How accurate are national stereotypes? A test of different methodological approaches

Citation for published version:
Hrebickova, M, Mottus, R, Graf, S, Jelínek, M & Realo, A 2018, 'How accurate are national stereotypes? A test of different methodological approaches', *European Journal of Personality*, vol. 32, no. 2, pp. 87-99. https://doi.org/10.1002/per.2146

Digital Object Identifier (DOI):
10.1002/per.2146

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
European Journal of Personality

How accurate are national stereotypes? A test of different methodological approaches

Martina Hřebíčková[1], René Mõttus[2,3], Sylvie Graf[1,4], Martin Jelínek[1], & Anu Realo[3,5]

[1]Institute of Psychology, Czech Academy of Sciences

[2]Department of Psychology, University of Edinburgh

[3]Department of Psychology, University of Tartu

[4]Institute of Psychology, University of Bern

[5]Department of Psychology, University of Warwick

Authors' Note: Correspondence concerning this article should be addressed to Martina Hřebíčková, Institute of Psychology, Academy of Sciences of the Czech Republic, Veveří 97, 602 00 Brno, Czech Republic. Tel: +420 532290258; Fax: +420 549244667; E-mail: **martina@psu.cas.cz**. The research was supported by the grant 17-14387S by the Czech Science Foundation awarded to the first and third author and by RVO: 68081740 of the Institute of Psychology, Czech Academy of Sciences. Anu Realo was supported by institutional research funding (IUT2-13) from the Estonian Ministry of Education and Science.

Abstract

We compared different methodological approaches in research on the accuracy of national stereotypes that use aggregated mean scores of real people's personality traits as criteria for stereotype accuracy. Our sample comprised 16,713 participants from the Central Europe and 1,090 participants from the Baltic Sea region. Participants rated Participants rated national stereotypes of their own country using the National Character Survey (NCS) and their personality traits using either the Revised NEO Personality Inventory (NEO PI-R) or the NCS. We examined the effects of different (a) methods for rating of real people (NEO PI-R vs. NCS) and national stereotypes (NCS); (b) norms for converting raw scores into *T*-scores (Russian vs. international norms); and (c) correlation techniques (intraclass correlations vs. Pearson correlations vs. rank-order correlations) on the resulting agreement between the ratings of national stereotypes and real people. We showed that the accuracy of national stereotypes depended on the employed methodology. The accuracy was the highest when ratings of real people and national stereotypes were made using the same method and when rank order correlations were used to estimate the agreement between national stereotypes and personality profiles of real people. We propose a new statistical procedure for determining national stereotype accuracy that overcomes limitations of past studies. We provide methodological recommendations applicable to a wider range of cross national stereotype accuracy studies.

Keywords: national stereotypes, stereotype accuracy, intraclass correlations, rank order correlations

How accurate are national stereotypes? A test of different methodological approaches

Stereotypes can be defined as characteristics associated with certain social categories such as age, gender, or ethnicity. Stereotypes based on gender, age, religion or nationality are often linked to prejudice, social problems and intergroup conflicts. Consequently, social psychologists have approached stereotypes as a kind of social toxins: for decades, stereotypes in social psychology were predominantly seen as something inaccurate although empirical data was long lacking (Allport, 1954; Campbell, 1967; Mackie, 1973). However, the situation has since changed considerably and stereotype accuracy is one of the most noticeable and replicable effects in social psychology (for reviews see Jussim, Cain, Crawford, Harber, & Cohen, 2009; Jussim et al., 2016).

As compared to social psychologists, the interest of personality psychologists in stereotype accuracy is relatively recent. The surge of interest has been enhanced by a wide-spread adoption of the NEO Personality Inventories that measure five broad personality dimensions (Five-Factor Model, FFM; McCrae & Allik, 2002; McCrae at al., 2005; Terracciano et al., 2005). The NEO Personality Inventories (NEO-PI R; Costa & McCrae, 1992) have provided a common, multi-trait framework for collecting information on both real people and stereotypes. The correlation approach then allows for a comparison of stereotypical traits with personality traits of real people. The resulting agreement determines how well beliefs about groups' characteristics correspond to what members of these groups are actually like. Although stereotypes can encompass various characteristics, in personality psychology research stereotypes are operationalized as *personality traits* of typical representatives of a given group. Consequently, studies examine to what extent group stereotypes (e.g., typical personality traits of adult German men) correspond to chosen criteria (e.g., self-reports or observer ratings of personality traits of German adult men).

