), 669-674. doi: 10.1073/pnas.1611874114

Eckes, T. (2002). Paternalistic and envious gender stereotypes: Testing predictions from the stereotype content model. *Sex Roles*, *47*(3-4), 99-114. doi:10.1023/A:1021020920715

Eurostat. (2016). *Asyl in den EU-Mitgliedsstaaten – Rekordzahl von über 1,2 Millionen registrierten erstmaligen Asylbewerbern im Jahr 2015* [Asylum in EU member states – record number of more than 1.2 million registered first-time asylum seekers in 2015]. Retrieved from http://ec.europa.eu/eurostat/documents/2995521/7203837/3-04032016-AP-DE.pdf/9fcd72ad-c249-4f85-8c6d-e9fc2614af1b

Eurostat. (2018). *– Asylum report quarterly – First time asylum applicants and first instance decisions on asylum decisions: Fourth quarter 2017.* Retrieved from http://ec.europa.eu/eurostat/statistics-explained/index.php/Asylum_quarterly_report#cite_note-3

Fiske, S. T. (2015). Intergroup biases: A focus on stereotype content. *Current Opinion in Behavioral Sciences, 3*, 45-50. doi:10.1016/j.cobeha.2015.01.010

Fiske, S. T. (2017). Prejudice in cultural contexts: Shared stereotypes (gender, age) versus variable stereotypes (race, ethnicity, religion). *Perspectives on Psychological Science, 12*(5), 791-799. doi:10.1177/1745691617708204

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science, 27*(2), 67-73. doi:10.1177/0963721417738825

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:10.1037/0022-3514.82.6.878

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576. doi:10.1146/annrev.psych.58.110405.085530

Hackensberger, A., Kalnoky, B., & Smirnova, J. (2016, June 09). *Warum Flüchtlinge aus diesen Ländern oft kriminell werden [Why refugees from these countries often become delinquent]. Die Welt*. Retrieved from https://www.welt.de/politik/ausland/article156077320/Warum-Fluechtlinge-aus-diesen-Laendern-oft-kriminell-werden.html

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

Juran, S., & Broer, P. N. (2017). A profile of Germany's refugee populations. *Population and Development Review*, *43*(1), 149-157. doi:10.1111/padr.12042

Kervyn, N., Fiske, S., & Yzerbyt, V. (2015). Forecasting the primary dimension of social perception*. Social Psychology, 46*(1), 36-45. doi:10.1027/1864-9335/a000219

Kline, R. B. (2010). *Principles and practise of structural equation modelling* (3rd ed.). New York: Guilford.

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:10.1027/1864-9335/a000312

Kotzur, P. F., Schäfer, S. J., & Wagner, U. (2018). Meeting a nice asylum seeker: Intergroup contact changes stereotype content perceptions and associated emotional prejudice, and encourages solidarity-based collective action intentions. *British Journal of Social Psychology.* Advance online publication. doi: 10.1111/bjso.12304*.*

Kotzur, P. F., Tropp, L. R., & Wagner, U. (2018). Welcoming the unwelcome: How contact shapes contexts of reception for new immigrants in Germany and the United States. *Journal of Social Issues, 74*(4), 812-832. doi:10.1111/josi.12300

Kruglanski, A. W., Chernikova, M., & Jasko, K. (2017). Social psychology circa 2016: A field on steroids. *European Journal of Social Psychology, 47*(1), 1-10. doi:10.1002/ejsp.2285

Lee, T. F., & Fiske, S. T. (2006). Not an outgroup, not yet an ingroup: Immigrants in the stereotype content model. *International Journal of Intercultural Relations, 30*(6), 751-768. doi:10.1016/j.ijintrel.2006.06.005

Muthén, B. O., & Asparouhov, T. (2013, January 11). BSEM measurement invariance analysis. *Mplus Web Notes* (Number 17). Retrieved from http://www.statmodel.com/examples/webnotes/webnote17.pdf

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

n-tv (2016). *Athen: Kaum noch Kriegsflüchtlinge.* [Athens: Hardly any war refugees anymore] Retrieved from http://www.n tv.de/politik/Athen-Kaum-noch-Kriegsfluechtlinge-article17144711.html

Sadler, M. S., Meagor, E. L., & Kaye, K. E. (2012). Stereotypes of mental disorders differ in competence and warmth. *Social Science & Medicine, 74*(6), 915-922. doi:10.1016/j.socscimed.2011.12.019

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. doi:10.1016/j.ssresearch.2003.11.003

Sørli, M. E., Gleditsch, N. P., & Strand, H. (2005). Why is there so much conflict in the Middle East?. *Journal of Conflict Resolution, 49*(1), 141-165. doi:10.1177/0022002704270824

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

Stanciu, A., Cohrs, J. C., Hanke, K., & Gavreliuk, 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:10.1080/00224545.2016.1262812

"The Fiske lab: Intergroup relations, social cognition, and social neuroscience". (n.d.) Retrieved from www.fiskelab.org/cross-cultural-wc-maps

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. O. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology, 4,* 770. doi:10.3389/fpsyg.2013.00770

van de Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement invariance. *Frontiers in Psychology, 6*, 1064. doi:10.3389/fpsyg.2015.01064

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:10.1177/109442810031002

Tables and Figures

Table 1

Single Group Confirmatory Factor Analysis Model Fit

#	Group	*N*	χ^2	*df*	*p*	χ^2/df	*RMSEA*	*SRMR*	*CFI*
1	Refugees	264	18.008	8	.021	2.251	.069	.037	0.975
2	Christian Refugees	188	18.859	8	.016	2.357	.085	.048	0.955
3	Syrian Refugees	188	11.104	8	.196	1.388	.045	.033	0.991
4	Germans	264	27.566	8	<.001	3.446	.096	.043	0.940
5	Turkish Migrants	264	24.084	8	.002	3.011	.087	.035	0.967
6	Muslim Refugees	264	11.808	8	.160	1.476	.042	.030	0.991
7	Afghan Refugees	188	15.376	8	.054	1.922	.070	.039	0.977
8	Rich People	264	38.872	8	<.001	4.859	.121	.068	0.881
9	War Refugees	174	7.219	8	.513	0.902	.000	.029	1.000
10	Refugees from the Balkans	174	23.923	8	.002	2.990	.107	.044	0.953
11	Iraqi Refugees	174	30.485	8	<.001	3.811	.127	.057	0.929
12	Elderly People	264	15.995	8	.043	1.999	.062	.040	0.969
13	Economic Refugees	166	11.044	8	.199	1.381	.048	.035	0.988
14	Refugees from Eritrea	166	3.145	8	.925	0.393	.000	.019	1.000
15	Refugees from North Africa	166	3.226	8	.917	0.403	.000	.017	1.000
16	Homeless People	264	20.019	8	.010	2.502	.075	.038	0.962

Note. N = Number of participants; *df* = degrees of freedom; *p* = probability value; *RMSEA* = root mean square error of approximation; *SRMR* = standardized root mean square residual; *CFI* = comparative fit index. Acceptable model fit is indicated if all following requirements were fulfilled:

$\chi^2/df < 3$; *RMSEA* < .08; *SRMR* < .10; *CFI* > .95 (Schermelleh-Engel et al., 2003). *Christian refugees, Germans, Turkish migrants, Rich people, Refugees from the Balkans*, and *Iraqi refugees* indicated poor model fit.

Table 2

Rank Order of Latent Mean Values for Warmth Assessment Across Social Groups

Rank	#	Group	Latent Mean Value	Groups with Significantly Smaller Factor Means
1	12	Elderly People	1.931	9, 1, 3, 16, 14, 7, 6, 15, 13
2	9	War Refugees	0.548	1, 3, 16, 14, 7, 6, 15, 13
3	1	Refugees†	0.000	14, 7, 6, 15, 13
4	3	Syrian Refugees	0.000	14, 7, 6, 15, 13
5	16	Homeless People	-0.108	7, 6, 15, 13
6	14	Refugees from Eritrea	-0.335	6, 15, 13
7	7	Afghan Refugees	-0.375	6, 15, 13
8	6	Muslim Refugees	-0.639	13
9	15	Refugees from North Africa	-0.854	
10	13	Economic Refugees	-1.105	

Note. Significance testing was conducted at a 5% significance level (two-sided). †Due to the fixed alignment optimisation model, this mean value was constrained to be zero.

Table 3

Rank Order of Latent Mean Values for Competence Assessment Across Social Groups

			Latent Mean	Groups with Significantly
Rank	#	Group	Value	Smaller Factor Means
1	12	Elderly People	1.301	13, 3, 6, 9, 1, 7, 14, 15, 16
2	13	Economic Refugees	0.559	3, 6, 9, 1, 7, 14, 15, 16
3	3	Syrian Refugees	0.216	14, 15, 16
4	6	Muslim Refugees	0.043	14, 15, 16
5	9	War Refugees	0.015	15, 16
6	1	Refugees†	0.000	15, 16
7	7	Afghan Refugees	-0.018	15, 16
8	14	Refugees from Eritrea	-0.214	16
9	15	Refugees from North Africa	-0.348	16
10	16	Homeless People	-0.733	

Note. Significance testing was conducted at a 5% significance level (two-sided). †Due to the fixed alignment optimisation model, this mean value was constrained to be zero.

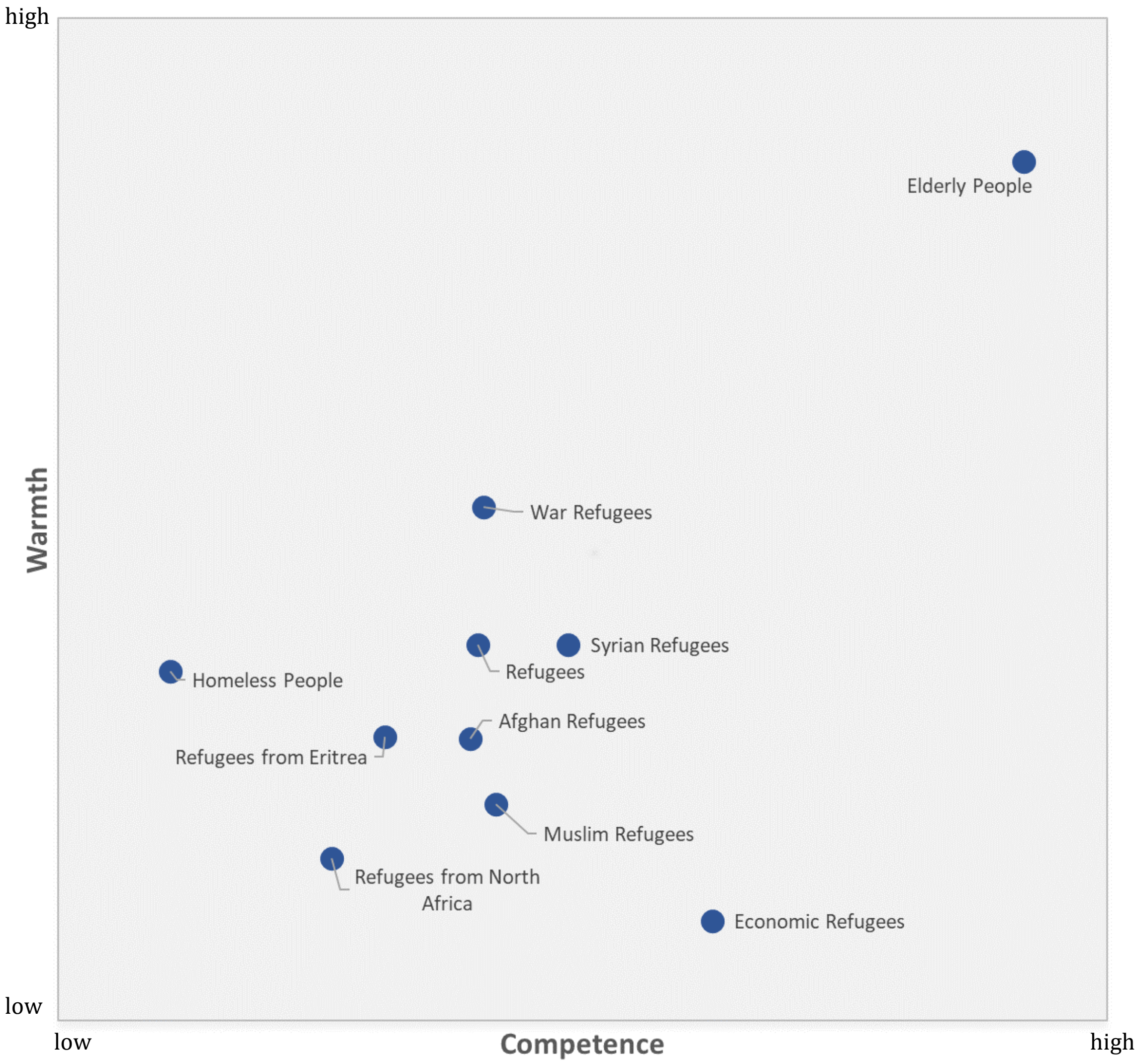


Figure 1. Latent warmth (Y-axis) and competence (X-axis) mean values for social groups. Scaling was achieved by constraining the latent mean values of warmth and competence of the group *Refugees* to zero.

Supplementary Materials for Manuscript # 3

The supplementary materials for Manuscript # 3 are stored in the Open Science Framework, see https://osf.io/5j7t6/.

They include:

- The used questionnaire of the main study;
- The open answers of the pilot study;
- An overview of the allocation of the target groups to the planned missingness design;
- Overall sample descriptive statistics;
- A detailed overview of the estimated parameters in the alignment optimisation procedure;
- Target-group-specific descriptive statistics and intercorrelations of the warmth and competence indicators;
- Warmth and competence scale intercorrelations between the different target groups;
- Sample-specific confirmatory factor analysis results;
- Results of the alignment optimisation procedure for both samples separately;
- A detailed overview of the estimated parameters in the alignment optimisation procedure for both samples separately;
- Analysis outputs for all analysis reported in the manuscript;
- Analysis outputs for the additional analysis separating both samples.

Preregistration for Manuscript # 4

The preregistration was competed using an AsPredicted template on April 27, 2020.

OSF preregistration link: https://osf.io/pmjgf/?view_only=353ea72e07fc4cb7b33b0beba8fe4842

Have any data been collected for this study already? Note: 'Yes' is a discouraged answer for this preregistration form.

- Yes, we already collected the data.
- **No, no data have been collected for this study yet.**
- It's complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid pre-registration nevertheless.

What's the main question being asked or hypothesis being tested in this study? (optional)

The present research project will investigate the social perception of different occupational groups and occupation-related social groups in Germany applying the Stereotype Content Model (SCM; Fiske et al., 2002). The SCM proposes that the social perception of social groups in general can be attributed to evaluations on two fundamental dimensions, which are warmth (i.e., the intentions of the other group; Fiske et al., 2007), and competence (i.e., the ability to act on those intentions; Fiske et al., 2007). The surveyed occupational and social groups are: Bankers, child care workers, craftsmen, farmers, firefighters, hospital and elderly care nurses, judges, physicians, police officers, politicians, retirees, teachers, and unemployed people (all using gender-neutral labels).

In the tradition of previous SCM literature, our research focusses on the exploration of latent mean differences in social perception of the surveyed occupational and social groups in the two-dimensional warmth and competence space. More specifically, we formulate the following expectations:

- E1: The used items will allow for latent modelling of warmth and competence factors, which is expressed in acceptable model fits of confirmatory factor analyses of the warmth and competence measurement model in all surveyed occupational and social groups.

- E2: There will be significant latent mean differences in perceived warmth and competence between the surveyed occupational and social groups (Fiske et al., 2002).

Previous SCM studies have identified well-investigated “typical” social groups which score in all four extremes of the two-dimensional warmth-competence-space. The “typical” groups and their expected locations are: “Physician” for high warmth - high competence (e.g., Asbrock, 2010; He et al., 2019), “retirees” for high warmth - low competence (e.g., Asbrock, 2010; Eckes, 2002), “bankers” for low warmth – high competence (e.g., He et al., 2019), and “unemployed people” for low warmth – low competence (e.g., Asbrock, 2010, He et al., 2019). These “typical” groups will serve as anchor groups in our research, against whose ratings we will compare the ratings of all other occupational groups. Given these previous findings, we further formulate the following expectations:

- E3: The social group “retirees” will be rated high in warmth and low in competence compared to the other surveyed occupational and social groups.
- E4: The social group “unemployed people” will be rated low in both warmth and competence compared to the other surveyed occupational and social groups.
- E5: The occupational group “physicians” will be rated high in both warmth and competence compared to the other surveyed occupational and social groups.
- E6: The occupational group “teachers” will be rated high in both warmth and competence compared to the other surveyed occupational and social groups and will not show significant differences in terms of warmth and competence compared to “physicians”.
- E7: The occupational group “politicians” will be rated less positive in warmth compared to “physicians”, “teachers” and “retirees” (Eckes, 2002; Wagner et al., in press).

