Special Collection: Challenges in Translating and Adapting Psychological Measures to Spanish/Portuguese

Original Article

Comparing Social Perceptions of Culturally Emic Protagonists Using the Stereotype Content Model

A Scale Development and Adaption Process Across Four Languages and Eight Countries

Moritz Schemmerling[1], Maria-Therese Friehs[1], Patrick F. Kotzur[2], Franco Bastias[3], Jonas De Keersmaecker[4], Francisco G. Macedo[5], Felix Neto[6], Joana Neto[7], Agnieszka Pietraszkiewicz[8], Katharina Schmid[4], Sabine Sczesny[8], Claudio Torres[5], and Klaus Boehnke[9]

[1]Fakultät Psychologie, FernUniversität in Hagen, Germany
[2]Department of Psychology, Durham University, UK
[3]Cluster of Excellence "The Politics of Inequality", Universität Konstanz, Germany
[4]Department of People Management and Organization, Ramon Llull University, Esade, Spain
[5]Department of Social and Work Psychology, University of Brasilia, Brazil
[6]Faculdade de Psicologia e de Ciências da Educação, University of Porto, Portugal
[7]Research on Management and Information Technologies (REMIT), Universidade Portucalense, Porto, Portugal
[8]Institute of Psychology, University of Bern, Switzerland
[9]Bremen International Graduate School of Social Sciences (BIGSSS), Constructor University, Bremen, Germany

Abstract: Cross-cultural comparisons are often based on a single itemset that is used in several cultures and languages being translated semantically correct. In contrast, a new, emic, approach measures the same construct with individually created items for each culture and language. To test this emic approach, the current paper used the stereotype content model (SCM) with its dimensions, warmth, and competence. It is used to compare perceptions of people, residing in different countries, speaking different languages. The current paper reports a study (N = 2,901) that tests whether an adapted scale allows reliable and structurally valid measurement and comparisons of culturally emic protagonists on SCM dimensions across four languages (English, German, Portuguese, Spanish) in eight countries (United States, United Kingdom, Germany, Switzerland, Portugal, Brazil, Spain, Argentina). The warmth dimension emerges as largely universal, but the competence dimension is a more culture-specific construct. Cross-cultural comparisons as to the competence dimension should be treated with care.

Keywords: stereotype content model, measurement invariance, emic, etic, multilingual

Our modern, globalized world is constantly growing together. Countries become increasingly diverse, different cultures can be travelled to and experienced within hours, and people speak several languages to ease communication. However, a well-connected world opens up new potentials of misunderstanding. The thumbs-up gesture, for example, is known as a positive symbol in the Westerly world, whereas it will be a quite offensive gesture in parts of the Middle East. To maneuver along the world without conflict, a broad knowledge of the unique cultural specifics has to be gained. Cross-cultural research aims to find both aspects of cultural specifics over different countries, cultures and languages, similarities, and dissimilarities.

https://doi.org/10.1027/2698-1866/a000067

When researchers set up a new survey-based study, they usually seek to understand and then narrow down the research area, generate research questions and hypotheses, and finally set up the questionnaire (Bordens & Abbott, 2010). Each research process is often anchored in the regional, national, or ethno-cultural context of the researcher. Most parts of this research process are implemented using the local and/or researchers' native language in non-English-speaking countries, although the majority of scientific communication is in English. Thus, language and context need to be considered when selecting research instruments for a survey using one of the following options: There is a suitable instrument in the required language, there is a suitable instrument in a different language that requires translation, or a new instrument has to be developed (Fenn et al., 2020). Instruments frequently have to be translated and sometimes adapted to the cultural context. It is, however, rarely substantiated that the original and adapted instruments are indeed equivalent (Boehnke, 2022; Friehs, Kotzur, Böttcher, et al., 2022). "Equivalent" in the current context means that the researcher has ensured that instruments measure the same construct and that results are, thus, comparable across different languages and cultures. As equivalence is rarely corroborated and reported in full (Boer et al., 2018; Friehs, Kotzur, Böttcher, et al., 2022), we see the necessity to shed further light on this issue by attempting to assess the equivalence of an established measure of social perception: the Stereotype Content Model's *warmth* and *competence* measures (Fiske et al., 2002). As a further novelty, we present the first replication of a novel, emically rooted approach to the assessment of equivalence (Boehnke et al., 2014). To gain a comprehensive overview of the degree of equivalence, we applied the procedure originally suggested by Boehnke et al. (2014) in our eight-country study (Argentina, Brazil, Germany, Portugal, Spain, Switzerland, United Kingdom, United States) using four different languages (English, German, Portuguese, Spanish).

Theoretical Background

When people meet other people, they constantly evaluate themselves, the encountered others, the respective groups they belong to, etc. This evaluative process is continuous and guides behavior and protection by distinguishing whether people might be beneficial or harmful. When it comes to social evaluation, this basic process targets social groups rather than individuals (Abele et al., 2021). The stereotype content model (SCM; Fiske et al., 2002) structures the evaluation process by comparing the social perception of different social groups on two basic dimensions: *warmth*, which is associated with trustworthiness and friendliness, and *competence*, which is associated with capability and assertiveness (Fiske, 2018). Warmth can also be integrated into an overall model of social evaluation (Abele et al., 2021), which differentiates a horizontal dimension, i.e., the aspect of *getting along with peers* which is also associated with communion (Abele et al., 2008) and morality and sociability (Leach et al., 2016). The vertical dimension of this overall model, i.e., the aspect of *ascending in the status hierarchy* (Abele et al., 2021), integrates the SCM dimension competence and agency (Koch et al., 2016).