Previous research in personality psychology has shown that age and gender stereotypes are highly accurate across different cultures (Chan et al., 2012; Löckenhoff et al., 2014). For instance, women score consistently higher on Neuroticism and Agreeableness than men and indeed, women are also perceived as higher on those traits than men. In contrast, the accuracy of national stereotypes appears to be lower than the accuracy of age and gender stereotypes and consequently, some researchers have concluded that national stereotypes are inaccurate (McCrae et al., 2013; Terracciano et al., 2005). Other researchers, however, have arrived at a different conclusion showing that national stereotypes in fact did converge with the characteristics of real people, at least to a certain extent (e.g., Allik, Mõttus, & Realo, 2010; Hřebíčková & Graf, 2014; Lönnqvist, Konstabel, Lönnqvist, & Verkasalo, 2014; Realo et al., 2009).

The revived interest in research comparing stereotypes of different nations with characteristics of inhabitants of these countries was sparked by the international project Personality Profiles of Cultures that encompassed 49 cultures (PPOC; Terracciano et al., 2005) and by its subsequent criticism (Krueger & Wright, 2006). In the PPOC, personality traits typical for one's own country representatives were rated on the National Character Survey (NCS), and personality traits of real people from the given countries were rated on the Revised NEO Personality Inventory (NEO PI-R; Costa & McCrae, 1992). The ratings of real people comprised both self-reports (i.e., ratings of one's own personality characteristics) and observer-ratings (i.e., well-acquainted raters provided assessments of the target's personality). Self-reports of real people converged with national stereotypes only in two countries (Japan and Poland). Observer ratings of real people agreed with national stereotypes in four countries (Australia, Lebanon, New Zealand, and Poland). Based on the comparison of national stereotypes with both self-reports and observer-ratings of real people, the authors of the PPOC project concluded that national stereotypes were generally inaccurate (Terracciano et al., 2005). McCrae et al. (2013) presented an overview of the empirical findings concerning factors (metaphoric thinking, economic prosperity) and cognitive biases (ethnocentric bias, mirroring effect) that influence forming and accuracy of national stereotypes**.** For instance, national stereotypes are associated with climatic temperature, which appears to have an effect of metaphoric thinking (McCrae et al., 2007; Zhong & Leonardelli, 2008). People from warmer climates are evaluated as personally warmer than people from colder climates. Thus metaphoric thinking can have an effect on national stereotype accuracy. Some stereotypical personality traits seem to be related to the Gross Domestic Product (McCrae at al., 2007). Therefore, people from wealthy countries are seen as more conscientious than people from poor countries, even though in reality there is no difference between real people in this particular personality trait (Hřebíčková & Graf, 2014). Other studies documented that regional (Terracciano et al., 2005) and national stereotypes (Realo et al., 2009; Hřebíčková & Graf, 2014) exaggerate real differences between people.

The difference in accuracy of national stereotypes established by previous studies can also be due to the fact that ratings of national stereotypes can be affected by cognitive biases. One example is the ethnocentric bias: a type of culture-level self-enhancement. Allik, Mõttus, and Realo (2010) reported that Russian ingroup stereotype converged with Russian´s ratings of an ideal person. Another type of cognitive bias is the contrast or mirroring effect in which ingroup stereotype is rated in contrast to stereotype of a relevant outgroup. For example, Canadians rated their national stereotype in contrast to their ratings of the US national stereotypes especially in traits of warmth and assertiveness (Terracciano et al., 2005). Estonians and Finns rated their national stereotype in contrast to their ratings of Russian national stereotype: Estonians and Finns are quiet and modest but Russians are active and talkative (Realo et al., 2009). Hřebíčková, Graf, Tegdes, & Brezina, 2017 documented the mirroring of ingroup versus outgroup stereotypes in different intergroup contexts – not only in national, but also in regional and ethnic stereotypes.