Concerning the occupational group “politicians”, conflicting evidence has been presented concerning the perceived competence: While Eckes (2002) found relatively high competence ratings, Wagner et al. (in press) found very low competence ratings. Thus, the competence rating of politicians, as well as the ratings of the remaining occupational groups in terms of both warmth and competence, will be explored openly without previous expectations.

Describe the key dependent variable(s) specifying how they will be measured. (optional)

Our key dependent variables are the warmth and competence ratings of the different occupational and social groups. We will assess warmth and competence from a societal perspective (i.e., "from the perspective of most Germans"; Fiske et al., 2002). Warmth and competence will be measured by four semantic differential scales with five points respectively: For warmth, the items are the German equivalents of "*honest - dishonest", "friendly - unfriendly", "good-natured - ill-natured", "warm - cold";* for competence, the items are the German equivalents of *"thorough – careless", "competent - incompetent", "hard-working - lazy", "efficient - inefficient"*. These items are an adapted and extended version of the German SCM scale by Asbrock (2010).

How many and which conditions will participants be assigned to? (optional)

The survey does not contain any experimental manipulation requiring participants' assignment into different conditions. However, to reduce participants' strain, we will apply a planned-missing-data design (three-form design; Graham, 2009; Graham et al., 2006), which allocates participants to different conditions in which a subset of occupational and social groups was excluded, thus reducing the survey length. All participants will rate the four anchor groups (i.e., "physicians", "banker", "retirees", "unemployed people"). The combination of the remaining occupational groups into excluded subsets is as follows: (1) "craftsmen", "firefighters" and "police officers"; (2) "hospital and elderly care nurses", "child care workers" and "judges"; (3) "teachers", politicians", and "farmers".

Specify exactly which analyses you will conduct to examine the main question/hypothesis. (optional)

We aim to assess mean value differences, thus correcting for measurement error. Thus, our analyses will include the following steps:

1. Confirmatory factor analyses of the measurement models will be computed for each occupational and social group separately. The measurement model will include a latent warmth and a latent competence factor which may be correlated with each other, and which each load onto the observed indicator items listed above. No cross-loadings or residual covariations will be allowed. Acceptable model fit will be based on the criteria of Schermelleh-Engel et al. (2003): Root mean standard error of approximation (RMSEA) < .08, standardised root mean square residual (SRMR) < .10, comparative fit index (CFI) > .95.
2. All occupational and social groups with acceptable baseline model fit will be subjected to the fixed alignment optimization procedure (Asparouhov & Muthén, 2014). Alignment optimization serves the dual purpose of discovering the most optimal pattern of scalar measurement invariance (thereby assuring comparability of the measurement models as a precondition of latent mean value comparison, Davidov et al., 2014) as well as computing latent mean value differences in warmth and competence for all occupational and social groups. The findings will be assumed to be acceptable if the presented alignment solution features less than 25% non-invariant parameters (Asparouhov & Muthén, 2014).

Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.

The data will be screened for implausible values, which will be coded as missing data. Uni- or multivariate outlier values will not be examined; instead, we will use a robust maximum likelihood estimator (MLR in Mplus) to account for non-normality and non-independence in the data (Muthén & Muthén, 1998-2017). What is more, participants will be asked at the beginning if they identify with one or more of the surveyed occupational and social groups: If participants indicate to self-identify with one of the surveyed groups, we will exclude the ratings of that particular group (pairwise deletion) to avoid potential ingroup bias.

How many observations will be collected or what will determine the sample size? No need to justify decision, but be precise about exactly how the number will be determined. (optional)

We will survey a minimum of 300 participants (i.e., at least 200 observations per occupational and social group due to the planned-missingness design), as this is an adequate sample size to conduct the alignment optimization procedure (Asparouhov & Muthén, 2014).

Anything else you would like to pre-register? (e.g., data exclusions, variables collected for exploratory purposes, unusual analyses planned?) (optional)

Treatment of missing data

We might explore the usage of multiple imputation to fill the missing data points created by the planned-missingness design. This is dependent on the compatibility of multiple imputed data with the subsequent analyses.

Exploration of clustering

Many prominent SCM studies (e.g., Asbrock, 2010; Eckes, 2002; Fiske et al., 2002) present *k*-means cluster analyses to identify clusters of social groups which are perceived similar in terms of warmth and competence and which are distinct from other social groups in other clusters. We might explore the applicability of equivalent latent analyses, such as latent profile analyses or factor mixture models.

Exploration of robustness effects

We might explore the robustness of our findings by comparing the results of a sample in which ingroup-bias is reduced (i.e., by excluding the ratings of self-identified social or occupational groups) against the results of the whole sample. This is based on the reasoning of Fiske et al. (2002) that the used societal perspective reflects societally shared rather than personal stereotypes, which are not affected by ingroup bias.

Additional variables in the survey

Additionally to the warmth and competence measures for the different occupational and social groups, the survey will contain a number of demographic information as well as two items assessing the gender typicality of the surveyed occupational groups. These variables are irrelevant for the study at hand.

Pilot test to identify occupational groups

The surveyed occupational and social groups were chosen based on a literature review, a pilot study, and due to current political and societal relevance in Germany (see procedure in Fiske et al., 2002).

Sample and recruitment

The data will be collected using an Unipark online survey targeting a heterogeneous adult sample. Participants will be compensated for their participation by donating 50ct to one of four charity organisations of the participants' choice.

The usage of the word "expectation"

When formulating the research questions, we chose the term "expectation" instead of the more commonly used terms "hypothesis" or "exploratory question". The reason for that is that, on the one hand, we evaluate the theoretical basis of our research not strong enough for strict hypotheses, as all referenced research has either been collected in another national context (He et al., 2019) or more than ten years ago in the German context (Asbrock, 2002; Eckes, 2002), and thus, applicability to the current German context is limited. On the other hand, the previous research is too manifold to justify purely exploratory research without any assumptions.

Literature:

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

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

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

Eckes, T. (2002). Paternalistic and envious gender stereotypes: Testing predictions from the stereotype content model. *Sex Roles*, *47*(3-4), 99-114. https://doi.org/10.1023/A:1021020920715

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. 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. https://doi.org/10.1037/0022-3514.82.6.878

Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, *60*(1), 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323–343. https://doi.org/10.1037/1082-989X.11.4.323

He, J. C., Kang, S. K., Tse, K., & Toh, S. M. (2019). Stereotypes at work: Occupational stereotypes predict race and gender segregation in the workforce. *Journal of Vocational Behavior*, *115*, 103318. https://doi.org/10.1016/j.jvb.2019.103318

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). *Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures*. *8*(2), 52.

Wagner, U., Friehs, M.-T., & Kotzur, P. F. (in press). Das Bild der Polizei bei jungen Studierenden [The image of the police in the eyes of German students]. *Polizei und Wissenschaft*.

Give a title for this AsPredicted pre-registration. Suggestion: use the name of the project, followed by study description.

Occupational Stereotypes in Germany in Terms of Warmth and Competence

For record keeping purposes, please tell us the type of study you are pre-registering.

- Class project or assignment
- Experiment
- **Survey**
- Observational/archival study
- Other (describe below)

Manuscript # 4

Stereotype Content of Occupational Groups in Germany

Maria-Therese Friehs[1], ORCID: 0000-0002-5897-8226

Felicia Aparicio Lukassowitz[2]

Ulrich Wagner[3]

[1] Department of Psychology, University of Koblenz-Landau, Germany

[2] Institute for Psychology, Osnabrück University, Germany

[3] Department of Psychology, Philipps-University Marburg, Germany

Corresponding author:

Maria-Therese Friehs, University of Koblenz-Landau, Department of Psychology, Developmental and Educational Psychology, Fortstraße 7, 76829 Landau, Germany, friehsm@uni-landau.de

Stereotype Content of Occupational Groups in Germany

Open Science Framework Project:

https://osf.io/gxz49/?view_only=5f6a34d009ae4741ad1dd4142ffd8d7d

Preregistration:

https://osf.io/pmjgf/?view_only=32e35099f0b2450abe4e4c3c48d88632

Word Count:

Abstract: 140

Title, Main text, Notes: 8395

Graphics: 5 Tables, 2 Figures

Supplementary Materials: 12 documents labelled “OSM A” to “OSM K”

Abstract

The Stereotype Content Model (SCM) is a prominent model of social perception, proposing two universal dimensions of evaluation: Warmth and competence. Occupational stereotypes have rarely been assessed in this model, though they have an important impact on how individuals experience gainful occupation and navigate everyday social interactions. Responding to recent methodological critiques regarding the SCM's scale performance, we developed a context- adapted, well-performing German-language SCM scale and assessed warmth and competence ratings of 13 occupational groups in a heterogeneous sample. Using the alignment optimisation procedure to allow for more reliable latent mean value comparisons, we found occupational stereotypes to differ significantly, with *Firefighters* presenting the most favourable and *Politicians* and *Unemployed people* showing the least favourable evaluations. We discuss our findings in terms of their content-wise and methodological meaning as well as their implications for research and in occupational contexts.

Keywords: Occupational Stereotypes, Stereotype Content Model, Factor Analyses, Scale Development, Alignment Optimisation

Stereotype Content of Occupational Groups in Germany

Theoretical Background

The Stereotype Content Model (SCM)

At all times, humans are required to navigate the social world. On the one hand, this involves constant social evaluation processes of oneself, other individuals, or social in- and outgroups. On the other hand, information-reduction strategies are required to avoid cognitive overload (Allport, 1954; Abele, Ellemers, Fiske, Koch, & Yzerbyt, 2020; Macrae & Bodenhausen, 2000). Avoidance of cognitive overload may be achieved through processes of categorisation and stereotyping (i.e., beliefs of characteristic traits of typical group members; Dovidio, Hewstone, Glick, & Esses, 2010). Individuals may be categorised into social groups based on a variety of characteristics, such as gender, age, race, and occupation (Imhoff, Koch, & Fade, 2018), and these categories are connected with specific attributes (i.e., stereotypes).

The Stereotype Content Model (SCM; Fiske, Cuddy, Glick, & Xu, 2002) is a prominent model describing these stereotyping processes. It assesses the evaluation of different social groups from a shared societal perspective and, in line with various other theories of social perception, proposes two basic dimensions of social perception (Abele, Cuddy, Judd, & Yzerbyt, 2008; Abele & Wojciszke, 2007; Abele et al., 2020; Yzerbyt, Provost, & Corneille, 2005). These dimensions are *warmth*, defined as the intentions of the other group, and *competence*, meaning the group's ability to act on those intentions (Fiske, Cuddy, & Glick, 2007). Social structure predicts stereotype content, with competition and threat negatively predicting warmth assessments, and status positively predicting competence perceptions (Fiske, 2015, 2018; Fiske et al., 2002; Kervyn, Fiske, & Yzerbyt, 2015). Evaluations on these dimensions are assumed to be independent of each other (Fiske et al., 2007). Being the cognitive facet of attitudes, stereotype content is proposed to predict emotional (i.e., feelings of pride, envy, contempt, pity) and behavioural responses (i.e., active and passive, facilitative and harmful behaviour; Cuddy, Fiske, & Glick, 2007). The applications of the SCM are manifold, focusing mostly on the description of stereotype content of social groups of high societal relevance (including, but not exclusive to occupations) in various countries (e.g., Asbrock, 2010; Eckes, 2002; Durante et al., 2013; The Fiske Lab, n.d.). In Germany, the SCM has been applied descriptively to explore the stereotype content of different social groups (e.g., Asbrock, 2010; Eckes, 2002; Kotzur, Friehs, Asbrock, & van Zalk, 2019), as well as experimentally to compare different target groups or instructions (e.g., Imhoff, Woelki, Hanke, & Dotsch, 2013; Kotzur, Veit, Namyslo, Holthausen, Wagner, & Yemane, 2020).

Warmth and competence have been proposed to be universal and fundamental dimensions of social perception "across perceivers, stimuli, cultures, and time" (Cuddy, Fiske, & Glick, 2008, p. 137; Fiske, 2018; Fiske et al., 2007). Methodologically, this claim would be achieved if the SCM measures would result in warmth and competence scales that, statistically independent from one another, assess all kinds of target groups in all cultures and contexts validly, reliably, and independent of the raters' characteristics. However, this assumption of universality was empirically challenged: Firstly, because the items used to assess warmth and competence often did not form reliable scales of acceptable dimensionality in confirmatory factor analyses (Blinded for peer review A, 2020; Blinded for peer review B, 2020; Kotzur et al., 2019, 2020). And secondly, because the statistical preconditions for meaningful (latent) mean value comparison (i.e., (partial) scalar measurement invariance; Vandenberg & Lance, 2000) are oftentimes not fulfilled (Blinded for peer review A, 2020; Blinded for peer review B, 2020; Kotzur et al., 2020). In the absence of (partial) scalar measurement invariance, comparisons on SCM dimensions between social groups are biased because the scales' conceptual understanding, measurement units, or points of zero may differ between groups (Boer, Hanke, & He, 2018). These findings contradict the idea of a universal assessment and functioning of warmth and competence (Cuddy et al., 2008; Fiske, 2018; Fiske et al., 2007). As a consequence, a call has been put forward for the development of well-performing stereotype content scales, a careful examination of the measures' dimensionality, reliability and invariance, and the application of more suitable analytical approaches in Stereotype Content research (Blinded for peer review A, 2020). In this study, we aim to comply with this call.

Occupational Stereotypes

In this article, we will focus on the stereotype content of occupational groups. In accordance with He and colleagues, we define occupational stereotypes as “stereotypes about the specific professions or jobs that people hold, as well as the individuals who are employed in those occupations” (He, Kang, Tse, & Toh, 2019, p. 2). These stereotypes might apply to the actors who self-identify or are identified by others as practitioners of a specific occupation, to the actions that are required by the occupational role, and the structural and organisational systems upholding the occupation (Anteby, Chan, & Dibenigno, 2016). A rich pool of research exists to examine occupational stereotypes (e.g., Abele & Petzold, 1998; Philbin, 2016; Rutjens & Heine, 2016). Understanding these occupational stereotypes is important for a variety of reasons: For a start, occupation is a meaningful part of life, as it serves to provide financial means, self-concept and societal standing (Crößmann & Günther, 2018). Occupational stereotypes influence self- and other-perception, as well as impressions formed of particular occupations (Oswald, 2003); consequently, occupational group membership may be a relevant element of an individual’s social identity (Christiansen, 1999). Thus, occupational stereotypes may affect the people’s sense of self and well-being (e.g., van Vuuren, Teurlings, & Bohlmeijer, 2015). Moreover, stereotypes predict and shape emotional and behavioural reactions towards occupational groups (Cuddy et al., 2007), which might be applied for example to identify occupations with a high risk of experiencing misconduct and abuse in their professional activities. Finally, stereotypes about different occupational groups may affect career choices (e.g., by females systematically choosing or being chosen for occupations stereotyped as communal, social, and caring), or predict promotion and segregation in the workforce (i.e., the distribution of individuals from various demographic categories across different occupations; He et al., 2019; Ragins & Sundstrom, 1989).

Stereotypes in general and occupational stereotypes in particular can be assessed along a variety of dimensions, the principal and most frequently applied ones being gender associations and perceptions of status (e.g., Glick, Wilk, & Perreault, 1995; Miller & Hayward, 2006; Oswald, 2003).