There is a still ongoing discussion about the facets of the vertical (communion; SCM-warmth) and horizontal dimension (agency; SCM-competence), which may be split into facets (Abele et al., 2016, 2021) across current models of social perception, including the SCM. Bick et al. (2022) revealed that facets within one dimension are highly intercorrelated and a deeper empirical differentiation within one dimension is hard to achieve. Considering this finding and the ongoing discussion on various facets within the SCM dimensions, we focused this paper on the two well-established overarching SCM dimensions. Thus, the main aim of this research lies in a procedure-related examination of the cross-cultural comparability and *measurement invariance* of the two SCM dimensions.

The SCM is often used in cross-cultural research and applied in many countries across the globe (Abele et al., 2021; Cuddy et al., 2007; Durante et al., 2012, 2017). But when it comes to cross-cultural or translingual comparisons, it is crucial to assure that equivalent theoretical constructs of warmth and competence are compared (Boehnke, 2022; Friehs, Kotzur, Böttcher, et al., 2022). Linguistically equal words and expressions can bear quite different connotations in different languages and cultures. The English expression of feeling *blue*, for example, indicates feelings of sadness or depression, whereas the equivalent German word *blau*, beyond designating a certain color, also refers to being drunk. In Russian, Ґолубóй can take the meaning of gay/homosexual, where in France, the color *bleu* will often be associated with the successful national soccer team.

We define the term *equivalence* as the abstract conceptual comparability across groups and cultures, which is a theoretical quality of a construct as manifested in a specific measurement instrument (Boer et al., 2018; Fontaine, 2005). In contrast, we understand the related term measurement invariance (MI) as a practical, statistical feature of an instrument defining how measurement operations under similar conditions yield measures of the same construct (Boer et al., 2018).

As for the corresponding quantitative approach to MI, the stepwise hierarchical procedure for MI testing distinguishes four levels of invariance (Brown, 2015). There are:

1. *configural* MI, testing for an equal form of the construct by applying the same factor structure of latent and manifest variables;
2. *metric* MI, testing for equal meaning of the construct by constraining the factor loadings of a given item across different groups to equality;
3. *scalar* MI, testing for equal item functioning by additionally constraining the indicator intercepts of a given item across different groups to equality; and
4. *strict factorial* MI, testing for equal reliability of items by additionally constraining the residual variances of a given item across the target groups to equality (Boer et al., 2018; Brown, 2015).

Boehnke et al. (2014) take a different view on equivalence. Items are considered as *etically* equivalent, when they have a linguistically equal wording (e.g., the items of a measurement instrument are worded as similar as possible in different languages). Linguistic equivalence does not necessarily mean that the same construct is measured in different countries, languages, and cultures (Boehnke et al., 2014). In his programmatic article, Boehnke (2022) once again proposed that to provide evidence for the cultural comparability of a psychological construct across different languages and cultures, *emic* equality should be tested. Emic equivalence can only be achieved through a bottom-up approach. Whenever a psychological construct is to be measured in different languages or cultures, items should be independently created in the respective language or culture to assure that construct and context are properly covered and then reduced by statistical analysis to ascertain that equivalence across the countries/cultures is achieved (Boehnke, 2022).

If both emic and etic equivalence are given, it can be assumed that the same construct is tested across countries and cultures. However, when emic equality is corroborated, it is not required to achieve etic equality as well, if the emically developed instruments are agreed upon by the academic community as operationalizing the same construct (Boehnke et al., 2014). However, what happens if – otherwise etically equivalent – items address different targets in different cultures? For instance, a head of state or executive leader of a state can be found in almost every country, but the role and relevance of this position and person can vary considerably between countries. When a scale and its *tried-and-tested* translation to another language are accepted as being etically equivalent, what happens if different targets are addressed in items either by a role descriptor (Head of State) or even by a name (e.g., Charles III, Luiz Inácio Lula da Silva, Vladimir Putin, or Emmanuel Macron)? In this case, emic equivalence has to be corroborated before any cross-cultural comparisons are possible.

Testing for equivalence is a crucial step when it comes to cross-cultural comparisons, but such tests are often performed insufficiently (Friehs, Kotzur, Böttcher, et al., 2022). Friehs, Kotzur, Böttcher, et al. (2022) recently found that SCM scales in German and English perform poorly in terms of MI. Without having established certain levels of MI, especially when applying an emic approach with per-culture created items, the results of the comparison should be seen critical as it is not assured that the data are actually comparable without bias.

The SCM has frequently been used in Spanish and Portuguese studies (see Cuddy et al., 2007; Durante et al., 2012, 2017). However, we found some reports regarding the general structural validity of SCM scales for Spanish-speaking countries (see López-Rodríguez et al., 2013; Sayans-Jiménez et al., 2017), but we did not find reports regarding MI for Spanish and Portuguese settings. It is yet unknown how they perform within their respective language and countries or what levels of equivalence can be established when tested between the languages and countries. We know that scales in German and English perform poorly in terms of structural validity (Friehs, Kotzur, Böttcher, et al., 2022), so we aimed to provide this kind of insight for the Ibero-Romance languages.

Furthermore, scopus.com has listed more than 400,000 peer-review articles containing the keyword *scale* that are related to English-speaking countries, whereas the numbers for Spanish and Portuguese countries are below 50,000. In contrast, there are around 475 million native Spanish speakers, around 230 million native Portuguese speakers, and 375 million native English speakers worldwide. As Spanish and Portuguese seem underrepresented in the available research literature, we put our focus on these countries. Still, we did not exclusively limit our research to Spanish and Portuguese, but also included English and German and tested the quantitative equivalence within and between the countries of concern.

In our study, we tested for MI in eight different countries (Portugal, Brazil, Spain, Argentina, Germany, Switzerland, United States, United Kingdom) and four languages (Spanish, Portuguese, English, German) to gain insight into which levels of equivalence of MI can be established for the warmth and competence scales and, as a result, to what extent etic items can establish emic comparability.