Until recently only limited attention was paid to methodological problems in research on the accuracy of national stereotypes. There is evidence that differences in instruments used for ratings of stereotypes and real people may have contributed to the apparent lack of stereotype accuracy (Realo et al., 2009). Another problem potentially contributing to the various level of accuracy established by previous studies is the use of different norms for standardization of ratings on national stereotypes and real people. To examine the impact of these methodological discrepancies, we compared and contrasted the effect of *methods* used for the rating of national stereotypes and real people and *norms* typically employed for transformation of raw scores of stereotypical and real people ratings. Additionally, we tested whether the accuracy of national stereotypes can be influenced by methods of statistical inference regarding the congruency between ratings of national stereotypes and real people.

After addressing issues complicating unequivocal conclusion about the accuracy of national stereotypes, we turn to proposing a novel methodological approach that attempts to overcome the limitations of past studies. One of the main aims of our study is to provide guidelines for future research on the accuracy of national stereotypes.

Methods for ratings of National Stereotypes and Traits of Real People

Past studies on accuracy of national stereotypes took different approached to methods employed for ratings of national stereotypes and traits of real people. Terracciano and colleagues (2005) measured national stereotypes with the National Character Survey (NCS) and traits of real people with the NEO PI-R (Costa & McCrae, 1992). The 30-item NCS parallels the 30 facets of the 240-item NEO PI-R and measure the same constructs – characteristics included in the FFM of personality. However, the NCS and NEO PI-R differ with respects to their length (i.e., 30 vs. 240 items) and with respects to items' wording (i.e., isolated adjectives or groups of adjectives vs. phrases). However, Realo et al. (2009) and Allik et al. (2010) argued that the use of different methods can undermine the convergence between stereotypical and real-people profiles. Both studies examined the agreement between ratings of real people and national stereotypes of one's country representative in Russia and six Baltic Sea countries (Finland, Estonia, Latvia, Lithuania, Poland, and Belarus). They employed the NCS for ratings of both stereotypes and characteristics of real people and showed that national stereotypes were weakly or moderately related to personality characteristics of real people in four out of seven countries (Estonia, Finland, Poland, and Russia; Allik et al., 2010; Realo et al., 2009). Also Lönnqvist et al. (2014) used the same instrument for measuring both real people's personality traits and national stereotypes in Scandinavia and concluded that national stereotypes hold the kernel of truth based on their agreement with real people characteristics. Finally, Hřebíčková and Graf (2014) examined the effect of the same versus different methods employed for ratings of national stereotypes and characteristics of real people on stereotype accuracy in the Czech Republic. They showed that the Czech national stereotype agreed with self-rated personality traits of real Czechs when the same method (NCS) for both ratings was used but not when different method (NCS and NEO PI-R) were used for the two ratings. Thus, the outcomes of the studies from Europe did not support the conclusion of the PPOC project (Terracciano et al., 2005), suggesting an effect of methodological choices on the resulting level of accuracy.

In the present research, we compared accuracy of national stereotypes based on the same method for stereotypical and real people ratings with accuracy based on different methods for stereotypical and real people ratings. This way, we could establish to what extent methods employed for national stereotypes and the criterion of accuracy influence the resulting level of stereotype accuracy. Specifically, we tested two hypotheses that 1) using the same method for both ratings results in higher higher accuracy; and that 2) using different methods for the two ratings results in lower lower accuracy between stereotypical and real people profiles.

Standardization of Raw Scores into T-scores Based on Normative Samples

The level of accuracy of national stereotypes is usually established by comparing mean profiles of national stereotypes with mean profiles of characteristics of real people from the same country. However, when estimating the agreement between two profiles, a bias called “generalized other” can distort the estimated level of agreement (Cronbach, 1955). The “generalized other” describes a situation where any two profiles may be similar not only because their distinctive features are well-matched but also because they both reflect the profile of an average person – to some extent, most people are alike. For example, raters tend to endorse traits in the Neuroticism domain less than traits in the Extraversion, Openness, Agreeableness, or Conscientiousness domains, regardless of the specific targets they rate, and this results in intercorrelations among the rated profiles even when there is no congruence between the distinctive aspects of the both profiles (Allik, Mõttus & Realo, 2010). In other words, even randomly selected individual´s personality profiles tend to be similar. In order to separate distinctiveness from normativeness, profiles can be converted into standard scores such as *T*-scores [mean (*M*) = 50, standard deviation (*SD*) = 10], using mean scores and *SDs* from a reference norm sample.