Also, trait-based dimensions have been applied, such as hierarchy-enhancement vs. hierarchy-attenuation (e.g., Pratto, Stallworth, Sidanius, & Siers, 1997; Sidanius, Liu, Pratto, & Shaw, 1994) or theoretically derived vocational personality dimensions (e.g., Holland, 1985; Hollander & Parker, 1972). He and colleagues (2019) argued that these different dimensions show substantial communalities and proposed a comprehensive and parsimonious unified model of occupational stereotype dimensionality, which is based on warmth and competence, i.e., the Stereotype Content Model framework (Fiske et al., 2002). Warmth and competence correspond to two basic functions of behaviour (and therefore its assessment) which appear in various psychological research traditions and are highly relevant for the occupational context, namely, accomplishing tasks (i.e., competence) and forming bonds (i.e., warmth; Abele et al., 2020). Consequently, in line with He and colleagues (2019), we consider the SCM a suitable theoretical foundation for our research in occupational stereotypes.

Previous Research Findings

As the stereotype content of a particular group is dependent on the context of its assessment (e.g., Durante et al., 2013), we will focus primarily on German findings in the following, but we will also present relevant research from other cultural contexts. Generally, we wish to point out that stereotype content assessment is quite relational, i.e., the level of warmth and competence of one social group is determined by the relative difference in both dimensions compared to other social groups (rather than absolute values). As such, all findings for one particular occupational group are somewhat dependent on the other occupational or social groups it is compared with.

Research on occupational stereotypes applying the SCM is limited: Eckes (2002) and Asbrock (2010) included some occupational or occupation-related social groups in their assessments of SCM dimensions for societally relevant social groups in Germany. In both instances, the occupation-related social group *Pensioners* was evaluated as high in warmth and low in competence. Also, the occupation-related social group *Unemployed people* has been rated as low in competence (for reasons of brevity, in the following, we will refer to both *Pensioners* and *Unemployed people* as "occupational groups"). However, regarding warmth, the findings of *Unemployed people* ranged between relatively high (Eckes, 2002) and low ratings (Asbrock, 2010; The Fiske lab, n.d.). *Teachers* were perceived as high in both warmth and competence (Eckes, 2002). *Politicians* were evaluated as low in warmth, and either high (Eckes, 2002) or low (Imhoff et al., 2013; Wagner, Friehs, & Kotzur, 2020) in competence. *Physicians* were rated medium in warmth and high in competence (Asbrock, 2010; Imhoff et al., 2013). Also, *Child care workers*, *Hospital and elderly care nurses,* and *Teachers* on elementary school level showed high warmth and low to medium competence ratings (Imhoff et al., 2013). *Police officers* were rated as medium on both dimensions (Wagner et al., 2020).

Two recent surveys on German occupational stereotypes applied trait dimensions which are strongly associated with the SCM framework. The Gesellschaft für Konsum-, Markt- und Absatzforschung e.V. (GfK; 2018) assessed occupational trust for different professions, an established sub-facet of warmth (Cuddy et al., 2008; Fiske, 2018; Stanciu, 2015). The occupational groups *Craftspeople, Farmers, Firefighters, Hospital and elderly care nurses, Judges, Physicians, Police officers* and *Teachers* were rated high on trust(worthiness), while *Bankers* and *Politicians* were rated low. The survey institute forsa (2019) published a survey on occupational prestige in the German civil service. Prestige can be considered a sub-facet of status (Abele et al., 2020), which in turn strongly predicts competence perceptions in the SCM (Cuddy et al., 2008; Fiske, 2018; Fiske & North, 2014; Kervyn et al., 2015). Forsa (2019) found high prestige ratings for the occupational groups *Child care workers*, *Firefighters, Hospital and elderly care nurses, Judges, Physicians*, and *Police officers*, as well as medium prestige ratings for *Teachers* and low evaluations for *Bankers* and *Politicians*.

In the US-American context, He and colleagues (2019) assessed occupational stereotypes in a large sample using the SCM framework. Their findings indicate high warmth and high competence ratings for *Physicians* (referred to as "doctors") as well as high warmth and medium competence ratings for the occupational groups *Firefighters, Hospital and elderly care nurses* (referred to as "nurses"), and *Teachers*. *Child care workers* and *Farmers* were rated as high in warmth and low in competence, while *Politicians* were rated medium in warmth and low in competence. *Unemployed people* were rated low on both dimensions, *Police officers* and *Craftspeople* (referred to as "plumbers") were rated medium on both dimensions, and *Bankers* (referred to as "financial advisors") were rated low on warmth and high in competence.

The Present Research

We aim to advance the research on occupational stereotypes and the SCM in two ways: Methodically, by addressing methodological critiques of the SCM through the development of an adapted warmth and competence measure, the evaluation of the SCM's dimensionality, as well as the application of advanced analytical approaches for mean value comparison with the guarantee of measurement invariance as a precondition for the valid interpretation of results (Asparouhov & Muthén, 2014; Blinded for peer review A, 2020; Vandenberg & Lance, 2000); and content-wise, by presenting a current overview of occupational stereotypes using a large and heterogeneous sample and a parsimonious and comprehensive theoretical framework (Fiske et al., 2002; He el al., 2019). To this end, we conducted an online survey (N = 425) assessing participants' perceived warmth and competence assessments for 13 occupational groups using a new SCM scale adapted for the occupational context. After developing a well-performing stereotype content scale, we established the measurements' meaningful comparability between the different occupational groups (i.e., measurement invariance; Vandenberg & Lance, 2000) and compared their warmth and competence mean ratings on a latent level, thus correcting for measurement error (Asparouhov & Muthén, 2014). Our research was guided by the following expectations (E) based on previous SCM findings:

E1: Regarding the SCM scale development, we expected the used SCM items (or a sub-selection of these) to allow for the latent modelling of warmth and competence factors, which will be expressed in acceptable model fits in confirmatory factor analyses for all or most surveyed occupational groups (Blinded for peer review A, 2020; Blinded for peer review B, 2020).

E2: Regarding the assessment of occupational stereotypes, we expected to find significant variation on both warmth and competence dimensions between the surveyed occupational groups (Cuddy et al., 2008; Fiske et al., 2002).

E3: More in detail, we expected to find the following patterns of warmth and competence assessments for the various occupational groups:

a. For *Firefighters*, *Hospital and elderly care nurses*, *Judges* and *Physicians*, high evaluations of both warmth and competence (Asbrock, 2010; forsa, 2019; GfK, 2018; He et al., 2019; Imhoff et al., 2013);
b. For *Teachers*, high warmth and medium to high competence evaluations (Eckes, 2002; forsa, 2019; GfK, 2018; He et al., 2019; Imhoff et al., 2013);
c. For *Pensioners*, high warmth and low competence ratings (Asbrock, 2010; Eckes, 2002);
d. For *Craftspeople, Farmers* and *Child care workers*, high warmth ratings and unspecified competence ratings due to competing prior findings or lack of authoritative evidence (GfK, 2018; He et al., 2019; Imhoff et al., 2013);
e. For *Police officers*, medium ratings on both warmth and competence (He et al., 2019; Wagner et al., 2020);
f. For *Bankers* and *Politicians*, low warmth ratings and unspecified competence ratings due to conflicting prior findings (Eckes, 2002; forsa, 2019; GfK, 2018; He et al., 2019; Imhoff, 2013; Wagner et al., 2020);
g. For *Unemployed people*, low ratings on both warmth and competence (Asbrock, 2010; He et al., 2019).

Methods

This study received ethical approval from the ethics committee of the University Koblenz-Landau on March 25th, 2020 (reference number LEK-Kurzantrag 03-2020, 251). It was preregistered on April 27th, 2020 in the Open Science Framework (OSF;

https://osf.io/pmjgf/?view_only=32e35099f0b2450abe4e4c3c48d88632). Our data can be accessed openly in the OSF. Syntaxes and analyses outputs can be found in the Online Supplementary Materials (OSM). In accordance with the transparency statement by Simmons, Nelson and Simonsohn (2011), we report how we determined our sample size, all data exclusions, all manipulations, and all measures in this article and the corresponding OSM.

Identification of Relevant Occupational Groups

We aimed at nominating occupational groups that were familiar to all our survey participants from everyday discourse to foster the expression of ecologically valid stereotypes (Abele et al., 2020). Thus, we combined the two strategies of pilot test nominations and inclusion of groups based on their societal and political relevance (Fiske et al., 2002).

During a seminar session at a mid-size public German university, we conducted a pilot test with students pursuing a teaching Certificate (N = 45) who were asked to freely list *"Occupational groups that seem to be relevant in the German society"* (version A; n = 21) or *"Occupational groups of the civil service that seem to be relevant in the German society"* (version B, n = 24) without limitation of enumeration. We chose to include these two sets of instructions to broaden the scope of responses, but we pooled the answers for analysis. The nominations were summarised by merging different denominations for the same occupational groups (e.g., "physician", "medical practitioner" and "doctor" was summarised to *Physicians*). We then excluded nominations not referring to gainful occupation (e.g., "volunteers") and too general answers (e.g., "all occupations are relevant for a society to work well"). In total, 52 occupational groups were listed. We focused on 10 occupational groups which were mentioned by at least 20% of the complete sample (*Teachers*, *Police officers*, *Hospital and elderly care nurses*, *Physicians*, *Jurists*, *Social education workers*, *Public officials in the general administration*, *Firefighters*, *Politicians*, *Psychologists)*, as was done in previous SCM research (Asbrock, 2002; Fiske & North, 2014). From these, we removed one group due to its category broadness (*Public officials)* and two groups we assumed to be over-represented due to the context conditions of the pilot test (*Social education workers* and *Psychologists*; the data were generated in a University seminar session of students with purely educational background lead by a psychologist). Additionally, we specified the group *Jurists* as *Judges*.

These seven occupational groups were augmented by two occupation-related groups which were frequently used in previous research (*Pensioners*, *Unemployed people*; e.g., Asbrock, 2010; Eckes, 2002), and four groups that were in the focus of recent political and societal debates in Germany (*Child care workers*, *Craftspeople*, *Farmers*, *Bankers*). The final result included 13 occupational target groups (i.e., *Bankers*, *Child care workers*, *Craftspeople*, *Farmers*, *Firefighters*, *Hospital and elderly care nurses*, *Judges*, *Physicians*, *Police officers*, *Politicians*, *Teachers*), for which we used gender-neutral labels. A detailed documentation of the group selection process can be found in the online supplementary materials OSM A.

Measures

Stereotype Content. Previous research has found German and English SCM scales to perform poorly in confirmatory factor analyses (indicating that the dimensionality of warmth and competence was not given) and assessments of measurement invariance (indicating that (latent) mean value comparisons were biased; Blinded for peer review A, 2020; Blinded for peer review B; Kotzur et al., 2019, 2020). Consequently, we aimed at developing a new SCM scale with a special adaptation to assess warmth and competence of occupational groups. We screened all indicators that have previously been used in published SCM research and chose those traits we deemed most applicable in the context of assessing occupational stereotypes. Following the recommendations of Blinded for peer review A (2020), we included four indicators per dimension to increase the degrees of freedom of the latent models and allow for more potential adaptations of the measurement model. Each indicator was presented as two extremes of a semantic differential with five gradations: Warmth was assessed using the German equivalents for the items "*dishonest - honest", "unfriendly - friendly", "ill-natured - good-natured", "cold - warm",* while competence was rated with *"careless - thorough", "incompetent - competent", "lazy – hard-working",* and *"inefficient - efficient"*[1]. All items were coded so that higher values indicate higher warmth/competence ratings. In tradition with previous SCM research, we assessed warmth and competence from a societal perspective (Fiske et al., 2002), using an adapted version of the German SCM instructions by Asbrock (2010): *"From the perspective of most Germans, how [item] are the following occupational and social groups perceived?"*. In accordance with Abele et al. (2020) and Fiske et al. (2002), we assumed that by asking about societal beliefs, rather than personal assessments, possible social desirability bias would be minimised.

Other measures. Furthermore, demographic variables including gender, age, education, nationality, federal state of residence, migration background, occupational status and identification with any of the rated occupational groups were collected. If participants identified with any of the occupational groups, we excluded the ratings for said occupational group using pairwise deletion to avoid in-group preference (between 0.5 and 13.9% of the final sample, multiple answers possible, see Table 1 for further details). For further research purposes which were not subject of this study, we included two items on the gender-stereotypical perception of the occupational groups. These were assessed after the stereotype content items. A complete copy of the survey in German and English is presented in OSM B.1 and OSM B.2, respectively.

Procedure

Data were collected between April 28th and May 30th, 2020, using an online survey which took on average about 14 minutes to complete. The survey first assessed demographic variables and self-identification, followed by the SCM measures and other measures. We followed the procedure of previous SCM studies (e.g., Cuddy et al., 2009; Eckes, 2002; Kotzur et al., 2019) by presenting one indicator per survey page with all occupational groups randomly listed underneath the instruction and alternating warmth and competence indicators between pages (for the indicator order, please see OSM B).

To avoid fatigue effects due to survey length, we applied a planned-missing-data design (three-form design; Graham, 2009; Graham, Taylor, Olchowski, & Cumsille, 2006). Thus, participants were randomly allocated to three different conditions, with each one missing one subset of occupational groups respectively. All participants evaluated 10 occupational groups, all of them including the four groups *Unemployed people, Pensioners, Physicians*, and *Bankers*. For the remaining groups, the excluded subsets per condition were: Condition 1 (*n* = 143) - *Craftspeople*, *Firefighters, Police officers*; Condition 2 (*n* = 140) - *Hospital and elderly care nurses*, *Child care workers*, *Judges*; Condition 3 (*n* = 142) - *Teachers*, *Politicians*, *Farmers*.

Sample

We recruited a heterogeneous German sample with regard to age and occupation via websites, institutional e-mail lists, social media platforms, real-life advertisement (e.g., flyers, blackboards), and personal social networks (snowball procedure). Participants could take part in the survey if they were at least 18 years old and spoke German on native speaker level. All participants gave their informed consent to participate in the survey. As an incentive, after completing the survey, we offered participants to choose one out of four charitable organisations to which we donated 0.50€ per participant.

In total, 552 people started the online survey. After excluding those that had missing values on all relevant items (i.e., all SCM items; *n* = 127; 23.01% of overall participation), our final sample contained *N* = 425 participants. We aimed for a minimum sample size based on the minimal requirements for the alignment optimisation procedure (*n* > 100 per evaluated occupational group; Asparouhov & Muthén, 2014) while considering the planned-missing-data design. The final sample size (66.8% female, 0.5% diverse, 1.2 % missing; M_{Age} = 36.11 years, SD_{Age} = 13.97, $\text{Range}_{\text{Age}}$ = 18 - 83; 9.3% max. 10 years of school education, 33.4% university entrance diploma, 56.5% at least university graduate degree; 30.1% school or university students, 61.6% gainfully employed) surpassed these minimum requirements for all occupational groups. Further descriptive information is shown in Table 1 and in OSM C.

- Table 1 about here -

Analytical Strategy

To address the two different research aims of the article at hand, we applied two separate analytical procedures.

SCM scale development. In a first step, we aimed to develop a well-performing SCM scale using exploratory (EFA)[2] and confirmatory factor analytical (CFA) procedures (Brown, 2015). We aimed at establishing a common measurement model with acceptable model fit parameters for all occupational groups, or, if that failed, to maximise the number of occupational groups with acceptable model fit. Though we had a firm theoretical base concerning the scales' dimensionality and the interrelations of the different items (e.g., Fiske & North, 2014), we chose to start with an exploratory approach, because to our knowledge, neither the described combination of items nor the semantic differentials have previously been used in German SCM research. We randomly split our sample into one EFA sample (n_1 = 150), and one CFA sample (n_2 = 275). We used the smaller EFA sample to explore the scales dimensionality by running one- to four-factor EFAs with all eight items using oblimin[3] rotation for each of the 13 occupational groups. We focused on the number of eigenvalues > 1, the model fit and its improvement[4] when introducing additional factors, and the interpretability of the items' loading patterns to determine an optimal measurement model. In a second step, the most promising measurement model was applied to the larger CFA sample. CFA model fit was acceptable if the criteria of Schermelleh-Engel, Moosbrugger, and Müller (2003) were met: Root mean standard error of approximation (RMSEA) < .08; standardised root mean square residual (SRMR) < .10; comparative fit index (CFI) > .95. As we aimed at achieving acceptable model fit for a maximum number of occupational groups, we reserved the right to perform additional adjustments of the measurement model in the CFA context to increase the number of occupational groups with acceptable model fit.