Methods

The current study was pre-registered (Schemmerling et al., 2023a) and used a dataset taken from Friehs, Kotzur, Kraus, et al. (2022), which is provided openly on OSF

(see Schemmerling et al., 2023b). This project originally collected data in 35 countries asking participants to rate 12 different persons or groups (*protagonists*) related to the country's COVID-19 pandemic on two SCM dimensions, warmth and competence. Both dimensions were presented using six items each that were originally constructed in English and then translated to one of the official languages of the countries using a parallel translation procedure (TRAPD; Harkness, 2003). The items could, thus, be considered etic *sensu* Boehnke et al. (2014), as they were translated for linguistically equal meanings. The wording of the items was quite similar, as only the respective adjective and the protagonist were changed: *As viewed by society, how* [adjective] *are the following individuals, groups, organizations and movements?* – [protagonist]. To measure the SCM dimension warmth, the adjectives (1) *good-natured*, (2) *cooperative*, (3) *likable*, (4) *honest*, (5) *trustworthy*, and (6) *well-intended* were used and respectively for the SCM dimension competence (1) *capable*, (2) *competent*, (3) *efficient*, (4) *influential*, (5) *assertive*, and (6) *persistent*. Items were rated on a 5-point scale from (1) *not at all* to (5) *extremely*. The full item set and the corresponding translations can be found at https://osf.io/pvdqh/ (Schemmerling et al., 2023b). For details on scale construction, data acquisition, and preparation, see Friehs, Kotzur, Kraus, et al. (2022).

In the current study, we used the data of eight countries that speak four different languages, with German and English as common languages for scale construction and Spanish and Portuguese, about which we had little previous knowledge and on which we will therefore focus. Data collected in Spain and Argentina were used for the Spanish language, and data of Brazil and Portugal for Portuguese. Regarding the German language, Germany, Austria, and Switzerland were suitable as countries. The Austrian dataset used the term "Querdenker" as protagonist for COVID-Protest-Movement, and the respondents raised quite a few questions on this term because it had a strong relation to Germany and was often unknown in the Austrian context. Therefore, we decided upon including only Germany and Switzerland. We selected two of five English-speaking countries, the United States and the United Kingdom, to keep the number of countries per language equal throughout the study. The SCM originated in the United States and the items of Friehs, Kotzur, Kraus et al. (2022) were originally developed in the United Kingdom.

The final size of our grand sample was N = 2,901 with Germany (n = 351), Switzerland (n = 347), United States (n = 306), United Kingdom (n = 339), Brazil (n = 665), Portugal (n = 293), Spain (n = 327), and Argentina (n = 273). The data were mainly obtained from students (for more details, see https://osf.io/pvdqh/). A power analysis using pwrSEM (Wang & Rhemtulla, 2021) with six items on one latent factor, a minimum factor loading of .40 (which was used as cutoff value), and a sample size of 273 resulted in a power of .99–1.00 for equal means and .99–1.00 for equal residual variances, and thus suggested our sample sizes per country to be sufficient.

We created a total of 19 combinations of countries for our analyses. The primary combinations were countries speaking the same language and every combination of these language pairings. Additionally, we analyzed country combinations that were based on cultural and geographical proximity.

The dataset was reduced to three protagonists (P1 = head of state, P2 = physicians, P3 = anti-COVID protest movement) because these were the only protagonists included in every country's survey. As the protagonists semantically addressed the same groups and individuals in each country, but the actual protagonists itself are likely to have had a different standing, status, and role in the country's culture and society, they could be considered as emic. Thus, we used etic items and combined them with different protagonists across different countries, turning them into emic items.

To reduce complexity and as the performance of the two subscales, warmth and competence, rather than the overall construct SCM itself, was under investigation, we conducted separate analyses per dimension. This led to a total of six separate analyses per set of countries (2 SCM dimensions × 3 protagonists). Each analysis was assigned an identifier consisting of the SCM dimension label and the protagonist label (i.e., *warmth P1*).

The analyses were conducted using SPSS 28 for data preparation, Mplus 8.5 for exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) operated by an iterator script based on R 4.0.5, including the packages MplusAutomation (Hallquist & Wiley, 2018) for Mplus execution and openxlsx (Schauberger & Walker, 2022) for documentation. We used robust full-information maximum likelihood estimators (MLR) in CFAs to estimate missing values based on the observed variance covariance matrix (Enders & Bandalos, 2001). The iterator script applied the analytical strategy as described in the following (see https://osf.io/pvdqh/ for more details).

The analytical strategy was based on Boehnke et al. (2014): Initially, an exploratory factor analysis using forced one-factor principal component analysis was conducted country-wise. The items were then sorted by factor loading with items loading lower than .40 being excluded. If this resulted in an unequal number of items over all compared countries, the number of items in every country was harmonized by removing lowest loading items even if they were loading higher than the cutoff. Afterward, the items were renamed with technical labels starting with the highest loading item (i.e., Item 1 to Item 6). Next, a confirmatory factor analysis was run separately for each

country with the minimum requirement that all items loaded on one factor with acceptable fit indices (RMSEA ≤ .08, SRMR ≤ .10, CFI ≥ .95; Schermelleh-Engel et al., 2003). After resorting and reduction of the items, the data of the countries were combined into one dataset.

On this basis, the step-up procedure for testing MI was applied (Brown, 2015; Little, 1997; Meredith, 1993). Additional to the fit indices, the constrained models had to fulfil the difference in fit criteria (ΔRMSEA < .015, ΔCFI > −.010, $SRMR_{metric}$ < .030, $SRMR_{scalar/strict}$ < .010; Chen, 2007) to be accepted. On editor's and reviewers' request, we deviated from the pre-registration in applying the criteria suggested by Chen (2007) instead of using the Satorra–Bentler scaled $\Delta\chi^2$ test (Satorra & Bentler, 2001) as a criterion, as it is quite sensitive for large samples. Applying the Chen criteria and ignoring the Satorra–Bentler criterion led to same or slightly higher levels of MI compared to the Satorra–Bentler criterion only.