In previous studies on national stereotypes accuracy, different reference norms have been employed for standardization of national stereotypes and characteristics of real people. Ratings of national stereotypes on the NCS were usually converted into *T*-scores using the grand means for stereotype ratings of approximately 4,000 raters from 49 countries in the PPOC project (Allik et al., 2010; Realo et al., 2009; Terracciano et al., 2005). Since there are no published international norms for the NEO PI-R self- and observer-ratings available, real people's characteristics were usually standardized using American norms (Costa & McCrae, 1992), whereas the NCS self-ratings were transformed into *T*-scores using the unweighted means and standard deviations from seven countries (Allik et al., 2010; Realo et al., 2009). Standardization using distinct norms could then distort the resulting level of agreement.

An alternative to the use of distinct norms in situation where normative samples do not exist for all measurement instruments and types of ratings is to choose a reference country where all ratings – on the NEO PI-R and the NCS as well as ratings of stereotypes and of real people – are available. In the Russian Character and Personality Survey (RCPS; Allik et al., 2009, 2010, 2011), approximately 11,000 Russian students rated themselves using the NCS, more than 3,600 Russian students rated a typical Russian using the NCS and more than 7,000 students rated someone they knew well using the NEO PI-R. Furthermore, the normative Russian self-reported NEO PI-R data is available in the Russian NEO PI-R professional manual (Martin et al., 2002).

In the present research, we have employed the previously used international norms for the transformation of stereotypical ratings and the US norms for the transformation of real people ratings. At the same time, we transformed both stereotypical and real people ratings along Russian norms. Consequently, we could compare the accuracy of national stereotypes based on the correspondence between stereotypical and real people ratings that were both transformed along the same norms with accuracy based on the correspondence between stereotypical and real people ratings that were transformed along different norms.

Correlational Techniques for Comparing National Stereotypes and Real People

Intraclass correlations (ICC) with the double-entry method (Griffin & Gonzales, 1995) are usually used for the comparison of profiles of national stereotypes and real people ratings (Hřebíčková & Graf, 2014; Realo et al., 2009; Terracciano et al., 2005). The double-entry ICCs are calculated similarly to Pearson's correlations, but the profiles are duplicated in the opposite order and the standard Pearson's correlation is calculated for the doubled set. As a result, ICCs are sensitive not only to shapes of profiles of the compared variables but also to differences in profile elevation and scatter. Thus, the ICCs are more conservative than Pearson's *r*s (McCrae, 2008). Recently, objections were raised against the use of ICCs in profile similarity research since they represent an omnibus index confounding elevation, scatter and shape (for more details, see Furr, 2010). The discussion about suitability of one or the other index for estimation of profile similarity has not yet reached unequivocal conclusion (McCrae et al., 2013). In the current study, we calculated both the double-entry ICCs and Pearson's correlations in order to compare the agreement between profiles of national stereotypes and characteristics of real people yielded by different correlational techniques.

Another issue related to the use of correlations for estimates of the convergence between two profiles of personality traits is inferring their statistical significance. Testing for the statistical significance of a correlation is based on the assumption that observations are independent. The facets of the NEO PI-R and NCS, however, are not independent: The facets intended to cover the same broad domain (e.g., Neuroticism) are by definition variations of some at least partly common dimension. Consequently in each profile, there are maximum five independent observations corresponding to the (at least conceptually orthogonal) dimensions of the Five-Factor Model and not 30 observations corresponding to the particular facets. Therefore, the correct degrees of freedom are not 30 - 2 (28) as implied in previous studies but unknown values in the range from 3 (i.e., 5 – 2) to 28. Based on the assumption of non-independence, non-parametric procedures can be more appropriate for estimating the statistical significance of the profile correlations.