Latent mean comparison. Inspired by Kotzur and colleagues (2019), we applied the alignment optimisation procedure (Asparouhov & Muthén, 2014) to the entire sample (*N* = 425) to compare the social perception of the different occupational groups. Alignment optimisation is a procedure to compare reliability-corrected latent mean value differences, thus accounting for measurement error (Kline, 2010). This is a distinct advantage to many frequently used observed analytical approaches in SCM research (e.g., k-means cluster analysis, Asbrock, 2010; Eckes; 2002; Fiske et al., 2002). Compared to multiple-group confirmatory factor analysis (MGCFA), which is the traditional approach to compare latent means, alignment optimisation does not require the manual implementation of metric (i.e., equality assumption of factor loadings of identical indicators across identical measurement models of occupational groups to guarantee equal scaling; Boer et al., 2018) and scalar measurement invariance (i.e., additional equality assumption of indicator intercepts of identical indicators across occupational groups to ensure equal points of zero; Boer et al., 2018). Metric and scalar invariance are preconditions for valid, meaningful and free-of-bias latent mean value comparison (Davidov, Meulemann, Cieciuch, Schmidt, & Billiet, 2014). The alignment optimisation procedure automatically discovers the most optimal measurement invariance pattern based on a simplicity function similar to rotations in EFA, which minimises the number of non-invariant parameters (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2018). What is more, alignment optimisation is suitable for the within-person comparison of latent mean values (Asparouhov, 2020) and thus can be applied to the data at hand.

Following the procedure outlined by Asparouhov and Muthén (2014), we first assessed the general baseline measurement model fit for each occupational group using CFA in the entire sample. Subsequently, we performed a configural measurement invariance test (i.e., equal number of factors, and equal loading pattern of indicators onto factors; Boer et al., 2018) with all occupational groups presenting acceptable baseline model fit. In case of acceptable configural model fit, the occupational groups were entered into a fixed alignment optimisation analysis. The fixed alignment option sets the factor mean values of one occupational group to zero and the factor variances to one. All other latent mean values and variances are estimated freely. We chose *Unemployed people* as the fixed group, as previous research theorised this group to be rated very lowly on both SCM dimensions (Fiske, 2018), and consequently, the mean values of all other occupational groups should be scaled positively. Results were proposed to be robust and trustworthy if the share of non-invariant parameters in the alignment optimisation model did not exceed 25% (Asparouhov & Muthén, 2014).

Results

Data restructuring and descriptive analysis was carried out using IBM SPSS Statistics 25 (IBM Corporation, 2017). For EFA, CFA and alignment optimisation, we used Mplus version 8.3 with a robust maximum likelihood estimator (MLR) to account for missing values[5], non-normality and non-independence in the data (Muthén & Muthén, 1998-2017). The item-level descriptive statistics of the SCM indicators are displayed in Table 2, the item-intercorrelations per occupational group are displayed in OSM D.

- Table 2 about here -

SCM Scale Development

Exploratory factor analysis. The eigenvalues of the models in the EFA sample varied between one and three. In seven instances, we found a two-factor solution to be optimal for representing the empirical data structure; in the other cases, a one-factor-solution (one case), a three-factor solution (three cases) or a four-factor solution (two cases) was preferred on the base of χ^2-difference tests. Examining two-factor EFA solutions, we found that the warmth indicator *"dishonest-honest"* generally performed poorly, as it showed mis-specified factor loadings (i.e., significant loadings on both factors, non-significant loadings on any factor, or significant loadings only on the competence factor) in 10 out of 13 cases. Consequently, we decided to remove this indicator from further analysis. Detailed results of the EFA per occupational group, including the analysis syntaxes and complete outputs, can be found in OSM E.

Confirmatory factor analysis. Using the CFA sample, we modelled a CFA measurement model for each occupational group with the remaining three warmth indicators loading exclusively on a latent warmth factor, and the four competence indicators loading solely on a latent competence factor. Like in the EFA, we allowed for a latent correlation between the warmth and competence factors (for the reasoning, see footnote 2) and did not specify any residual covariations.

Nine out of 13 measurement models showed acceptable model fit. The model fit information per occupational groups, as well as detailed CFA solutions including syntaxes and complete outputs, can be found in OSM F.

To further improve model fit, we examined the standardised residual covariances and modification indices, which would indicate further potential empirically-driven model adaptations, of all models with a special focus on those occupational groups with non-acceptable model fit (see OSM G for detailed information). These information sources proposed two options: Either a cross-loading of the competence indicator *"thorough – careless"* on the warmth factor, which we discarded on theoretical grounds because warmth and competence are proposed as conceptually separate dimensions of social perception; or a residual covariation between the competence indicators *"hard-working - lazy"* and *"efficient - inefficient"*. We allowed for the latter residual covariation between the two competence indicators, which is in accordance with Stanciu (2015) proposing a distinct efficacy sub-facet of competence. Subsequently, we re-ran the CFA with the adapted measurement model (see Figure 1). Twelve out of 13 occupational groups (all groups except *Pensioners*, for which the CFI was slightly too low) showed acceptable model fit in the adapted version (for the detailed CFA solutions including syntaxes and complete output, see OSM H). This confirmed our expected scale performance expressed in E1. Therefore, we decided to apply this measurement model in the alignment optimisation procedure.

- Figure 1 about here -

Latent Mean Value Comparison

Baseline model fit and configural measurement invariance. We conducted full-sample CFA using the adapted measurement model depicted in Figure 1. The model fit is displayed in Table 3 (for detailed CFA syntaxes and outputs, see OSM I). All 13 occupational groups showed acceptable model fit, again supporting E1. Average scale reliability was ω_{Warmth} = 0.847 (range: 0.788 - 0.895) and $\omega_{\text{Competence}}$ = 0.850 (range: 0.805 - 0.932), which can be considered adequate (Raykov & Marcoulides, 2011). A configural measurement invariance test of all occupational groups also showed acceptable model fit, $\chi^2(156) = 234.091$, $p = 0.0001$, RMSEA = 0.041 [90% CI 0.030 - 0.051], CFI = 0.984, SRMR = 0.036 (for detailed results, including syntax and complete output, see OSM J). This indicated that the data were eligible for entering into the alignment optimisation procedure.

- Table 3 about here -

Alignment optimisation. We used the fixed alignment optimisation option, which means that we defined the latent warmth and competence means of the occupational group *Unemployed people* to be zero and the factor variances to be one. The resulting measurement invariance model showed non-invariance for one out of 91 factor loadings and for nine out of 91 indicator intercepts[6], resulting in a partial metric and scalar measurement invariance model with 5.49% non-invariant parameters. This indicated a trustworthy and robust estimation of latent means and variances in the alignment optimisation model (Asparouhov & Muthén, 2014), which can be interpreted without reservations.

The latent mean values of the different occupational groups are outlined in Table 4 for warmth and Table 5 for competence. All scores are graphically depicted in Figure 2. Further information on the alignment optimisation model (including syntax and compete output with factor loadings, indicator intercepts, factor means and variances, and factor covariations of warmth and competence) are provided in OSM K.

- Table 4 about here -

- Table 5 about here –

- Figure 2 about here -

We found significant latent mean differences between the occupational groups on both dimensions, supporting E2. The stereotype content ratings of the different occupational groups were as follows:

In line with E3a, *Firefighters* showed high warmth and competence scores, indeed the highest ratings on both dimensions. *Hospital and elderly care nurses* presented high warmth and medium competence ratings, which is partly consistent with our predictions in E3a. *Physicians* showed high competence ratings, but warmth ratings were lower than expected in E3a with medium to high ratings. *Judges* also showed medium to low warmth and high competence ratings, which deviates from E3a. For *Teachers*, we found medium warmth and medium to high competence scores, thus partly supporting E3b. We also found medium warmth and competence ratings for *Pensioners*, contradicting the expectation in E3c. *Child care workers* showed high warmth ratings, supporting E3d, and medium to high competence scores. *Farmers* and *Craftspeople* actually both scored in the medium range of warmth, against the expectation formulated in E3d. Also*, Farmers* presented medium to high competence scores, while *Craftspeople* received medium competence ratings. Consistent with E3e, *Police officers* showed medium ratings on both warmth and competence. The low warmth ratings of *Bankers* and *Politicians* proposed in E3f were confirmed; indeed, *Politicians*' warmth score was the lowest of all occupational groups. *Bankers* scored medium on the competence dimension, while *Politicians* scored low on competence. *For Unemployed people*, we found low warmth and low competence perceptions, as expected in E3g.

As an additional fact worth reporting, we found high and significant positive correlations between the warmth and competence factors, both within the individual measurement models of the occupational groups, r = 0.363 - 0.819, $p \leq 0.001$ (see OSM I for further details), and overall between the occupational groups, r = 0.594, p = 0.032.

Discussion

In this study, we pursued the two goals of developing a scale to measure perceived warmth and competence, the two fundamental dimensions of social perception as defined by the Stereotype Content Model (SCM; Fiske et al., 2002), and employing it to describe current occupational stereotypes in Germany. Using an online survey in a heterogeneous adult sample and applying the state-of-the-art alignment optimisation procedure to compare latent warmth and competence means, we found substantial differences between the perception of the 13 occupational groups included in the survey. The results as well as their implications will be discussed in the following.

Development of a Stereotype Content Scale to Assess Occupational Stereotypes

One goal of this article was to develop and apply a reliable and valid scale to assess stereotype content of occupational groups. Previous SCM scale development efforts did not fulfil our need for a scale to adequately assess human targets, compared to products or countries (Halkias & Diamantopoulos, 2020). What is more, the functionality of established German and English scales was challenged due to its unclear dimensionality and because preconditions for (latent) mean value comparison (i.e., (partial) scalar measurement invariance) were often not given (Blinded for peer review A, 2020; Blinded for peer review B, 2020; Kotzur et al., 2019, 2020). Thus, we carefully selected indicators suitable for the context of assessing occupational stereotypes, increased the number of indicators per scale and the overall sample size, and ran through a comprehensive factor-analytical scale development procedure. As a result, we can present a scale with a well-defined dimensionality, good model fit for all occupational groups we assessed, and with high reliability. We hope that this scale will help produce more valid and reliable SCM findings and will provide options for meta-analytical research on stereotype content using the same scale in the future.

Nonetheless, we believe more applications and careful examinations of the SCM scale are needed before using this scale without reservations. For one, we specifically tailored the scale to fit to the context of assessing occupational stereotypes. Most (German) SCM research, however, was not conducted to assess occupational stereotypes, but the perceptions of other target groups (e.g., social groups defined by gender, origin, or other features, or experimental conditions). Without further research applying the scale to non-occupational social groups, we cannot, at this point, attest to the general applicability of the scale to other contexts. For all further applications, we call for a careful examination of the measurement models using confirmatory factor analysis. Also, given that we used the alignment optimisation procedure, we cannot currently attest that the scales fulfil the precondition for meaningful latent mean value comparisons (i.e., (partial) scalar measurement invariance; Davidov et al., 2014) when tested with more traditional or conservative testing procedures, like multiple-group confirmatory factor analysis (MGCFA; Blinded for peer review A, 2020; Blinded for peer review B, 2020, Kotzur et al., 2020). Indeed, the data presented in Kotzur and colleagues (2019) showed considerably diverging results depending on the method of analysis (alignment optimisation in Kotzur et al., 2019; MGCFA in Blinded for peer review B, 2020). At the moment, we cannot exclude this dependency on analytical methods for our scale. Lastly, we would like to point out that allowing for a residual covariation between the two competence indicators, as we did in the study at hand, is new to (factor-analysis based) SCM research. Though it is in line with theoretical considerations about sub-dimensions of warmth and competence (Stanciu, 2015), it somewhat hinders applications of the scale to analyse observed means, because only modelling approaches can account for this residual covariation.

Occupational Stereotypes in Germany

Another goal was to describe occupational stereotypes in Germany using the well-established SCM. Previous research applying the SCM or related constructs provided ample empirical evidence to predict occupational stereotypes of some groups (e.g., Asbrock, 2010; Eckes, 2002; forsa, 2019; GfK, 2018; He et al., 2019; Imhoff et al., 2013; Wagner et al., 2020). As such, we were able to confirm our assumptions concerning both the warmth and competence assessments of *Firefighters* and *Police officers*, the warmth prediction of *Bankers*, *Child care workers* and *Politicians*, as well as the competence expectations concerning *Unemployed people*, *Physicians* and *Teachers*. Nonetheless, some of our hypotheses were contradicted outright, such as the high warmth rating of *Teachers* or the high competence perceptions of *Politicians*, or deviated slightly from our expectations, for example in the case of *Judges, Farmers* and *Craftspeople*. While an extensive discussion of all findings is beyond the scope of this article, we would like to draw the attention to some select and surprising findings.

Based on the results reported in the literature, we expected both *Physicians* (Asbrock, 2010; GfK, 2018; He et al., 2019; Imhoff et al., 2013) and *Teachers* (Eckes, 2002; GfK, 2018; He et al., 2019; Imhoff et al., 2013) to be perceived as highly warm. What is more, we would have assumed contextual circumstances (i.e., the COVID-19 pandemic circulating globally during the time of data collection; see below) to reinforce this positive warmth assessment due to an increased public salience and appreciation of these occupational groups' societal contributions. However, we found both occupational groups to score medium on warmth, with at least two occupational groups showing significantly higher warmth ratings. Our data cannot provide explanatory information for this deviation from theory; nonetheless, these issues might be worth investigating in future research.

We would also like to point out the prominently negative occupational stereotypes of *Politicians*, which were rated lowest on warmth and second-lowest on competence. These findings are not new (e.g., forsa, 2019; GfK, 2018; Wagner et al., 2020) and consistent with results focusing on other information sources, such as the screening of occupational groups mentioned frequently and negatively on the Internet (GfK, 2018). Nonetheless, they give rise to substantial societal concerns: Cuddy and colleagues (2007) proposed that warmth and competence stereotypes are predictive of emotional and behavioural responses. Consequently, the negative occupational stereotypes of politicians might in part be responsible for current political issues, such as the rise of right-wing populist parties, which proclaim their difference from established politicians and vote for fundamental changes in the political system, or the recent reports of hate mail threatening the lives of various politicians. In the long run, these negative perceptions of politicians might impair the functioning of the democratic system through a loss of interest and support for political parties and initiatives, reduced voter participations, and support for non-democratic movements and ambitions.

Finally, the SCM predicts warmth and competence dimensions to be independent and frequent observations of ambivalent stereotypes (i.e., high ratings on one dimension paired with low ratings on the other; Abele et al., 2020; Cuddy et al., 2009; Durante et al., 2013, 2017; Fiske, 2015; Fiske et al., 2002). In contrast, our findings showed overall strong and significant positive correlations between warmth and competence (also indicated graphically by the tendency of the occupational groups to cluster along the diagonal from low warmth-low competence to high warmth-high competence), both within and between occupational groups. There was also a distinct absence of ambivalently rated occupational occupational groups. Importantly, this pattern is not indicative of uni-dimensionality of the applied stereotype indicators, as the EFA results indicated an (at least) bi-dimensional solution to be preferred in all occupational groups except one. Obviously, from a statistical perspective, the between-group relation between the warmth and competence dimensions depends highly on the selection of occupational groups. Thus, our finding might just be explained by a tendency to select non-ambivalently stereotyped occupational groups for assessment. However, comparable findings have been reported elsewhere (Durante et al., 2013, 2017; Kervyn et al., 2015). One explanation might be by the fact that for both warmth and competence, it is assumed desirable to be rated highly, and therefore these dimensions correlate positively with general evaluations (Kervyn, Fiske, & Yzerbyt, 2013; Osgood, Suci, & Tannenbaum, 1957). In fact, Sayans-Jímenez, Cuadrado, Rojas and Barrada (2017) found support of a bi-factor-model of Stereotype Content featuring both the SCM dimensions and an independent global evaluation factor. On the other hand, high correlations between warmth and competence factors within and across occupational groups could be indicative of acquiescence of halo-effects (Judd, James-Hawkins, Yzerbyt, & Kashima, 2005). The lack of ambivalently evaluated groups is also in line with the findings of Durante and colleagues (2013, 2017), which predict little ambivalent stereotypes in societies with low inequality and low conflict, such as Germany.