As a first step, a baseline model was created that tested for equal form of the model in all countries of concern to establish configural MI. To test for equal form, the latent variable-indicator relation was set equal for each country with all other parameters being freely estimated. The baseline model was accepted when meeting the fit indices criteria mentioned above. If the criteria were not met, the model was rejected and no further analyses were performed with this subset.

Second, the model was tested for metric MI by setting the factor loadings equal across all countries. If the fit indices were met and the difference in fit criteria was met in relation to the configural model, full metric MI was established. Otherwise, we tested for partial MI in an iterative process using the highest modification index of the Mplus results, which was used to free the most promising parameter in one of the countries. The highest modification index was only applied if the proposed adjustment was not in conflict with the requirements for partial MI. To achieve partial MI, a minimum of two items had to remain constrained to equality across all countries and all (established) MI levels (Byrne et al., 1989). When no further modification was left and the model was still not acceptable, the analysis was stopped.

Whenever partial or full metric MI could be established, we tested for (partial) scalar MI by additionally constraining the item intercepts to equality. As acceptance criteria, we used the same model fit criteria and the difference in fit criteria referencing to the accepted (partial) metric MI model version.

As the last and highest level of MI, we tested for (partial) strict factorial MI by additionally constraining the residual variance of the items to equality and using the mentioned acceptance criteria in conjunction with the difference in fit criteria referencing to the accepted (partial) scalar MI model version.

Results

Depending on the combinations of countries, different item sets were used, as the item count was harmonized throughout the different country combinations, and some countries had less items loading ≥. 40 than others in the initial exploratory factor analyses. In general, analyses of the SCM dimension warmth used five or six items, whereas analyses of the dimension competence contained minimum three items, depending on country and protagonist. For details, see Tables 1 and 2 and the materials at https://osf.io/pvdqh/.

Reliability was calculated per country, SCM dimension, and protagonist, with the reliability concerning the warmth dimension ranging from good to excellent (ω = .81 –.95) and for competence between acceptable and excellent (ω = .70–.95; see Table 1 and materials as https://osf.io/pvdqh/).

Analysis of Measurement Invariance

After reduction and reorganization of the data country by country, we conducted the analyses for several country combinations. For reasons of brevity, we only report the highest corroborated level of measurement invariance, starting with the basic language pairings and then focusing on combinations with *Portuguese* and *Spanish* scale versions, and also omit exact fit indices. For further results and information, see materials at https://osf.io/pvdqh/ and Table 2.

Countries With English and German Language

The items had originally been designed in English and were later translated. Thus, we started the analyses with English-speaking countries (United States and United Kingdom). Partial strict factorial MI was shown for warmth P1–P3 and competence P1. Competence P3 reached partial metric MI, whereas competence P2 failed in establishing configural MI.

Analyses for the German country pairing (Germany and Switzerland) revealed similar findings. Partial strict factorial MI could be established for all warmth scales, P1–P3. Regarding competence, partial strict factorial MI was only found for competence P3, whereas configural MI was rejected for competence P1 and P2.

Taken together, the SCM dimension warmth reached the highest levels of partial strict MI for all protagonists in German- and English-speaking countries and can be seen as equivalent regarding conceptual, mean-level, and reliability comparisons.

Table 1. McDonald's ω

SCM dimension	Protagonist	ARG	BRA	GER	POR	ESP	SUI	United Kingdom	United States
6 items									
Warmth	P1	.919	.944	.921	.906	.929	.916	.898	.945
	P2	.860	n.a.	.841	n.a.	.889	.918	.887	.861
	P3	.916	.902	.876	.813	.879	.911	.822	.943
Competence	P1	n.a.	.929	n.a.	.871	n.a.	n.a.	n.a.	n.a.
	P2	n.a.	.880	.809	.859	n.a.	.847	.840	.798
5 items									
Warmth	P2	.844	.874	.832	.887	.883	.911	.877	.844
Competence	P1	.838	.934	.863	.880	n.a.	.846	n.a.	.910
	P2	.818	.872	.808	.879	.751	.840	.828	.787
	P3	.870	.873	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
4 items									
Competence	P1	.888	.946	.866	.897	.880	.842	.848	.914
	P3	.878	n.a.	n.a.	n.a.	.737	n.a.	.695	.855
3 items									
Competence	P3	.857	.867	.796	.825	.803	0.784	.739	.852

Note. Values marked with n.a. were not calculated as this item combination was not used in any of the analyses. Countries are Argentina (ARG), Brazil (BRA), Germany (GER), Portugal (POR), Spain (ESP), Switzerland (SUI), United Kingdom, and United States. Protagonists are head of state (P1), physicians (P2), and anti-COVID protest movement (P3).

Ibero-Romance Language Pairings

In the analyses of the Spanish-speaking countries Spain and Argentina, partial strict factorial MI was found for warmth P1 and P3 and competence P3. Partial scalar MI was found for warmth P2 and competence P2. No level of MI was established for competence P1.

In the analyses for Portugal and Brazil, partial scalar MI was shown for warmth P2 and competence P2. Configural MI was established for competence P3 but failed for warmth P1 and P3 as well as for competence P1.

Spanish- and Portuguese-Speaking Countries

After combining the data for Portuguese- and Spanish-speaking countries, partial strict factorial MI could be shown for warmth P1 and P3 as well as for competence P3. Warmth P2 and competence P2 reached partial metric MI. Competence P1 could not reach configural MI.

English-, Spanish-, and Portuguese-Speaking Countries

In the next step, we included English, the language most research is based on, published in, and used as source language for the translation of scales, as a third language to Spanish and Portuguese. With this combination of countries, partial strict MI was established for warmth P1, whereas warmth P2 and P3 reached partial scalar MI. Partial metric MI was established for competence P2 and full configural MI for competence P3. Competence P1 failed to achieve configural MI.