Additionally, using such a small number of degrees of freedom (even 28) means that only moderate to strong correlations are considered statistically significant ($r > .36$ in case of $n = 30$), which may leave stereotypes without a fair opportunity to appear accurate. Moreover, a study on agreement between self-reports on NCS and NEO PI-R found that using different measures for the same type of rating yielded correlations between .19 and .74, median .55 (Konstabel, Lönnqvist, & Walkowitz, 2012). If the maximum correlation between national stereotypes and characteristics of real people is limited by the maximum agreement of an instrument if the same phenomenon is being measured, then .55 (and not 1.00) may be in fact the maximum possible correlation between the two types of ratings. These statistical considerations further complicate the interpretation of past findings on the accuracy of national stereotypes. We propose a novel way of addressing the outlined shortcomings below.

A Country Ranking-Based Approach

Potentially, the above mentioned methodological issues may be remedied by using non-parametric rank order correlations based on rankings of the considered countries in raw scores instead of scores transformed according to external norms. We suggest a novel way of estimating stereotype accuracy whereby we compare positions of the countries among themselves on stereotypical and real people ratings. This procedure means asking that if Germans score the highest and Czechs the lowest on the N1: Anxiety facet scale in terms of national stereotype, for instance, do they also rank the same in terms of the personality traits of real people? An advantage of this methodological approach is that it does not require the transformation of raw data into *T*-scores. Using ranks of countries instead of absolute scores also helps to avoid the problem of the “generalized other” and the disputed choice between *ICC*s and *Pearson’s r*s.

Moreover, in order to estimate the significance of results obtained by rank order correlations, we can compare them with correlations obtained by bootstrapping. Specifically, we propose using randomly reshuffled data, whereby the profiles of national stereotypes and real people traits arise from random groupings of people, yielding pseudo country profiles. Bootstrapping also overcomes the problem of non-independence of observations because it influences the real and the pseudo profile correlations alike. Furthermore, bootstrapping mitigates the problem of the unknown upper bounds of correlations. All in all, the suggested procedure estimates the likelihood of obtaining the observed correlations randomly, taking into account the inter-dependence of observations and instrument differences.

The Present Study

The aim of our study was to examine whether different methodological approaches to design and data analysis influence the resulting level of agreement between ratings of national stereotypes and real people. More specifically, we examined the effects of (a) methods for rating of real people (i.e., NEO PI-R vs. NCS); (b) norms for converting raw scores into *T*-scores (i.e., Russian vs. international and US norms); and (c) correlation techniques (i.e., intraclass correlations vs. Pearson's correlation vs. Spearman's rank-order correlations) on agreement between profiles of national stereotypes and real people's personality traits.

We have partly reanalyzed data published in Hřebíčková and Graf (2014) and Realo et al. (2009). In Central Europe, we employed the existing ratings of national stereotypes from Austria (AU), the Czech Republic (CZ), Germany (GE), Poland (PL), and Slovakia (SK); the existing ratings of real people on both NCS and NEO-PI-R from the Czech Republic and the existing ratings of real people on NEO-PI-R in the four other Central European countries. We additionally collected self-reports on the NCS in the four remaining Central European countries in order to carry out a full-range test of the methodological issues at stake. In the Baltic Sea region, we have used existing self-reports and stereotypical ratings on NCS from Belarus (BLR), Estonia (EST), Finland (FIN), Latvia (LAT), and Lithuania (LIT).[1] Although we use already published data, we employ a novel analytical approach. Realo et al. (2009) and Hřebíčková & Graf (2014) estimated stereotype accuracy using international norms and ICC

[1] We are referring to the two projects based on the geographical regions where most of the included countries belong. We are aware of the fact that Belarus included in the Baltic Sea Project does not belong to the Baltic Sea Region.

correlations. In the present research, we compare the previous findings with accuracy based on transformation of both stereotypical and real people ratings along Russian norms. Furthermore, we estimate the convergence between stereotypical and real people ratings using not only the ICC but also Pearson´s and Spearman´s correlations. Based on a thorough analysis of all relevant methodological issues that can influence the accuracy of national stereotypes, we ultimately aim at suggesting optimal methodological choices for future research on stereotype accuracy.