Relevance of the Research Results

Our findings might be applied in the investigation of social interactions and processes in specific work contexts: Oftentimes, work places are characterised by the intimate collaboration of differently stereotyped occupational groups (e.g., nurses and physicians in hospitals, teachers and child care workers in schools). Employees holding low-status positions may be stereotyped as incompetent and, therefore, passed over, so that information exchange, and thus collaboration, is disturbed (Abele et al., 2020). Indeed, research found that if employees occupying high-status positions in the hospital hierarchy would consider the input of employees in lower positions, medical errors could be reduced drastically (Sutcliffe, Lewton, & Rosenthal, 2004). Reversely, employees in high-status positions being stereotyped as highly competent might not be informed about smaller issues and problems due to strong perceptions of hierarchy, which might lead to “blind spots” and impaired decision-making processes based on incomplete information (Tourish, 2005). Acknowledging the ways different staff members could potentially be biased by social perception processes is crucial for well-functioning team work, which is a necessity in most contemporary working environments.

As mentioned before, occupational stereotypes might also strengthen occupational segregation (i.e., the distribution of individuals from different demographic backgrounds across occupations; He et al., 2019). Groups such as women, physically or mentally impaired people, or those with low socio-economic status might be underrepresented in occupations scoring high on competence (He et al., 2019), a circumstance by which occupational stereotyping is reinforced. Occupational segregation can be reduced by predicting the social groups that might be underrepresented in a particular job and subsequently encouraging and promoting their access to and performance in that occupation (e.g., by phrasing job advertisements non-discriminatory). By knowing about occupational stereotypes and intervening accordingly, future labour shortages might be prevented (He et al., 2019). Thus, our research might also be applied to define and examine strategies to change occupational stereotypes (He et al., 2019).

Limitations and Future Directions

Naturally, we observe some issues that might limit the interpretability and generalisability of our findings, both content-wise and methodically. In the following, we wish to discuss these aspects and offer orientation for future research.

Methodically, our work can be criticised due to the fact that we did not use an independent sample to develop our scale before testing for mean differences, as is often proposed in scale development literature (e.g., Brown, 2015). We believe that our approach is justifiable both because establishing well-fitting baseline measurement models is an inherent part of the alignment optimisation procedure (Asparouhov & Muthén, 2014), meaning that we would have needed to determine an adequate measurement model in any case, and because the scale we used was adapted especially for the context of occupational stereotypes. Nonetheless, further applications of the proposed scale using non-occupational social groups is desirable.

Content-wise, as discussed above, we acknowledge that our findings are relational and dependent on the specific other occupational groups we assessed. Other research comparing different occupational groups might thus come to somewhat different conclusions. What is more, our study contained only a small number of occupational groups (compared to other SCM research, e.g., Asbrock, 2010; Eckes, 2002; He et al., 2019), thus limiting the descriptive and comparative informational value. Most certainly, the number and choice of groups in our study does not reflect the full range of occupational groups relevant in any society. Nonetheless, we collected data from a heterogeneous sample, most of whom had no substantial prior experience with filling in online surveys. Thus, we needed to keep survey length and participant strain to a minimum (Halkias & Diamantopoulos, 2020). Further research might investigate the stereotypes associated with more or other occupational groups. When assessing many different groups in one survey, we recommend allocating subsamples to a selection of target groups to avoid fatigue (e.g., He et al., 2019).

We would also like to draw the attention to the potential influence of the context this study was conducted in, an this may influence the occupational stereotypes of some groups. During the data collection period, Germany just experienced a relaxation of severe restrictions of everyday life and personal freedom due to the COVID-19 pandemic as well as a decreasing number of severe medical treatments for the lung disease. This context might impact the evaluations of some occupational groups, such as *Physicians* and *Hospital and elderly care nurses*. Likewise, schools and nurseries were closed for the most part, and parents were forced to care for their children at home, which might affect the assessments of *Teachers* and *Child care workers*. This period was also marked by a large number of short-term and extensive political decisions, mainly to stabilise Germany's economy and to provide more extended health-care, potentially impacting the stereotype content of *Politicians*. Our data collection period also overlapped somewhat with the lamentable incident of George Floyd's death in the US on May 25th, 2020, which initiated a wave of protests and a fierce public debate about racism in the police force both in the U.S. and in Germany. Consequently, the social perception of *Police officers* might be influenced by these circumstances. Previous SCM research has not, to the best of our knowledge, focused on the impact of relevant external circumstances, nor on the change of occupational stereotypes over time. Thus, further research applying repeated cross-sectional or longitudinal surveys might help answer these questions.

Finally, future research could apply the assumption that warmth and competence perceptions are predictive of emotional and behavioural reactions towards the assessed occupational groups (Cuddy et al., 2007). Thus, on the base of the presented findings, future research could predict and investigate the affective and conative responses certain occupational group memberships might elicit in professional interactions or societal discourses. This approach might be employed on a variety of contemporary problems, such as the striking contrast between the highly positive social perceptions of some occupational groups (e.g., professions in the child, hospital or elderly care sector) on the one hand, and their precarious working conditions and insufficient remuneration on the other hand (DGB Niedersachsen, 2020). Another application might lie in the investigation of reported phenomena of actively harming or hindering representatives of different occupational groups fulfilling their occupational role (e.g., attacking firefighters and paramedics in action). The SCM and related theories may be put to the test as a theoretical framework to describe and explain these phenomena.

Conclusion

In this article, we assessed occupational stereotypes in Germany applying the Stereotype Content Model of Fiske and colleagues (2002). Based on exploratory and confirmatory factor analyses, we developed a well-functioning, reliable and valid scale to assess warmth and competence of 13 different occupational groups. We compared their occupational stereotype content on a latent level using the alignment optimisation procedure proposed by Asparouhov and Muthén (2014). We found occupational stereotypes to differ significantly on both dimensions, ranging from the high warmth-high competence extreme (e.g., *Firefighters*) to the low warmth-low competence extreme (e.g., *Unemployed people*, *Politicians*) with little ambivalent stereotyping. We advocate for the exploration of occupational stereotypes as an important field of research, as they might shape the experiences of people working in or interacting with these occupations, as well as impact career decisions.

Notes

[1] Recently, Halkias and Diamantopoulos (2020) presented a thoroughly developed SCM measure for the usage in marketing research. Our item selection shows some overlap with this scale, but we focused specifically on building a scale to optimally describe the actions of professional individuals in an occupational context, rather than describing characteristics of a product or country of origin, which was the purpose of Halkias and Diamantopoulos (2020).

[2] Please note that the EFA were not part of the original preregistration. Nonetheless, during our data analysis, we deemed conducting EFAs an important analytical step of the scale development. If not reported otherwise, all other analyses were conducted as presented in the pre-registration.

[3] Warmth and competence are theorised to be independent (i.e., the factor correlation should be zero; Fiske, 2015; Fiske et al., 2007). Nonetheless, empirically, the scales often correlate quite highly (e.g., Durante et al., 2013; Kervyn et al., 2015; Kotzur et al., 2019), so we allowed for an oblique rotation (i.e., for the factors to correlate with each other).

[4] Mplus reports model fit criteria for EFAs (i.e., χ^2, RMSEA, CFI, SRMR), which we used to evaluate model fit.

[5] In the preregistration, we indicated that we might explore multiple imputation for the estimation of missing values, depending on the compatibility of multiple imputation with the statistical analyses. We refrained from applying multiple imputation because conducting exploratory factor analyses as well as examining standardised residual covariances and modification indices in CFA are not supported when using multiple imputation.

[6] The non-invariant parameters were: The factor loading of *"careless - thorough"* for *Child care workers*, the indicator intercepts of *"unfriendly - friendly"* for *Bankers* and *Politicians*, of *"cold - warm"* for *Unemployed people* and *Pensioners*, of *"careless - thorough"* for *Child care workers*, of *"lazy - hard-working"* for *Unemployed people* and *Farmers*, and of *"inefficient - efficient"* for *Unemployed people* and *Pensioners.*

Data Accessibility Statement

All data presented in this article are openly accessible on the website of the article's Open Science Framework project, see https://osf.io/gxz49/?view_only=5f6a34d009ae4741ad1dd4142ffd8d7d.

Acknowledgements

The authors have no competing interests to declare.

The authors contributed to the production of this manuscript in the following ways:

Maria-Therese Friehs – Conceptualisation and Design, Methodology, Data curation, Formal analysis, Interpretation of results, Writing – Original draft;

Felicia Aparicio Lukassowitz – Conceptualisation and Design, Data curation, Interpretation of results, Writing – Critical review and editing.

Ulrich Wagner – Conceptualisation and Design, Interpretation of results, Writing – Critical review and editing.

We thank Patrick Kotzur, Julia Karbach, and Thilo Friehs for their advice and comments on earlier drafts of this article.

We thank the *Fachgruppe Sozialpsychologie* of the *Deutsche Gesellschaft für Psychologie* for providing a short-term research grant which financed the participants’ compensation in this study.

References

Abele, A. E., Cuddy, A. J. C., Judd, C. M., & Yzerbyt, V. Y. (2008). Fundamental dimensions of social judgment. *European Journal of Social Psychol*ogy, *38*(7), 1063–1065. https://doi.org/10.1002/ejsp.574

Abele, A., E., Ellemers, N., Fiske, S. T., Koch, A., & Yzerbyt, V. (2020). *Navigating the social world: Shared horizontal and vertical dimensions for evaluating self, individuals, and groups.* PsyArXiv. https://doi.org/10.31234/osf.io/b5nq6

Abele, A. E., & Petzold, P. (1998). Pragmatic use of categorical information in impression formation. *Journal of Personality and Social Psychology, 75*(2), 347–358. https://doi.org/10.1037/0022-3514.75.2.347

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. https://doi.org/10.1037/0022-3514.93.5.751

Allport, G. W. (1954). *The nature of prejudice.* Reading, MA: Addison-Wesley.

Anteby, M., Chan, C. K., & DiBenigno, J. (2016b). Three lenses on occupations and professions in organizations: Becoming, doing, and relating. *The Academy of Management Annals, 10*(1), 183–244. https://doi.org/10.1080/19416520.2016.1120962.

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

Asparouhov, T. (2020, July 07). Re: Multigroup alignment method [Online forum comment]. Retrieved from http://www.statmodel.com/discussion/messages/9/13900.html?1543970865

Asparouhov, T., & Muthén, B. O. (2014). Multiple-group factor analysis alignment. *Structural Equation Modelling, 21*, 1-14. https://doi.org/10.1080/10705511.2014.919210

Blinded for peer review A. (2020). The examination of the structural validity of English stereotype content measures – Problems and possible solutions. *Manuscript in preparation.*

Blinded for peer review B. (2020). Equal scales for everyone? – A preregistered examination of the structural validity of German stereotype content model data. *Manuscript in preparation*.

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: 10.1177/0022022117749042

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: Guilford Press.

Christiansen, C. H. (1999). Defining lives: Occupation as identity: An essay on competence, coherence, and the creation of meaning. American Journal of Occupational Therapy, *53*(6), 547–558. https://doi.org/10.5014/ajot.53.6.547

Crößmann, A., & Günther, L. (2018). Arbeitsmarkt. In *Datenreport 2018. Ein Sozialbericht für die Bundesrepublik Deutschland* (pp. 149–165). Bundeszentrale für politische Bildung. Retrieved from https://www.bpb.de/system/files/dokument_pdf/dr2018_bf_pdf_ganzes_buch_online.pdf

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology, 92*(4), 631-648. https://doi.org/10.1037/0022-3514.92.4.631

Cuddy, A. j. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology, 40,* 61-149. https://doi.org/10.1016/S0065-2601(07)00002-0

Cuddy, A. J. C., Fiske, S. T., Kwan, V. S. Y., Glick, P., Demoulin, S., Lexens, ... Ziegler, R. (2009). Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology, 48*(1), 1-33. https://doi.org/10.1348/014466608X314935

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

DGB Niedersachsen (2020, April 02). Systemrelevante Berufe: Kostenloser Applaus reicht nicht [System-relevant occupations: Applause is not enough]. Retrieved July 02, 2020 from https://niedersachsen.dgb.de/themen/++co++76b18518-74b6-11ea-8b82-52540088cada

Dovidio, J. F., Hewstone, M., Glick, P., & Esses, V. M. (2010). Prejudice, stereotyping and discrimination: Theoretical and empirical overview. In J. F. Dovidio, M. Hewstone, P. Glick, & V. M. Esses (Eds.), *The SAGE Handbook of Prejudice, Stereotyping and Discrimination* (pp. 3-52). London: SAGE Publications Ltd. https://doi.org/10.4135/9781446200919.nl

Durante, F., Fiske, S. T., Kervyn, N., Cuddy, A. J., Akande, A. D., Adetoun, B. E., ... Barlow, F. K. (2013). Nations' income inequality predicts ambivalence in stereotype content: How societies mind the gap. *British Journal of Social Psychology, 52*(4), 726-746. https://doi.org/10.1111/bjso.12005

Durante, F., Fiske, S. T., Gelfand, M. J., Crippa, F., Suttora, C., Stillwell, A., … 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. https://doi.org/10.1073/pnas.1611874114

Eckes, T. (2002). Paternalistic and envious gender stereotypes: Testing predictions from the stereotype content model. *Sex Roles*, *47*(3-4), 99-114. https://doi.org/10.1023/A:1021020920715

Fiske, S. T. (2015). Intergroup biases: A focus on stereotype content. *Current Opinion in Behavioral Sciences, 3*, 45-50. https://doi.org/10.1016/j.cobeha.2015.01.010

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science, 27*(2), 67-73. https://doi.org/10.1177/0963721417738825

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. 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. https://doi.org/10.1037//0022-3514.82.6.878

Fiske, S. T., & North, M. S. (2014). Social psychological measures of stereotyping and prejudice. In J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of Personality and Social Psychological Constructs.* Oxford, UK: Academic Press.

forsa Politik- und Sozialforschung GmbH. (2019). *dbb Bürgerbefragung Öffentlicher Dienst. Einschätzungen, Erfahrungen und Erwartungen der Bürger* [dbb citizen survey on the public service. Assessments, experiences and expectations of the citizens]. Retrieved July 07, 2020 from https://www.dbb.de/fileadmin/pdfs/2019/forsa_2019.pdf

Gesellschaft für Konsum-, Markt- und Absatzforschung e.V. [GfK]. (2018). *Trust in Professions 2018—Eine Studie des GfK Vereins*. [Trust in professions 2018 – A study of the GfK society]. Retrieved July 07, 2020 from https://www.nim.org/sites/default/files/medien/135/dokumente/2018_-_trust_in_professions_-_deutsch.pdf

Glick, P., Wilk, K., & Perreault, M. (1995). Images of occupations: Components of gender and status in occupational stereotypes. *Sex Roles, 32*(9–10), 565–582. https://doi.org/10.1007/BF01544212

Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, *60*(1), 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323–343. https://doi.org/10.1037/1082-989X.11.4.323

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*, advance online publication. https://doi.org/10.1016/j.ijresmar.2020.02.004

He, J. C., Kang, S. K., Tse, K., & Toh, S. M. (2019). Stereotypes at work: Occupational stereotypes predict race and gender segregation in the workforce. *Journal of Vocational Behavior*, *115*, 103318. https://doi.org/10.1016/j.jvb.2019.103318

Holland, J. L. (1985). *Making vocational choices: A theory of vocational personalities and work environments* (2nd ed.). New Jersey: Prentice-Hall, Inc.

Hollander, M. A., & Parker, H. J. (1972). Occupational stereotypes and self-descriptions: Their relationship to vocational choice. *Journal of Vocational Behavior, 2*(1), 57–65. https://doi.org/10.1016/0001-8791(72)90007-3

IBM Corporation. (2017). IBM SPSS Statistics for Macintosh (Version 25). [Computer software]. Armonk, NY: Author.

Imhoff, R., Koch, A., & Flade, F. (2018). (Pre)occupations: A data-driven model of jobs and its consequences for categorization and evaluation. *Journal of Experimental Social Psychology*, *77*, 76-88. https://doi.org/10.1016/j.jesp.2018.04.001

Imhoff, R., Woelki, J., Hanke, S., & Dotsch, R: (2013). Warmth and competence in your face! Visual encoding of stereotype content. *Frontiers in Psychology, 4*, 386. https://doi.org/10.3389/fpsyg.2013.00386

Judd, C. M., James-Hawkins, L., Yzerbyt, V., and Kashima, Y. (2005). Fundamental dimensions of social judgment: understanding the relations between judgments of competence and warmth. *Journal of Personality and Social Psychololology, 89*(6), 899–913. doi: 10.1037/ 0022-3514.89.6.899

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. https://doi.org/10.1002/ejsp.1978

Kervyn, N., Fiske, S., & Yzerbyt, V. (2015). Forecasting the primary dimension of social perception. *Social Psychology, 46(*1), 36-45. https://doi.org/10.1027/1864-9335/a000219

Kline, R. B. (2010). *Principles and practice of structural equation modelling* (3rd ed.). New York, NY: Guilford.