Analysis of the Full Set of Countries

The analysis of all countries established partial strict factorial MI for warmth P1, partial scalar MI for warmth P3, and partial metric MI for warmth P2. Competence P3 reached configural MI, but no level of MI could be established for competence P1 or P2. With partial metric MI being established for warmth P1–P3, it can be assumed that the meaning of the warmth concept could be seen as equivalent over eight countries speaking four languages, but mean-value comparisons would be biased due to the lack of (partial) scalar MI.

Analysis of Geographically Created Country Combinations

Besides the language-based combinations, we tested some geographically or culturally anchored combinations. The Latin-American combination Brazil–Argentina reached partial strict factorial MI for all warmth and competence sets. Adding the United States to these countries (thereby

Table 2. Measurement invariance levels for different country combinations

					Measurement invariance			
Pairing/language	Countries	Label	Protagonist	Items	Configural	Metric	Scalar	Strict
English	United Kingdom, United States	Warmth	P1	6	Full	Partial (5)	Partial (5)	Partial (5)
			P2	6	Full	Full	Partial (4)	Partial (3)
			P3	6	Full	Full	Partial (5)	Partial (4)
		Competence	P1	4	Full	Partial (3)	Partial (3)	Partial (3)
			P2	6	n.a.			
			P3	3	Full	Partial (2)		
German	GER, SUI	Warmth	P1	6	Full	Partial (5)	Partial (4)	Partial (4)
			P2	6	Full	Full	Partial (4)	Partial (3)
			P3	6	Full	Full	Partial (2)	Partial (2)
		Competence	P1	5	n.a.			
			P2	6	n.a.			
			P3	3	Full			
Spanish	SPA, ARG	Warmth	P1	6	Full	Partial (5)	Partial (2)	Partial (2)
			P2	6	Full	Full	Partial (5)	
			P3	6	Full	Partial (5)	Partial (4)	Partial (4)
		Competence	P1	4	n.a.			
			P2	5	Full	Full	Partial (2)	
			P3	4	Full	Full	Partial (3)	Partial (3)
Portuguese	POR, BRA	Warmth	P1	6	n.a.			
			P2	5	Full	Full	Partial (2)	
			P3	6	n.a.			
		Competence	P1	6	n.a.			
			P2	6	Full	Partial (5)	Partial (2)	
			P3	3	Full			

(Continued on next page)

Table 2. (Continued)

					Measurement invariance			
Pairing/language	Countries	Label	Protagonist	Items	Configural	Metric	Scalar	Strict
Spanish–Portuguese	SPA, ARG, POR, BRA	Warmth	P1	6	Full	Partial (5)	Partial (2)	Partial (2)
			P2	5	Full	Partial (4)		
			P3	6	Full	Partial (4)	Partial (2)	Partial (2)
		Competence	P1	4	n.a.			
			P2	5	Full	Partial (4)		
			P3	3	Full	Partial (2)	Partial (2)	Partial (2)
English–Spanish–Portuguese	United Kingdom, United States, SPA, ARG, POR, BRA	Warmth	P1	6	Full	Partial (5)	Partial (2)	Partial (2)
			P2	5	Full	Partial (4)	Partial (2)	
			P3	6	Full	Partial (4)	Partial (2)	
		Competence	P1	4	n.a.			
			P2	5	Full	Partial (2)		
			P3	3	Full			
German–English–Spanish–Portuguese	GER, SUI, United Kingdom, United States, SPA, ARG, POR, BRA	Warmth	P1	6	Full	Partial (4)	Partial (2)	Partial (2)
			P2	5	Full	Partial (4)		
			P3	6	Full	Partial (4)	Partial (2)	
		Competence	P1	4	n.a.			
			P2	5	n.a.			
			P3	3	Full			
Latin-America	ARG, BRA	Warmth	P1	6	Full	Partial (5)	Partial (3)	Partial (3)
			P2	5	Full	Partial (4)	Partial (2)	Partial (2)
			P3	6	Full	Full	Partial (5)	Partial (5)
		Competence	P1	5	Full	Full	Full	Partial (4)
			P2	5	Full	Full	Partial (4)	Partial (3)
			P3	5	Full	Full	Partial (4)	Partial (4)

(Continued on next page)

Table 2. (Continued)

					Measurement invariance			
Pairing/language	Countries	Label	Protagonist	Items	Configural	Metric	Scalar	Strict
Americas	United States, ARG, BRA	Warmth	P1	6	Full	Partial (5)	Partial (2)	Partial (2)
			P2	5	Full	Partial (4)	Partial (2)	Partial (2)
			P3	6	Full	Partial (5)	Partial (3)	Partial (3)
		Competence	P1	5	Full	Partial (4)	Partial (3)	Partial (3)
			P2	5	Full	Partial (4)		
			P3	4	Full	Partial (3)		
Europe	GER, SUI, United Kingdom, SPA, POR	Warmth	P1	6	Full	Partial (4)	Partial (2)	Partial (2)
			P2	5	Full	Partial (4)	Partial (2)	
			P3	6	Full	Partial (4)		
		Competence	P1	4	n.a.			
			P2	5	n.a.			
			P3	3	Full			
Western-Europe	SPA, POR	Warmth	P1	6				
			P2	5	Full	Partial (4)	Partial (4)	Partial (3)
			P3	6	n.a.			
		Competence	P1	4	n.a.			
			P2	5	n.a.			
			P3	3	Full	Full	Full	

Note. This table only contains an excerpt of the results, and the complete table can be found in the ESM. "Full" stands for full MI in this level. "Partial(x)" means partial MI on the respective level with "x" items being invariant through all levels up to the level of concern. "n.a." indicates that no configural MI could be established and the analysis was terminated. Countries are Argentina (ARG), Brazil (BRA), Germany (GER), Portugal (POR), Spain (ESP), Switzerland (SUI), United Kingdom, and United States. Protagonists are head of state (P1), physicians (P2), and anti-COVID protest movement (P3). Configural MI at competence P3 was often just identified (3 items); we accepted the "perfect" result on the configural MI level as no qualified decision could be made on basis of the fit indices, and we relied on the results starting with the metric MI level.

looking at all included countries from the Americas), partial strict factorial MI was shown for warmth P1–P3 and competence P1. Competence P2 and P3 reached partial metric MI.