Method

Participants

Ratings of national stereotypes. 2,241 university students (75% women)[2] from five countries in Central Europe (Austria, the Czech Republic, Germany, Poland, and Slovakia) and 1,090 university students (68% women) from five countries in the Baltic Sea region (Belarus, Estonia, Finland, Latvia, and Lithuania) described a 'typical' member of their own nation using the National Character Survey (NCS). The *Austrian* sample consisted of 396 students, age range 18 – 65 years (*M* = 25.02, *SD* = 7.20; 75% women). The *Czech* sample consisted of 726 students, age range 18 – 54 years (*M* = 23.16, *SD* = 4.98; 75% women), the *German* sample consisted of 329 students, age range 18 – 63 years (*M* = 23.73, *SD* = 5.00; 70% women). The *Polish* sample consisted of 281 students, age range 17 – 53 years (*M* = 22.7, *SD* = 3.64; 86% women). The *Slovak* sample consisted of 509 university students, age range 16 – 66 years (*M* = 24.39, *SD* = 6.58; 76% women). The *Belarusian* sample consisted of 200 students, age range 18 – 26 years (*M* = 20.9 years, *SD* = 1.5, 50 % women). The *Estonian* sample consisted of 201 university students, age range 18 – 35 years (*M* = 21.24 years, *SD* = 2.29, 64% women). The *Finnish* sample consisted of 286 social science students, age range 19 – 55 years (*M* = 24.44 years, *SD* = 5.80, 86 % women). The *Latvian* sample

[2] Our previous analysis documented that women and men rate national stereotypes in a similar way. High level of congruence between women and men were found in rating of a typical German (ICC = .99), Czech (ICC = .95), Austrian (ICC = .91), Slovak (ICC = .88) and Pole (ICC = .84; Hřebíčková & Kouřilová, 2012).

consisted of 200 university students, age range 19 – 42 years (*M* = 22.14, *SD* = 3.03, 79 % women). The *Lithuanian* sample consisted of 203 students, age range 19 – 24 years (*M* = 20.57, *SD* = 0.91, 53% women).

Ratings of real people. In Central Europe, different participants provided real people ratings either on NEO PI-R or on NCS. The NEO PI-R data of real people´s personality traits were taken from the different versions of NEO PI-R professional manuals in the countries under study or from previously published materials. For NEO PI-R ratings of real people living in *Austria*, data were taken from a publication by McCrae (2002, p. 120–125), including self-reports by 444 college-age and adult people (66% women), further details on age of participants were not given. Self-reports on NCS were provided by 119 Austrian participants, age range 20 –72 years (*M* = 34.75, *SD* = 12.72; 79% women). For ratings of real people living in the *Czech Republic*, self-reports on NEO PI-R were taken from the Czech NEO PI-R professional manual (Hřebíčková, 2004), including self-reports by 2,288 people, age range 14 – 83 years (*M* = 26.04, *SD* = 12.21, 55% women). Self-reports on NCS were provided by 938 Czech university students, age range 19 – 70 years (*M* = 26.4, *SD* = 9.38, 76% women). For ratings of real people living in *Germany*, data were taken from the German NEO PI-R professional manual (Ostendorf & Angleitner, 2004), including self-reports by 11,724 people, age range 16 – 91 years (*M* = 29.92, *SD* = 12.08, 64% women). Self-reports on NCS were provided by 230 German participants, age range 18 – 57 years (*M* = 23.87, *SD* = 5.25, 81% women). For ratings of real people living in *Poland*, data were taken from the Polish NEO PI-R professional manual (Siuta, 2007; p. 85), including self-reports by 324 people, age range 30 – 79 years (a mean age and standard deviation were not specified in the manual, 57% women). Self-reports on NCS were provided by 180 Polish participants, age range 16 – 53 years (*M* = 26.02, *SD* = 7.70, 86% women). For ratings of real people living in *Slovakia*, self-reports were not available. Thus, we used observer-ratings on NEO PI-R provided by 238 Slovak university students, age range 18 – 23 (*M* = 20.16, *SD* = 1.40, 50% women).[3] Participants rated 240 targets (a mean age and standard deviation were not specified in the available materials, 50% women). Self-reports on NCS were provided by 228 Slovak participants, age range 19 – 32 years (*M* = 26.02, *SD* = 7.70, 86% women). Self-reports on NCS in the five *Baltic Sea countries* (Realo et al., 2009) were provided by the same participants who provided the ratings of national stereotypes (see above). Thus the social projection (Robbins & Kruger, 2005) alternatively termed false consensus effect (Ross, Greene, & House, 1997) can affect agreement between stereotype rating and self-report in Baltic Sea counties.