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

Kotzur, P. F., Veit, S., Namyslo, A., Holthausen, M.-A., 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,* advance online publication.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology, 51*(1), 93–120. https://doi.org/10.1146/annurev.psych.51.1.93

Muthén, B. O., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociological Methods & Research, 47*(4), 637-664. https://doi.org/10.1177/0049124117701488

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Miller, L., & Hayward, R. (2006). New jobs, old occupational stereotypes: Gender and jobs in the new economy. *Journal of Education and Work, 19*(1), 67–93. https://doi.org/10.1080/13639080500523000

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The Measurement of Meaning*. University of Illinois Press.

Oswald, P. A. (2003). Sex-typing and prestige ratings of occupations as indices of occupational stereotypes. *Perceptual and Motor Skills, 97*(3), 953–959. https://doi.org/10.2466/pms.2003.97.3.953

Philbin, C. (2016, May 13). Too geeky for girls? Tech industry stereotypes are hindering equality. *The Guardian.* Retrieved July 02, 2020 from https://www.theguardian.com/sustainable-business/2016/may/13/geeky-image-girls-tech-industry-stereotypes-hinder-equality

Pratto, F., Stallworth, L. M., Sidanius, J., & Siers, B. (1997). The gender gap in occupational role attainment: A social dominance approach. *Journal of Personality and Social Psychology, 72*(1), 37-53. https://doi.org/10.1037/0022-3514.72.1.37

Ragins, B. R., & Sundstrom E. (1989). Gender and power in organizations: A longitudinal perspective. *Psychological Bulletin, 105*(1)*,* 51-88.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to Psychometric Theory*. New York: Routledge.

Rutjens, B. T., & Heine S. J. (2016). The immoral landscape? Scientists are associated with violations of morality. *PLoS ONE, 11* (4), e0152798. https://doi.org/10.1371/journal.pone.0152798

Sayans-Jímenez, 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. 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.

Sidanius, J., Liu, J. H., Pratto, F., & Shaw, J. S. (1994). Social dominance orientation, hierarchy attenuators and hierarchy-enhancers: Social dominance theory and the criminal justice system. *Journal of Applied Social Psychology, 24*(4), 338-366. https://doi.org/10.1111/j.1559-1816.1994.tb00586.x

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359-1366. https://doi.org/10.1177/0956797611417632

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

Sutcliffe, K. M., Lewton, E., & Rosenthal, M. M. (2004). Communication failures: An insidious contributor to medical mishaps. *Academic Medicine, 79*(2), 186-194. https://doi.org/10.1097/00001888-200402000-00019

The Fiske Lab. (n.d.). *Intergroup relations, social cognition, and social neuroscience.* https://www.fiskelab.org/cross-cultural-wc-maps/

Tourish, D. (2005). Critical upward communication: Ten commandments for improving strategy and decision making. *Long Range Planning, 38*(5), 485-503. https://doi.org/10.1016/j.lrp.2005.05.001

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.

Van Vuuren, M., Teurlings, J., & Bohlmeijer, E. (2012). Shared fate and social comparison: Identity work in the context of a stigmatized occupation. *Journal of Management & Organization, 18*(2), 263-280. doi:10.5172/jmo.2012.18.2.263

Wagner, U., Friehs, M.-T., & Kotzur, P. F. (2020). Das Bild der Polizei bei jungen Studierenden [The image of the police in the eyes of German students]. *Polizei und Wissenschaft*, *3*, 16-23.

Yzerbyt, V. Y., 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. https://doi.org/10.1177/1368430205053944

Tables and Figures

Table 1

Sample Composition Regarding Gender, Educational Background, Occupational Status, and Identification with Surveyed Groups

	Total	Percentage
Gender		
Female	284	66.8
Male	134	31.5
Diverse	2	0.5
Missing	5	1.2
Educational Background		
No school leaving certificate	1	0.2
Certificate of Secondary Education (9th grade)	4	0.9
General Certificate of Secondary Education (10th grade)	35	8.2
University entrance diploma	142	33.4
Undergraduate degree	78	18.4
Graduate degree	127	29.9
Doctoral degree	35	8.2
Missing	3	0.7
Main occupation		
School student	5	1.2
University student	123	28.9
Gainfully employed	262	61.6
Pensioners	14	3.3
Unemployed	13	3.1
Other	8	1.9

Table 1 (continued)

	Total	Percentage
Identification with (multiple answers possible)		
Unemployed people	8	1.9
Pensioners	9	2.1
Physicians	19	4.5
Bankers	3	0.7
Hospital and elderly care nurses	24	5.6
Childcare workers	30	7.1
Judges	2	0.5
Teachers	59	13.9
Politicians	5	1.2
Farmers	7	1.6
Craftspeople	23	5.4
Firefighters	6	1.4
Police officers	3	0.7
None of the above	275	64.7

Table 2

Item Level Descriptive Statistics of the Stereotype Content Measures per Occupational Group

Item	M	SD	Min	Max
Group *Unemployed people* (*n* = 417)				
Dishonest (1) – Honest (5)	2.30	0.98	1.00	5.00
Unfriendly (1) – Friendly (5)	2.42	0.89	1.00	5.00
Ill-natured (1) – Good-natured (5)	2.63	0.80	1.00	5.00
Cold (1) – Warm (5)	2.65	0.82	1.00	5.00
Careless (1) – Thorough (5)	2.04	0.95	1.00	5.00
Incompetent (1) – Competent (5)	1.95	0.95	1.00	5.00
Lazy (1) – Hardworking (5)	1.66	0.82	1.00	5.00
Inefficient (1) – Efficient (5)	1.72	0.90	1.00	5.00
Group *Pensioners* (*n* = 416)				
Dishonest (1) – Honest (5)	3.78	0.90	1.00	5.00
Unfriendly (1) – Friendly (5)	3.18	1.03	1.00	5.00
Ill-natured (1) – Good-natured (5)	3.54	0.96	1.00	5.00
Cold (1) – Warm (5)	3.58	0.95	1.00	5.00
Careless (1) – Thorough (5)	3.30	0.94	1.00	5.00
Incompetent (1) – Competent (5)	2.96	0.86	1.00	5.00
Lazy (1) – Hardworking (5)	2.83	0.95	1.00	5.00
Inefficient (1) – Efficient (5)	2.21	0.94	1.00	5.00
Group *Physicians* (*n* = 406)				
Dishonest (1) – Honest (5)	4.07	0.87	1.00	5.00
Unfriendly (1) – Friendly (5)	3.55	0.83	1.00	5.00
Ill-natured (1) – Good-natured (5)	3.81	0.81	1.00	5.00
Cold (1) – Warm (5)	3.17	0.99	1.00	5.00
Careless (1) – Thorough (5)	4.16	0.81	1.00	5.00
Incompetent (1) – Competent (5)	4.33	0.73	2.00	5.00
Lazy (1) – Hardworking (5)	4.43	0.73	1.00	5.00
Inefficient (1) – Efficient (5)	4.60	0.61	2.00	5.00

Table 2 (continued)

Item	M	SD	Min	Max
Group *Bankers* (*n* = 422)				
Dishonest (1) – Honest (5)	2.36	1.05	1.00	5.00
Unfriendly (1) – Friendly (5)	3.19	1.13	1.00	5.00
Ill-natured (1) – Good-natured (5)	2.35	0.90	1.00	5.00
Cold (1) – Warm (5)	1.81	0.84	1.00	5.00
Careless (1) – Thorough (5)	3.47	1.12	1.00	5.00
Incompetent (1) – Competent (5)	3.52	1.03	1.00	5.00
Lazy (1) – Hardworking (5)	3.39	1.01	1.00	5.00
Inefficient (1) – Efficient (5)	3.49	1.04	1.00	5.00
Group *Hospital and elderly care nurses* (*n* = 271)				
Dishonest (1) – Honest (5)	4.16	0.85	1.00	5.00
Unfriendly (1) – Friendly (5)	3.80	0.94	1.00	5.00
Ill-natured (1) – Good-natured (5)	4.23	0.79	2.00	5.00
Cold (1) – Warm (5)	4.22	0.84	1.00	5.00
Careless (1) – Thorough (5)	3.83	0.91	1.00	5.00
Incompetent (1) – Competent (5)	3.79	0.91	1.00	5.00
Lazy (1) – Hardworking (5)	4.46	0.79	1.00	5.00
Inefficient (1) – Efficient (5)	4.34	0.81	2.00	5.00
Group *Childcare workers* (*n* = 271)				
Dishonest (1) – Honest (5)	4.06	0.85	1.00	5.00
Unfriendly (1) – Friendly (5)	4.29	0.72	1.00	5.00
Ill-natured (1) – Good-natured (5)	4.37	0.73	2.00	5.00
Cold (1) – Warm (5)	4.52	0.69	1.00	5.00
Careless (1) – Thorough (5)	3.85	0.71	2.00	5.00
Incompetent (1) – Competent (5)	3.72	0.84	1.00	5.00
Lazy (1) – Hardworking (5)	3.90	0.89	2.00	5.00
Inefficient (1) – Efficient (5)	3.87	0.92	2.00	5.00

Table 2 (continued)

Item	M	SD	Min	Max
Group *Judges* (*n* = 284)				
Dishonest (1) – Honest (5)	4.11	0.91	1.00	5.00
Unfriendly (1) – Friendly (5)	2.95	0.88	1.00	5.00
Ill-natured (1) – Good-natured (5)	3.07	0.82	1.00	5.00
Cold (1) – Warm (5)	2.16	0.82	1.00	4.00
Careless (1) – Thorough (5)	4.23	0.83	1.00	5.00
Incompetent (1) – Competent (5)	4.32	0.78	1.00	5.00
Lazy (1) – Hardworking (5)	4.05	0.86	1.00	5.00
Inefficient (1) – Efficient (5)	4.12	0.88	1.00	5.00
Group *Teachers* (*n* = 244)				
Dishonest (1) – Honest (5)	3.75	0.84	1.00	5.00
Unfriendly (1) – Friendly (5)	3.53	0.81	1.00	5.00
Ill-natured (1) – Good-natured (5)	3.51	0.81	1.00	5.00
Cold (1) – Warm (5)	3.51	0.89	1.00	5.00
Careless (1) – Thorough (5)	3.63	0.81	1.00	5.00
Incompetent (1) – Competent (5)	3.50	0.90	1.00	5.00
Lazy (1) – Hardworking (5)	3.27	1.07	1.00	5.00
Inefficient (1) – Efficient (5)	3.47	0.95	1.00	5.00
Group *Politicians* (*n* = 280)				
Dishonest (1) – Honest (5)	2.19	0.96	1.00	5.00
Unfriendly (1) – Friendly (5)	2.83	0.93	1.00	5.00
Ill-natured (1) – Good-natured (5)	2.44	0.81	1.00	5.00
Cold (1) – Warm (5)	1.90	0.82	1.00	5.00
Careless (1) – Thorough (5)	2.63	1.01	1.00	5.00
Incompetent (1) – Competent (5)	2.51	1.03	1.00	5.00
Lazy (1) – Hardworking (5)	2.86	1.12	1.00	5.00
Inefficient (1) – Efficient (5)	2.96	1.11	1.00	5.00

Table 2 (continued)

Item	M	SD	Min	Max
Group *Farmers* (*n* = 279)				
Dishonest (1) – Honest (5)	3.69	0.93	1.00	5.00
Unfriendly (1) – Friendly (5)	3.28	1.03	1.00	5.00
Ill-natured (1) – Good-natured (5)	3.63	0.87	1.00	5.00
Cold (1) – Warm (5)	3.25	0.99	1.00	5.00
Careless (1) – Thorough (5)	3.48	0.93	1.00	5.00
Incompetent (1) – Competent (5)	3.73	0.95	1.00	5.00
Lazy (1) – Hardworking (5)	4.51	0.74	1.00	5.00
Inefficient (1) – Efficient (5)	4.27	0.85	1.00	5.00
Group *Craftspeople* (*n* = 264)				
Dishonest (1) – Honest (5)	3.27	0.92	1.00	5.00
Unfriendly (1) – Friendly (5)	3.17	0.91	1.00	5.00
Ill-natured (1) – Good-natured (5)	3.20	0.78	1.00	5.00
Cold (1) – Warm (5)	2.92	0.87	1.00	5.00
Careless (1) – Thorough (5)	3.34	0.86	1.00	5.00
Incompetent (1) – Competent (5)	3.50	0.91	1.00	5.00
Lazy (1) – Hardworking (5)	3.75	0.95	1.00	5.00
Inefficient (1) – Efficient (5)	4.00	0.85	2.00	5.00
Group *Firefighters* (*n* = 280)				
Dishonest (1) – Honest (5)	4.57	0.64	2.00	5.00
Unfriendly (1) – Friendly (5)	4.22	0.75	1.00	5.00
Ill-natured (1) – Good-natured (5)	4.31	0.71	3.00	5.00
Cold (1) – Warm (5)	3.90	0.82	1.00	5.00
Careless (1) – Thorough (5)	4.55	0.61	3.00	5.00
Incompetent (1) – Competent (5)	4.51	0.67	2.00	5.00
Lazy (1) – Hardworking (5)	4.51	0.70	2.00	5.00
Inefficient (1) – Efficient (5)	4.61	0.62	2.00	5.00

Table 2 (continued)

Item	M	SD	Min	Max
Group *Police officers* (*n* = 280)				
Dishonest (1) – Honest (5)	3.80	0.92	1.00	5.00
Unfriendly (1) – Friendly (5)	3.05	0.90	1.00	5.00
Ill-natured (1) – Good-natured (5)	3.10	0.85	1.00	5.00
Cold (1) – Warm (5)	2.47	0.87	1.00	5.00
Careless (1) – Thorough (5)	3.86	0.86	1.00	5.00
Incompetent (1) – Competent (5)	3.74	0.91	1.00	5.00
Lazy (1) – Hardworking (5)	3.92	0.83	1.00	5.00
Inefficient (1) – Efficient (5)	4.05	0.92	1.00	5.00

Note. N = Number of participants, M = mean value, SD = standard deviation, Min = minimum value, Max = maximum value. All scales ranging from 1 to 5. German translation of the items in the order of the table: *Unaufrichtig – Aufrichtig, Unfreundlich – Freundlich, Bösartig – Gutmütig, Kühl – Warmherzig, Nachlässig – Sorgfältig, Inkompetent – Kompetent, Arbeitsscheu – Fleißig, Leistungsschwach – Leistungsfähig.*

.

Table 3

Baseline Model Fit Indices for each Occupational Group

Group	N	AIC	BIC	χ^2	df	*p*	RMSEA [90% CI]	CFI	SRMR
Unemployed people	397	5,504.833	8,896.463	13.898	12	0.304	0.020 [0.000 - 0.057]	0.997	0.020
Pensioners	395	6,733.358	3,824.873	25.956	12	0.011	0.054 [0.025 - 0.083]	0.960	0.043
Physicians	385	5,459.668	5,550.593	29.621	12	0.003	0.062 [0.034 - 0.090]	0.967	0.037
Bankers	401	7,035.168	7,127.029	16.325	12	0.177	0.030 [0.000 - 0.063]	0.992	0.031
Hospital and elderly care nurses	253	3,899.757	3,981.025	11.826	12	0.460	0.000 [0.000 - 0.063]	1.000	0.028
Childcare workers	253	3,522.220	3,603.488	21.776	12	0.040	0.057 [0.012 - 0.094]	0.974	0.042
Judges	266	4,123.852	4,206.272	20.296	12	0.062	0.051 [0.000 - 0.088]	0.961	0.048
Teachers	234	3,652.702	3,732.174	8.491	12	0.746	0.000 [0.000 - 0.048]	1.000	0.026
Politicians	267	4,578.423	4,660.930	20.795	12	0.054	0.052 [0.000 - 0.089]	0.970	0.043
Farmers	266	4,195.271	4,277.692	15.904	12	0.196	0.035 [0.000 - 0.076]	0.989	0.033
Craftspeople	253	3,952.339	4,033.607	13.788	12	0.314	0.024 [0.000 - 0.071]	0.995	0.033
Firefighters	269	3,440.670	3,523.349	22.391	12	0.033	0.057 [0.016 - 0.093]	0.963	0.040
Police officers	269	4,311.303	4,393.981	15.716	12	0.205	0.034 [0.000 - 0.075]	0.988	0.028

Note. N = number of participants, AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion, CFI = comparative fit index, χ^2 = Chi square-value, *df* = degrees of freedom, *p* = probability value, RMSEA = root mean square error of approximation, CI = confidence interval, SRMR= standardized root mean square residual.