Moving across the Atlantic, the pairing of the European countries (Spain, Portugal, Germany, Switzerland, United Kingdom) revealed partial strict MI for warmth P1, partial scalar MI for warmth P2, and partial metric MI for warmth P3. Configural MI could be established for competence P3, but not for competence P1 and P2.

In contrast to the Latin-American pairing, the analysis of the Western-European countries located on the Iberian Peninsula, Spain, and Portugal showed partial strict factorial MI for warmth P2 and full scalar MI for competence P3. We were not able to establish configural MI for warmth P1 and P3, nor for competence P1 and P2.

Discussion

The SCM has been well-researched in a range of different languages and countries and widely applied in cross-cultural studies (Abele et al., 2021; Cuddy et al., 2007; Durante et al., 2012, 2017). A recent study showed that German and English SCM data perform poorly in terms of structural validity, and measurement invariance (MI) testing is rarely done or reported (Boer et al., 2018; Friehs, Kotzur, Böttcher, et al., 2022). When adapting scales to different languages and cultures, the main focus typically lies on semantic equivalence and less on the performance of structural validity and measurement invariance that validates the comparability of the culturally adapted versions of scales. Boehnke (2022) showed an alternative approach to cross-cultural adaption of scales with an emphasis on data-driven validation of equivalence by MI testing and careful cultural adjustment (emic approach), which puts less focus on linguistic equivalence and parallelism of (back)translated item sets (etic approach).

We followed the new procedure as suggested by Boehnke (2022). Our design included eight countries speaking four languages, which allowed us to create a great variety of combinations. The overarching goal was to examine the performance in terms of statistical validity and comparability of the SCM dimensions warmth and competence in different countries and languages with a special focus on Portuguese and Spanish. Our design enabled a wide range of combinations to test for equivalence between countries. We analyzed 19 different combinations: 14 based on languages and five based on geographical aspects. With six MI testing procedures per country combination (2 SCM dimensions × 3 protagonists), we ran a total of 114 full analytical proceedings.

We were able to establish partial strict factorial MI for the SCM dimension warmth in the language pairings of Spanish, English, and German, with the exception of warmth P2 in the Spanish combination that only reached partial scalar MI. That means that warmth can be considered sufficiently equivalent in terms of meaning of the latent factor (metric MI), item functioning (scalar MI), and reliability (strict factorial MI) within each of the language combination, and the respective scale can be considered emic. These results were not supported by the Portuguese language pairing, which only reached partial scalar MI in one of three warmth scales and failed to establish configural MI in the other two. Combining Spanish- and Portuguese-speaking countries and extending this set of countries with English- and/or German-speaking countries, we were mainly able to establish partial metric MI for all three protagonists and partial scalar MI for the warmth scales of P1 and P3 or partial strict factorial MI for warmth P1. With metric MI standing for equal factor loadings, this result indicates that the meaning of the measured construct warmth is comparable over the countries, but the absolute scores cannot be compared.

Leaving the logic of combining countries by language toward a geographically based selection, the combination of Brazil and Argentina resulted in partial strict factorial MI for all three protagonists in the warmth dimension. After adding the United States to this analysis, strict factorial MI was still maintained. Although three different languages were combined, the construct of warmth was highly comparable within the Latin-American and North-plus-South-American dataset in terms of meaning and mean-value comparability. In contrast to these findings, the combination of Spain and Portugal, which might be seen as analogous to the Latin-American one, solely reached partial strict MI for P2 and no level of MI for the other two protagonists.

Taken together, the concept of warmth can be seen as largely universal in the countries and languages of concern, but the extent and shape of the construct warmth might be culturally bound, which hinders unbiased mean-value comparisons. It is crucial to test if and to what extent the measured constructs can be considered equivalent. This requirement is obvious in multilanguage studies and implied within the process of translating scales. But it is also necessary when applying one scale to same-language groups because equivalence cannot be taken for granted as the Portuguese combination showed.

The findings for warmth were not transferable to the other SCM dimension competence, as it did not perform as well as warmth throughout the analyses. In total, 25 of 57 analyses (44%) of the competence scales failed to

achieve configural MI, which means that we were not even able to test whether the conceptualization of the construct competence can be seen as equivalent. The majority of tests that reached configural MI or higher were related to competence P3 (19 of 31; 61%). Only 42% (8 of 19) of the analyses of competence P3 revealed (partial) metric MI and 32% reached MI on scalar level or higher. This limited performance of competence P3 might be caused by the model structure that mainly consisted of three items only or the fact that competence scores for this protagonist were generally very low (for further information, see Friehs, Kotzur, Kraus, et al., 2022). It thus remains unclear whether the limited performance of competence is based on technical or conceptual causes.

Head of states (P1) and physicians (P2) are positions that are often in focus because of the results and impact of their work, which are classical indicators of competence. Nonetheless, only 12 of the 38 analyses (32%) of P1 and P2 competence reached metric MI or higher. The best overall performance of competence scales was found in the Latin-American and the North-plus-South-American combinations. According to our data, the concept of competence is context-specific and not comparable for head of states and physicians throughout the countries and languages. This might be an indication that the SCM dimension competence is a culture-specific construct rather than a universal one. During our analysis, we found some potential reasons why the competence scales might have reduced comparability:

When taking a closer look at the data preparation, some of the country-wise CFAs we conducted to prepare the data were outside the limits of our fit indices criteria, and most of them were related to the competence scales (for details, see materials at https://osf.io/pvdqh/). As we used an available dataset, we kept these scales in the analysis to be able to compare the countries. Further research with new data is required to gain insight into whether this procedure has an impact on the performance of the competence scales.