Table 4

Ranking of Occupational Groups, Factor Means, Variances and Significant Mean Differences between Occupational Groups in Terms of Warmth

			Warmth Factor											
Ranking	Group #	Group Label	Mean	Variance	Groups with Significantly Smaller Factor Mean									
1	12	Firefighters	3.119	0.734	3	10	8	2	11	13	7	1	9	4
2	6	Childcare workers	3.051	0.654	8	2	11	13	7	1	9	4		
3	5	Hospital and elderly care nurses	2.907	1.024	3	10	8	2	11	13	7	1	9	4
4	3	Physicians	2.032	1.062	2	11	13	7	1	9	4			
5	10	Farmers	1.831	1.632	11	13	7	1	9	4				
6	8	Teachers	1.687	1.129	11	13	7	1	9	4				
7	2	Pensioners	1.433	1.400	13	7	1	9	4					
8	11	Craftspeople	1.048	0.950	1	9	4							
9	13	Police officers	0.781	0.906	1	9	4							
10	7	Judges	0.690	0.725	1	9	4							
11	1	Unemployed people	0.000*	1.000	9	4								
12	9	Politicians	-0.290	0.673										
13	4	Bankers	-0.481	1.208										

Note. * The factor mean of the occupational group *Unemployed people* was fixed to zero in the fixed alignment optimization approach. *p* < .05.

Table 5

Ranking of Occupational Groups, Factor Means, Variances and Significant Mean Differences between Occupational Groups in Terms of Competence

Competence Factor														
Ranking	Group #	Group Label	Mean	Variance	Groups with Significantly Smaller Factor Mean									
1	12	Firefighters	3.299	0.329	6	10	5	8	4	13	11	2	9	1
2	7	Judges	2.860	0.479	10	5	8	4	13	11	2	9	1	
3	3	Physicians	2.696	0.391	6	10	5	8	4	13	11	2	9	1
4	6	Childcare workers	2.154	0.651	2	9	1							
5	10	Farmers	2.057	0.501	11	2	9	1						
6	5	Hospital and elderly care nurses	2.055	0.498	2	9	1							
7	8	Teachers	1.905	0.567	2	9	1							
8	4	Bankers	1.896	0.845	2	9	1							
9	13	Police officers	1.847	0.590	2	9	1							
10	11	Craftspeople	1.731	0.634	2	9	1							
11	2	Pensioners	1.145	0.482	9	1								
12	9	Politicians	0.748	0.830	1									
13	1	Unemployed people	0.000*	1.000										

Note. * The factor mean of the occupational group *Unemployed people* was fixed to zero in the fixed alignment optimization approach. $p < .05$.

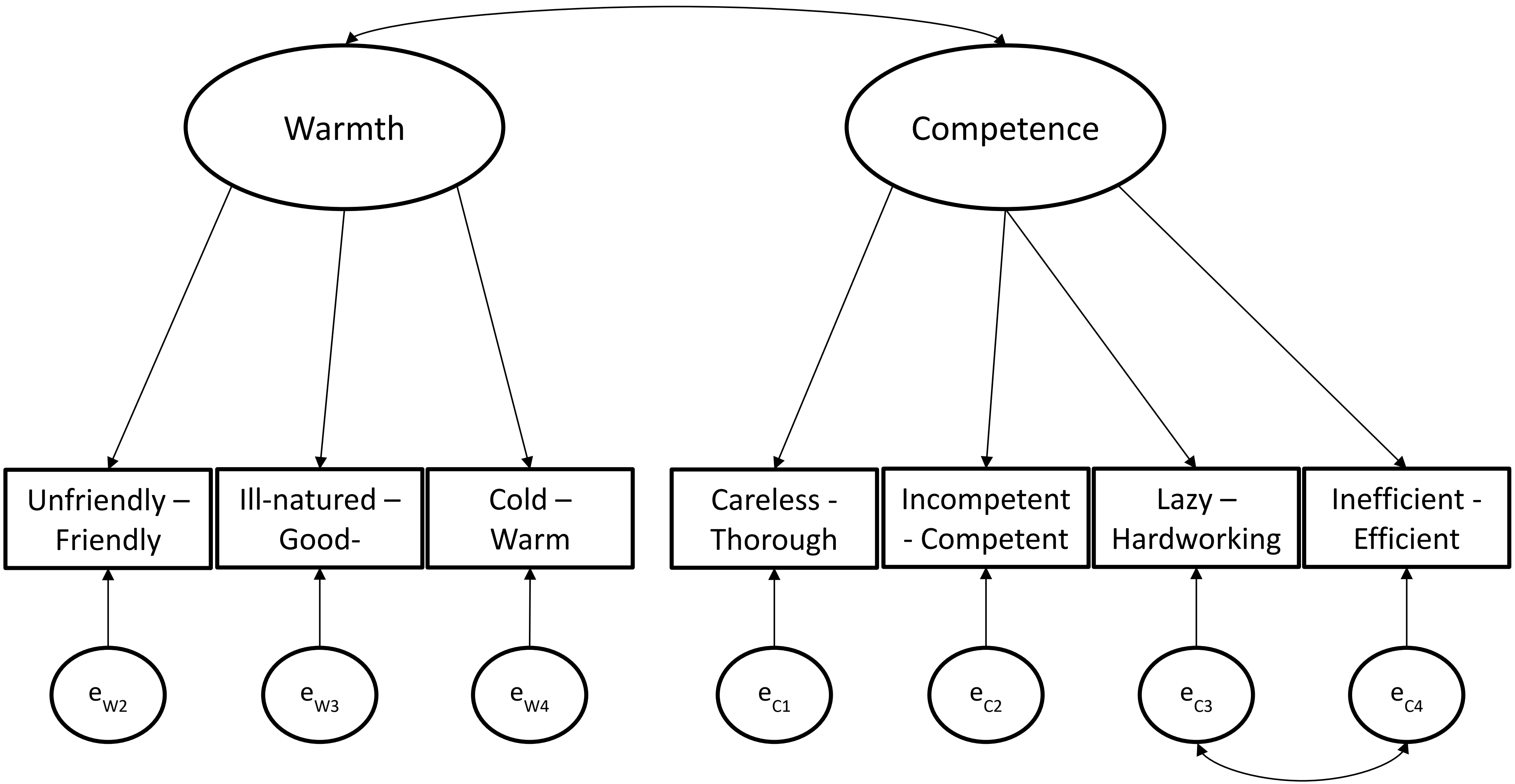


Figure 1. Final measurement model of each occupational group in the alignment optimization procedure. e = measurement error, W = warmth indicator, C = competence indicator.

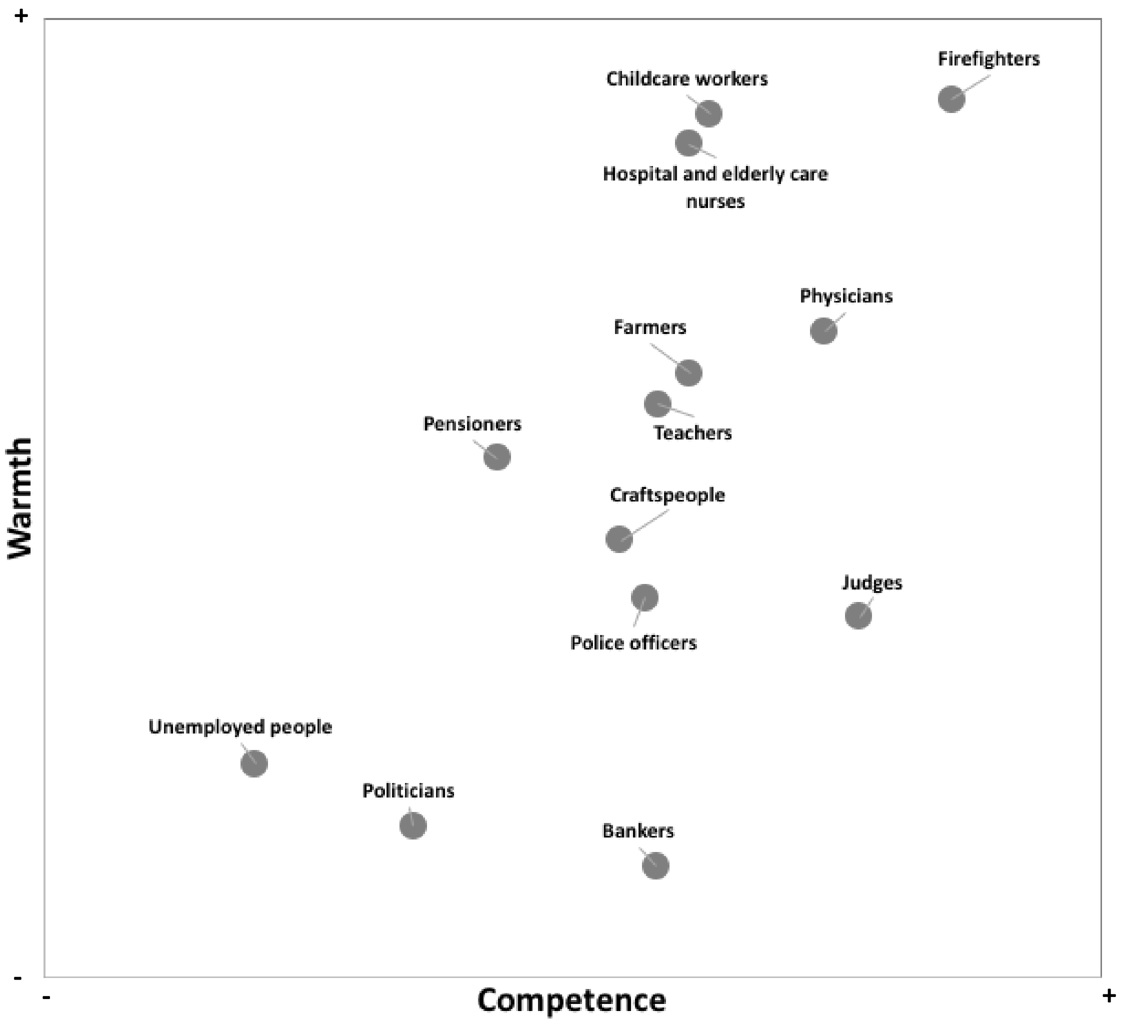


Figure 2. Latent warmth (y-axis) and competence (x-axis) mean values for each occupational group. Scaling was achieved by constraining the latent mean values of warmth and competence for the occupational group *Unemployed people* to zero.

Supplementary Materials for Manuscript # 4

The supplementary materials for Manuscript # 4 are stored in the Open Science Framework, see https://osf.io/gxz49/?view_only=99f65516c497496b89b089b0c032a088.

They include:

- All used datasets;
- The results of the pilot study to identify relevant occupational groups;
- A German and an English version of the used survey;
- Further descriptive information of the sample;
- Item-intercorrelations per occupational group;
- Detailed EFA results including analysis code and complete outputs;
- The initial CFA (with all items and without residual covariance) model fit results;
- Further information on potential CFA model adaptations;
- Detailed CFA solutions of the final measurement model including analysis code and complete outputs for the split CFA sample and complete sample;
- The detailed configural measurement invariance model, including analysis code and complete outputs;
- The detailed results of the alignment optimization procedure; including analysis code and complete outputs.

Maria-Therese Friehs

Curriculum vitae

(née Wiemer)

Born 27.11.1992
In Ribnitz-Damgarten

Friedrich-Ebert-Straße 25
76829 Landau (Pfalz)
Germany

Maria-Therese.Friehs@fernuni-hagen.de

ORCID: 0000-0002-5897-8226
OSF: https://osf.io/8awfx/

Academic Career

Since 04/2021	**Research Associate** Chair for Psychological Methods and Evaluation, *FernUniversität in Hagen, Germany*
02/2020 – 03/2021	**Research Associate** Project *Implementationsprozesse in formalen Bildungssettings (ImproBis)*, Work Unit Developmental and Educational Psychology, *University Koblenz-Landau, Germany*
04/2018 – 03/2020	**Research Associate** Working Unit Developmental Psychology, *Osnabrück University, Germany*
09/2017 – 12/2018	**Referee for Evaluation** Project *ProPraxis* of the *Qualitätsoffensive Lehrerbildung,* *Philipps-University of Marburg, Germany*

Education

10/2016 – 03/2018	**Master of Science Psychology (Grade 0.8)** Master thesis: "Testing the universal meaning of stereotype content dimensions as basic organisers of social perceptions of groups within and across regional samples" (Grade 0.7), *Philipps-University of Marburg, Germany*
10/2012 – 09/2016	**Bachelor of Science Psychology (Grade 1.3)** Bachelor thesis: „You may choose your friends, but not your neighbours – The effects of contact on prejudice in the neighbourhood of an asylum seeker reception centre" (Grade 0.7[1]), *Philipps-University of Marburg, Germany*

Publications

In Press/Published with Peer-Review (sorted alphabetically)

Bohrer, B., **Friehs, M.-T.**, Schmidt, P., & Weick, S. (2019). Contact between natives and migrants in Germany: Perceptions of the native population since 1980 and an examination of the contact hypotheses. *Social Inclusion, 7*(4), 320-331. https://doi.org/10.17645/si.v7i4.2429

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

O'Donnell, A., **Friehs, M.-T**., Bracegirdle, C., Zuñiga, C., Watt. S. E., & Barlow, F. K. (2021). Technological and analytical advancements in intergroup contact research. *Journal of Social Issues, 77*, 171-196. https://doi.org/10.1111/josi.12424

Wagner, U., Kotzur, P. F., & **Friehs, M.-T.** (in press). Anti-immigrant prejudice and discrimination in Europe. In M. Augoustinos, K. Durrheim, & C. Tileage (Eds.), *International Handbook of Prejudice, Stereotyping and Discrimination*. Oxford: Routledge.

Wagner, U., Tachtnoglou, S., Kotzur, P. F., **Friehs, M.-T.**, & Kemmesies, U. (2020). Proportion of foreigners negatively predicts the prevalence of xenophobic hate crimes within German districts. *Social Psychology Quarterly, 83*(2), 195-205. https://doi.org/10.1177/0190272519887719

Submitted/under review with Peer-Review (sorted alphabetically)

Friehs, M.-T., Aparicio Lukassowitz, F., & Wagner, U. (2020). *Stereotype content of occupational groups in Germany.* Manuscript under review (First submission at the International Journal of Social Psychology).

Friehs, M.-T., Böttcher, J., Kotzur, P. F., Lüttmer, T., Asbrock, F., Wagner, U., Asbrock, F., & van Zalk, M. (2021*). Examining the Structural Validity of English Stereotype Content Measures – A Preregistered Re-Analysis of Published Data and Discussion of Possible Future Directions.* Manuscript under review (First submission at the Personality and Social Psychology Bulletin). https://doi.org/10.31234/osf.io/dej4m

Friehs, M.-T.*, Kotzur, P. F.*, Zöller, A.-K., Wagner, U., & Asbrock, F. (2021). *A Preregistered Replication of Scale Properties Stereotype Content Measures: The German case.* Manuscript under review (First submission at the International Journal of Social Psychology). https://doi.org/10.31234/osf.io/fa39w * Shared first authorship

Kotzur, P. F., **Friehs, M.-T.**, Schmidt, P., Wagner, U.., Pötzschke, S., & Weiss, B. (2020). *Attitudes towards refugees: Introducing a short three-dimensional scale.* Manuscript under review (First submission at the British Journal of Social Psychology).

Publications without Peer-Review, Book Chapters and Monographs (sorted alphabetically)

Schmidt, P., **Friehs, M.-T.**, Gloris, D., & Grote, H. (2020). Panel Conditioning or Socratic Effect revisited: 99 citations, but is there theoretical progress? In A. Mays, A. Dingelstedt, V. Hambauer, S. Schlosser, F., Berens, J. Leibold & J.-K. Höhne (Eds.), *Grundlagen – Methoden – Anwendungen in den Sozialwissenschaften,* 25-65. Wiesbaden: Springer VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-658-15629-9_2

Stellmacher, J., & **Friehs, M.-T.** (2020). Die Marburger Praxismodule: Eine evaluationsgestützte Reflexion der ersten Projektphase von ProPraxis (The Marburg Practical Modules: An evaluation-based reflection of the first project phase of ProPraxis). In S. Uhl (Ed.), *Das Lehramtsstudium und seine Praxisphasen: Konzepte, Strukturen, Erfahrungen*, 107-128*.* Bielefeld: wbv Media GmbH.

Wagner, U., **Friehs, M.-T.**, & Kotzur, P. F. (2020). Das Bild der Polizei bei jungen Studierenden (The image of the police in young students). *Polizei und Wissenschaft, 3,* 16-23.

Wagner, U., Suppmann, A., Siebert, R., **Friehs, M.-T.**, & Kotzur, P. F. (2019). Das Image der deutschen Polizei bei jungen Geflüchteten (The image of the police in young refugees). *Polizei und Wissenschaft, 1,* 10-13.

Stellmacher, J., & **Friehs, M.-T.** (2019). Ergebnisbericht im Rahmen des BMBF geförderten Vorhabens „ProPraxis – Gymnasiale Lehrerbildung in Marburg: professionell, praktisch, gut“ (Evaluation report of the project „ProPraxis“). https://www.uni-marburg.de/de/zfl/downloads/evaluation/2019-08-28-zweiter-evaluationsbericht-propraxis-final-v2-2.pdf

Conference Presentations (sorted by date)

09/2019 **Friehs, M.-T.**, Müller, L., & van Zalk, M. (2019). Subjective Intergroup Contact Experiences - A German Replication. Poster presented at the 17. Tagung der Fachgruppe Sozialpsychologie of the Deutsche Gesellschaft für Psychologie (DGPs), Cologne, Germany. See https://osf.io/hmyqa/

09/2019 Wagner, U., Kotzur, P. F., & **Friehs, M.-T.** (2019). *Anti-immigrant prejudice and discrimination in Europe.* Paper presented at the 17. Tagung der Fachgruppe Sozialpsychologie of the Deutsche Gesellschaft für Psychologie (DGPs), Cologne, Germany.

07/2019 **Friehs, M.-T.**, Schmidt, P., & van Zalk, M. (2019). Longitudinal intergroup contact effects - Disentangling within- and between-person variation in intergroup contact and outgroup attitudes. Paper presented at the 8th Conference of the European Survey Research Association (ESRA), Zagreb, Croatia. See https://osf.io/9cfpj/

05/2019 **Friehs, M.-T.**, & van Zalk, M. H. W. (2019). Longitudinal intergroup contact effects - Disentangling between-person differences and within-person processes over time. Paper presented at the SASP-SPSSI Group meeting on intergroup contact, Newcastle, Australia.

04/2019 **Friehs, M.-T.**, & Kotzur, P. (2019). The stereotype content model - Benefits and challenges of examining group-related differences on basic dimensions of social perception in a latent modelling framework. Paper presented at the 48th annual conference of the Society of Australasian Social Psychologists (SASP), Sydney, Australia.

10/2018 **Friehs, M.-T.**, Schmidt, P., Wagner, U., & Kotzur, P. F. (2018). *Intergruppenkontakt und Vorurteile (Intergroup contact and prejudice).* Paper presented at the GESIS-IEDI-Tagung zu Migration und interethnischen Beziehungen, Köln, Germany. See https://osf.io/un89f/

09/2018 **Friehs, M.-T.**, & Kotzur, P. F. (2018). *The stereotype content model - Testing for group differences on basic social perception dimensions in a latent modelling framework*. Paper presented at the 51. Kongress der Deutschen Gesellschaft für Psychologie (DGPs), Frankfurt, Germany. See https://osf.io/d4pve/

09/2018 Stellmacher, J., & **Friehs, M.-T.** (2018). *Evaluation der Marburger Praxismodule – Ergebnisse einer Längsschnittstudie zur Kompetenzentwicklung von Studierenden im Lehramt (Evaluation of the Marburg Practical Modules – Results of a longitudinal study of competence development of student teachers).* Poster presented at the 51. Kongress der Deutschen Gesellschaft für Psychologie (DGPs), Frankfurt, Germany. See https://osf.io/ubers/

09/2018 Wagner, U., **Friehs, M.-T.**, Kotzur, P. F., Lemmer, G., & Matick, E. (2018). *Effects of intergroup contact on attitudes towards refugees and minorities.* Paper presented at the 51. Kongress der Deutschen Gesellschaft für Psychologie (DGPs), Frankfurt, Germany.

06/2018 Wagner, U., Tachtsoglou, S., Kotzur, P. F., **Wiemer, M. T**., & Kemmesies, U. (2018). *Kriminalität gegen Zuwanderer: Fremdenfeindliche Straftaten sind dort häufiger, wo der Ausländeranteil niedrig ist (Crime against immigrants: Xenophobic hate crimes are more frequent where the proportion of foreigners is lower).* Paper presented at the Forum KI, Bundeskriminalamt, Wiesbaden, Germany.

09/2017 **Wiemer, M. T.**, Kotzur, P. F., Asbrock, F., & Wagner, U. (2017). The stereotype content model – Testing for differences in refugee groups' assessment in East- and West Germany. Paper presented at the 16. Tagung der Fachgruppe Sozialpsychologie of the Deutsche Gesellschaft für Psychologie (DGPs), Ulm, Germany.

07/2017 **Wiemer, M. T.**, Kotzur, P. F., Wagner, U., & Asbrock, F. (2017). The stereotype content model: Testing cross-group and regional comparability of warmth and competence assessments. Paper presented at the 9th regional conference of the International Association of Cross-Cultural Psychology (IACCP), Warsaw, Poland.

06/2017 **Wiemer, M. T**., Kotzur, P. F., Wagner, U., & Asbrock, F. (2017). *On the equality of social evaluations – Are the concepts of warmth and competence equivalent across social groups and different regions?* Paper presented at the 30. Tagung des Forums Friedenspsychologie, Chemnitz, Germany.

Teaching Experience

Extracurricular Teaching

08/2019
08/2018 **Tutor of the course "Factor analysis & Structural equation modeling with Mplus"**
Teacher: Prof. Dr. Peter Schmidt
Essex Summer School in Social Science Data Analysis,
University of Essex, *Colchester*, *United Kingdom*

Courses for Master's Degree Students (sorted by date)

2021 **Seminar „Multivariate Verfahren und computergestützte Datenanalyse I & II" (Multivariate and computer-based data analysis I & II)** for Psychology Master degree students,
FernUniversität in Hagen, Germany

2019/20
2018/19
2018 **Seminar „(Entwicklungs-) Herausforderungen im Schulkontext und Möglichkeiten zur Prävention und Intervention" ((Developmental) challenges in the context of schools and options für prevention and intervention)** for student teachers (B. Edu., M. Edu.),
Osnabrück University, Germany

2019 **MA Colloquium of the Work Unit Developmental Psychology,**
Osnabrück University, Germany

2019 **Seminar „Development and Culture"** for students of the Intercultural Psychology master degree,
Osnabrück University, Germany

Courses for Bachelor's Degree Students (sorted by date)

2020/21 **Tutor of the Empirical Practical Course** for Psychology Students, *University Koblenz-Landau, Germany*

2018/19 **Seminar „Die Psychologie der Lebensspanne" (Life-span psychology)** for Nursing Science students,
Osnabrück University, Germany

2018/19 **Tutor of the Empirical Practical Course** for Psychology Students,
Osnabrück University, Germany

2017/18 **Seminar „Diskriminierung" (Discrimination)** for Psychology Students,
Philipps-University of *Marburg, Germany*

2016/17 **Seminars „Intergroup Contact" and „Intergruppenprozesse" (Intergroup Processes)** for Psychology Students,
Philipps-University of *Marburg, Germany*

Supervision of Bachelor and Master Thesis and Research Internees

Master Theses (First Supervisor, sorted by date)

Aparicio Lukassowitz, F. (2020). *Occupational stereotypes in Germany in terms of warmth and competence.* Universität Osnabrück, Deutschland.

Bhatti, S. (2019). *Mental health and psychosocial support (MHPSS)– Challenges and chances for suicide prevention within Local and Syrian communities in Jordan.* Universität Osnabrück, Deutschland.

Giourga, A. M. (2019). *How do contact experiences with the majority affect ethnic minorities? – A study of the impact of positive and negative subjective contact experiences on majority stereotypes, acculturation orientation, and physical wellbeing of Greek minority members in Austria.* Universität Osnabrück, Deutschland.

Müller, L. (2019). *Positive and negative subjective intergroup contact experiences influence attitudes towards ethnic and non-ethnic outgroups.* Universität Osnabrück, Deutschland.

Speer, A. (2019). *Development and validation of and SDO and RWA state measure.* Universität Osnabrück, Deutschland.

Ehrler, L. (2019). *Does intergroup contact moderate moral licensing effects?* Universität Osnabrück, Deutschland.

Bachelor Thesesarbeiten (First Supervisor, sorted by date)

Masselmann, J. (2021). *Unobserved population heterogeneity of the syndrome of Group-Focused Enmity in Germany and its covariates*. Osnabrück University, Germany.

Schuka, K. (2021). *Kommunikationsverhalten unter Lehrkräften – Eine Validierungsstudie (Communication among teachers – A validation study).*
University Koblenz-Landau, Germany.

Research Internees

2021	**Johanna Böttcher,** *FernUniversität in Hagen, Germany*
2021	**Tamar Unger,** *FernUniversität in Hagen, Germany*
2020	**Johanna Böttcher**, *Osnabrück University, Germany*
2020	**Tabea Lüttmer**, *Osnabrück University, Germany*
2019	**Ann-Kristin Zöller**, *Osnabrück University, Germany*

Workshops Attendance (Selection)

05/2019 **"Life hacks for quantitative research methods lecturers – Data specialist training and course materials for your classroom"**
GESIS Cologne, Germany

07/2018
“Advanced social network analysis – Statistical analysis for cross-sectional and longitudinal SNA”
Teacher: Filip Agneessens
Essex Summer School in Social Science Data Analysis,
University of Essex, *Colchester*, *United Kingdom*

02/2018
“Strukturgleichungsmodelle für Längsschnittdaten in R“
(Structural equation modelling for longitudinal data in R)
Teachers: Elisabeth Prestele, Dorota Reis
University Koblenz-Landau, Germany

11/2017
„Mathematics for research: A refresher course in R”
Teacher: Michael Greenacre
GESIS Cologne, Germany

09/2017
„Angewandte Integration qualitativer und quantitativer Methoden“
(Applied integration of mixed methods)
Teacher: Jörg Stolz
GESIS Mannheim, Germany

12/2016
“Open science”
Georg-August University Göttingen, Germany

04/2015
“Meta-analysis and meta-analytic research”
Teacher: George A. Kelley
University of Limerick, Ireland

External Research Funding and Prices

12/2020
Open Science Badge Gold 2020
University Koblenz-Landau, Germany

09/2020
Support grant for young-career researchers from the German Psychological Society, Social Psychology Group (750€)

09/2019
Price for advising the best master thesis at the Institute for Psychology, (1.000€)
Osnabrück University, Germany

05/2019
DAAD Travelling grant for a conference visit in Newcastle, Australia (ca. 2.100€)

01/2012 – 03/2018
Study grant of the *German National Study Foundation (Studienstiftung des Deutschen Volkes)* (ca. 22.500€)

03/2015 – 06/2015
Erasmus Practical Placement (ca. 1.800€)
Grant for a research internship at the Department Cognitive Social Psychology, *University of Limerick*, *Ireland*

Further Academic Activities

12/2020 – 03/2021
Member of the Open Science Committee*, University Koblenz-Landau, Germany*

Since 12/2020
Member of the Selection Committee of the *German National Study Foundation*

Since 06/2020
Guest editor of the Focus Section „Group-focused enmity – Conceptual, longitudinal, and cross-national perspectives based on pre-registered studies” in the *International Journal of Conflict and Violence*

Since 10/2019
Co-Organisator of the International Summer School 2022, *University of Osnabrück, Germany*

Ad Hoc Review Activity

International Review of Social Psychology

Memberships

German Psychological Society (DGPs)

International Association for Cross-Cultural Psychology (IACCP)

Friehs

Landau, May 28, 2021

Maria-Therese Friehs

Eidesstattliche Erklärung

Hiermit erkläre ich, Maria-Therese Friehs, dass ich die Synopse der vorliegenden Dissertation eigenständig ohne unzulässige Inanspruchnahme Dritter verfasst habe und keine anderen als die angegebenen Hilfsmittel verwendet habe. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken habe ich als solche gekennzeichnet.

Für die vier im Rahmen dieser Dissertation verfassten Publikationen wurden folgende individuelle Beiträge von den einzelnen Autor*innen (definiert nach dem CRediT-System) erbracht:

Manuskript # 1

Friehs, M.-T., Böttcher, J., Kotzur, P. F., Lüttmer, T., Wagner, U., Asbrock, F., & van Zalk, M. H. W. (2021). *Examining the structural validity of stereotype content measures – A preregistered re-analysis of published data and discussion of possible future directions*. Manuscript under review.

Individuelle Beiträge der Autor*innen nach dem CRediT System:

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

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

PFK: Conceptualization, Methodology, Project administration, Supervision, 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

Manuskript # 2

Friehs, M.-T.*, Kotzur, P. F.*, Zöller, A.-K. C., Wagner, U., & Asbrock, F. (2021*). A preregistered examination of scale properties of stereotype content measures: The German case.* Manuscript under review. * Geteilte Erstautorenschaft

Individuelle Beiträge der Autor*innen nach dem CRediT System:

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

PFK: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing – Original draft

AKCZ: Data curation, Formal analysis, Methodology, Writing – Original draft

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

FA: Conceptualization, Resources, Writing – Review & editing

Manuskript # 3

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

Individuelle Beiträge der Autor*innen nach dem CRediT System:

PFK: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Validation, Writing – Original draft

MTF: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – Original draft

FA: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – Review & editing

MHWVZ: Resources, Writing – Review & editing

Manuskript # 4

Friehs, M. T., Aparicio Lukassowitz, F., & Wagner, U. (2020*). Stereotype content of occupational groups in Germany.* Manuscript under review.

Individuelle Beiträge der Autor*innen nach dem CRediT System:

MTF: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Methodology, Project administration, Resources, Supervision, Visualization, Writing – Original draft

FAL: Conceptualization, Data curation, Investigation, Validation, Writing – Review & editing

UW: Conceptualization, Supervision, Writing – Review & editing

Herr Patrick F. Kotzur, der geteilter Erstautor der Manuskripte # 2 und # 3 ist, bestätigt die genannten eigenständigen Beiträge (siehe nächste Seiten).

Diese Arbeit habe ich weder in Gänze noch in Teilen in gleicher noch in ähnlicher Form als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht.

Landau, der 28. Mai 2021

Friehs

Maria-Therese Friehs, M. Sc.

Manuscript # 3: Stereotype content of refugee subgroups in Germany

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

* Shared first authorship

Submitted on August 30, 2018 to the European Journal of Social Psychology

Accepted on March 11, 2019

Published online on October 29, 2019

Contributorship according to the CRediT system:

PFK: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Validation, Writing – Original draft

MTF: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – Original draft

FA: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – Review & editing

MVZ: Resources, Writing – Review & editing

I confirm the above displays my contributions accurately.

Patrick K

Patrick Kotzur, 21/05/2021

Manuscript # 2: A preregistered examination of scale properties of stereotype content measures: The German case

Friehs, M.-T.*, Kotzur, P. F.*, Zöller, A.-K. C., Wagner, U., & Asbrock, F. (2021). A preregistered examination of scale properties of stereotype content measures: The German case. *Manuscript under review.*

* Shared first authorship

Submitted on May 17, 2021 to the International Review of Social Psychology

<u>Contributorship according to the CRediT system:</u>

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

PFK: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing – Original draft

AKCZ: Data curation, Formal analysis, Methodology, Writing – Original draft

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

FA: Conceptualization, Resources, Writing – Review & editing

I confirm the above displays my contributions accurately.

Patrick K

Patrick Kotzur, 21/05/2021