Looking at the basic model configuration, fewer items had to be removed in the warmth (only one) than in competence (up to three) dimension because of low factor loadings (≤.40). The items that were excluded in the competence scales mainly addressed the in-discussion agency facet of the SCM dimension competence (Abele et al, 2021). As our aim is methodological rather than focusing on the concept of SCM itself, further research is required to clarify that picture of the cultural universality of the SCM dimension competence.

The number of removed items per scale can also be interpreted in view of information reduction. Boehnke (2022) proposed that a pool of cultural-based emic items should be created and then reduced by data analysis to the statistically relevant items. The better performance of competence P3 than the other competence scales might be because of the statistical item reduction. Competence P3 was reduced to a minimum of three, competence P1 to a minimum of four, and competence P2 to a minimum of five items because of low factor loadings. Further research is required to clarify the influence of information reduction and MI performance.

The emphasis of our study was to shed light on the performance of Spanish and Portuguese scales and their comparability. The Spanish-speaking country pairing (Spain, Argentina) reached high MI levels, but the Portuguese one (Portugal, Brazil) did not. When recombining those four countries, the Latin American combination (Argentina, Brazil) outperformed the Western European countries (Portugal, Spain). Combining all four countries, we were mainly able to establish partial metric MI to show that the SCM dimensions have an equal meaning in the Spanish- and Portuguese-speaking countries. These results additionally indicate that the comparability of the data of Portugal is not as good as the other countries, which requires further research.

As we did not use newly constructed emic scales but used the versatile dataset provided by Friehs, Kotzur, Kraus, et al. (2022), the approach *sensu* Boehnke (2022) was not applied fully. We considered our items as emic because we referenced them to targets with culturally different meaning, but the items were originally created in the classical etic way with a focus on semantic equivalence. The step taken toward emic cross-cultural assessment is relatively small in that vignettes (i.e., protagonists) are emically modified. This may, however, pave the way toward more comprehensively emic assessment in cross-cultural psychology. Developing entire vignettes within cultures followed by securing cross-cultural comparability in a way sketched by Boehnke (2022) must still be left to future research. This approach allows taking care of cultural specificities, such as social system, social roles, local dialects, etc., but assures cross-cultural comparability by statistical operations such as MI testing. In addition to that, more research is required that compares the classical approach and the emic approach when it comes to cross-cultural comparisons.

Conclusion

We took the emic approach into the field and compared eight countries and four languages. There exist few publications on cross-cultural comparability and scale performance in terms of structural validity, but assuring MI is a critical precondition for being able to compare different groups correctly. As we showed, it is not assured that one scale can be considered equivalent if the

groups of concern are within the same language, but we also showed that cross-language and cross-country comparisons are possible. It is more obvious that testing of MI and assuring equivalence is required when a study covers multiple languages, but it is also necessary within the same language. Independently of the language, it is crucial to test for MI and equivalence of the construct, whenever research focuses on cross-cultural or cross-group comparisons, that the researcher can assure the same psychological construct is compared.

References

Abele, A. E., Cuddy, A. J. C., Judd, C. M., & Yzerbyt, V. Y. (2008). Fundamental dimensions of social judgment. *European Journal of Social Psychology*, *38*(7), 1063–1065. https://doi.org/10.1002/ejsp.574

Abele, A. E., Hauke, N., Peters, K., Louvet, E., Szymkow, A., & Duan, Y. P. (2016). Facets of the fundamental content dimensions: Agency with competence and assertiveness–communion with warmth and morality. *Frontiers in Psychology*, *7*, 1810. https://doi.org/10.3389/fpsyg.2016.01810

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

Bick, N., Froehlich, L., Friehs, M.-T., Kotzur, P. F., & Landmann, H. (2022). Social evaluation at a distance - facets of stereotype content about student groups in higher distance education. *International Review of Social Psychology*, *35*, 12. https://doi.org/10.5334/irsp.686

Boehnke, K. (2022). Let's compare apples and oranges! A plea to demystify measurement equivalence. *American Psychologist*, *77*(9), 1160–1168. https://doi.org/10.1037/amp0001080

Boehnke, K., Arnaut, C., Bremer, T., Chinyemba, R., Kiewitt, Y., Koudadjey, A. K., Mwangase, R., & Neubert, L. (2014). Toward emically informed cross-cultural comparisons. *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

Bordens, K., & Abbott, B. B. (2010). *Research design and methods: A process approach*. McGraw-Hill Education.

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

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

Durante, F., Fiske, S. T., Kervyn, N., Cuddy, A. J. C., Akande, A. D., Adetoun, B. E., Adewuyi, M. F., Tserere, M. M., Ramiah, A. A., Mastor, K. A., Barlow, F. K., Bonn, G., Tafarodi, R. W., Bosak, J., Cairns, E., Doherty, C., Capozza, D., Chandran, A., Chryssochoou, X., ..., & Storari, C. C. (2012). 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., Tablante, C. B., & Fiske, S. T. (2017). Poor but warm, rich but cold (and competent): Social classes in the Stereotype Content Model. *Journal of Social Issues*, *73*(1), 138–157. https://doi.org/10.1111/josi.12208

Enders, C., & Bandalos, D. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(3), 430–457. https://doi.org/10.1207/s15328007sem0803_5

Fenn, J., Tan, C. S., & George, S. (2020). Development, validation and translation of psychological tests. *BJPsych Advances*, *26*(5), 306–315. https://doi.org/10.1192/bja.2020.33

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., & 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

Fontaine, J. R. (2005). Equivalence. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement*(pp. 803–813), https://doi.org/10.1016/b0-12-369398-5/00116-x

Friehs, M. T., Kotzur, P. F., Böttcher, J., Zöller, A. K. C., Lüttmer, T., Wagner, U., Asbrock, F., & Van Zalk, M. H. W. (2022). Examining the structural validity of stereotype content scales – A preregistered re-analysis of published data and discussion of possible future directions. *International Review of Social Psychology*, *35*(1), 1–18. https://doi.org/10.5334/irsp.613

Friehs, M. T., Kotzur, P. F., Kraus, C., Schemmerling, M., Herzig, J. A., Stanciu, A., Dilly, S., Hellert, L., Hübner, D., Rückwardt, A., Ulizcay, V., Christ, O., Brambilla, M., De keersmaecker, J., Durante, F., Gale, J., Grigoryev, D., Igou, E. R., Javakhishvili, N., ..., & Yzerbyt, V. (2022). Warmth and competence perceptions of key protagonists are associated with containment measures during the COVID-19 pandemic: Evidence from 35 countries. *Scientific Reports*, *12*(1), Article 21277. https://doi.org/10.1038/s41598-022-25228-9

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling*, *25*(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. J. R. van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–36). Wiley.

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

Leach, C. W., Carraro, L., Garcia, R. L., & Kang, J. J. (2016). Morality stereotyping as a basis of women's in-group favoritism: An implicit approach. *Group Processes & Intergroup Relations*, *20*(2), 153–172. https://doi.org/10.1177/1368430215603462

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues.

Multivariate Behavioral Research, *32*(1), 53–76. https://doi.org/10.1207/s15327906mbr3201_3
López-Rodríguez, L., Cuadrado, I., & Navas, M. (2013). Aplicación extendida del Modelo del Contenido de los Estereotipos (MCE) hacia tres grupos de inmigrantes en España [Application of the Stereotype Content Model (SCM) to three groups of immigrants in Spain]. *Estudios De Psicologia*, *34*(2), 197–208. https://doi.org/10.1174/021093913806751375
Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/bf02294825
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
Sayans-Jiménez, P., Cuadrado, I., Rojas, A. V., & Barrada, J. R. (2017). Extracting the evaluations of stereotypes: Bi-factor model of the stereotype content structure. *Frontiers in Psychology*, *8*, Article 1692. https://doi.org/10.3389/fpsyg.2017.01692
Schauberger, P., & Walker, A. (2022). *openxlsx: Read, write and edit xlsx files*. https://ycphs.github.io/openxlsx/index.html
Schemmerling, M., Friehs, M.-T., Kotzur, P., Bastias, F., De keersmaecker, J., Macedo, F. G. L., Neto, F., Neto, J., Pietraszkiewicz, A., Schmid, K., Sczesny, S., Torres, C., & Boehnke, K. (2023a). *Measuring and comparing the social perception of culturally emic protagonists using the stereotype content model: A scale development and adaptation process across four languages and eight countries* [Preregistration]. https://doi.org/10.17605/OSF.IO/6xp9w
Schemmerling, M., Friehs, M.-T., Kotzur, P., Bastias, F., De keersmaecker, J., Macedo, F. G. L., Neto, F., Neto, J., Pietraszkiewicz, A., Schmid, K., Sczesny, S., Torres, C., & Boehnke, K. (2023b). *Measuring and comparing the social perception of culturally emic protagonists using the stereotype content model: A scale development and adaptation process across four languages and eight countries* [Materials, Data]. https://doi.org/10.17605/OSF.IO/pvdqh
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*, *8*(2), 23–74.
Wang, Y. A., & Rhemtulla, M. (2021). Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial. *Advances in Methods and Practices in Psychological Science*, *4*(1), Article 251524592091825. https://doi.org/10.1177/2515245920918253

History
Received January 14, 2023
Revision received November 15, 2023
Accepted November 15, 2023
Published online December 23, 2023

Section: Methodological Topics in Assessment

Conflict of Interest
All authors declare the absence of any conflict of interest.

Publication Ethics
Informed consent was obtained from all participants included in the study.

Authorship
Moritz Schemmerling, Maria-Therese Friehs, and Patrick Kotzur conceptualized the study and conducted it through all its steps, including the writing of the first full draft of the current paper. Franco Bastias, Jonas de Keersmaecker, Francisco G. Macedo, Felix and Joana Neto, Gandalf Nicolas, Agnieszka Pietraszkiewicz, Katharina Schmid, Sabine Sczesny, and Claudio Torres secured data collection and acted as consultants on specifics of their countries of residence. Klaus Boehnke acted as senior conceptual consultant and took responsibility for a thorough editing of the penultimate draft. All authors approved the final version of the article.

Open Science
Open Data: The authors confirm that there is sufficient information for an independent researcher to reproduce all of the reported results. The data are available at https://osf.io/pvdqh/ (Schemmerling et al., 2023b).

Open Materials: The authors confirm that there is sufficient information for an independent researcher to reproduce all of the reported methodology. The materials are available at https://osf.io/pvdqh/ (Schemmerling et al., 2023b).

Pre-registration and Analysis Plan: The study was pre-registered on 2023-01-11 and is available at https://doi.org/10.17605/OSF.IO/6XP9W (Schemmerling et al., 2023a).

ORCID
Moritz Schemmerling
https://orcid.org/0000-0001-6099-5810
Maria-Therese Friehs
https://orcid.org/0000-0002-5897-8226
Patrick F. Kotzur
https://orcid.org/0000-0002-5193-3359
Franco Bastias
https://orcid.org/0000-0002-9477-1417
Jonas De Keersmaecker
https://orcid.org/0000-0002-8062-7422
Francisco G. Macedo
https://orcid.org/0000-0002-5785-3026
Felix Neto
https://orcid.org/0000-0003-0112-880X
Agnieszka Pietraszkiewicz
https://orcid.org/0000-0003-1009-9238
Katharina Schmid
https://orcid.org/0000-0001-6018-9245
Claudio Torres
https://orcid.org/0000-0002-3727-7391
Klaus Boehnke
https://orcid.org/0000-0002-5435-4037

Klaus Boehnke
Bremen International Graduate School of Social Sciences
Constructor University
Campus Ring 1
28759 Bremen
Germany
kboehnke@constructor.university