Materials and Procedure

The National Character Survey (NCS; Terracciano et al., 2005) was used for ratings of both national stereotypes and self-reports. The NCS consists of 30 bipolar items intended to parallel the facets of the NEO PI-R (Costa & McCrae, 1992). For example, a facet of Neuroticism, N3: Depression, is assessed by asking how likely, on a five-point scale, a typical country representative is depressed, sad and pessimistic vs. content and optimistic. Instructions and questionnaires were administered in corresponding languages. In the five Central European countries and in Finland, the questionnaires were administered online. In Central Europe, respondents first rated a typical member of their own nation and subsequently typical country representatives of four neighboring countries presented in a random order. In the Baltic Sea region, respondents first provided ratings of a typical member of their own nation, followed by ratings of a typical Russian, and finally rated their own personality, using the NCS. The NCS was used for self-ratings in all ten European countries.

[3] We are grateful to Emília Ficková from the Institute of Experimental Psychology, Slovak Academy of Sciences, for providing the raw scores and SDs on 30 facets and 5 domains of NEO PI-R observer-ratings from Slovak participants. A part of the Slovak data was already published in study by McCrae and Terracciano (2005).

The other measure used for self-ratings, the NEO PI-R (Costa & McCrae, 2002) is a 240-item measure based on the Five-Factor model. It contains 30 facets, 6 for each of the basic personality factors: Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness. Each of the 30 facets comprises eight-items. Responses are given on a five-point Likert-type scale from strongly disagree to strongly agree. The observer-rating form of the NEO PI-R (Form R) with items rephrased in the third person was used in the Slovak sample.[4]

Metric equivalence. While in previous studies (Realo et al., 2009; Hřebíčková & Graf, 2014) the metric equivalence of the NCS was formally tested on the pooled sample of all participants from the Baltic Sea and Central European countries, in the present study we tested the metric equivalence within countries.[5] We conducted 20 principal component analyses followed by a varimax normalized rotation on the 30 NCS items in stereotype ratings and in self-reports separately in ten countries. In order to examine how well the NCS factor solution in each country in both conditions (stereotype rating, self-report) replicated the NCS structure that was found in previous research (Terracciano et al., 2005), the Procrustes targeted rotations were carried out and an index of factor and total congruence coefficients (TCC) were computed in each country. More specifically, the varimax normalized factor loadings were targeted towards the factor structure of the NCS national stereotype ratings

[4] In the Slovak subsample where self-reports on the NEO PI-R were not available, we employed observer ratings on the NEO PI-R as the accuracy criterion. Since in the Czech and German subsamples self-reports and observer- ratings on the NEO PI-R average correlations were *r* = .81 and .62, respectively (Hřebíčková & Graf, 2014), we hypothesize that observer ratings could substitute the unavailable self-reports on the NEO PI-R.

[5] In our study we dealt only with metric equivalence. As for scalar invariance in personality inventories it is rarely tested. Zecca et al., (2013) observed that in nine French-speaking African countries and Switzerland, the NEO-PI R reached metric but not scalar invariance. The same was found in three culturally different countries, the USA, Mexico, and the Philippines (Church, Alvarez, Mai, French, Katibak & Oritz, 2011). Rossier with coauthors (2016) tested cross-cultural generalizability of the alternative Five-factor model using the Zuckerman-Kuhlman-Aluja Personality Questionnaire (ZKA-PQ) in 23 cultures and came to the same conclusion. ZKA-PQ has metric but not scalar equivalence across cultures. Thus the criticism regarding measurement invariance is valid and caution is needed in drawing conclusions regarding cross-cultural comparison of personality profiles.

obtained from the international sample of 3,989 respondents from 49 different nations (Terracciano et al., 2005). After the target rotations were carried out, the factor congruence coefficients for the five factors and the total congruence coefficients with the international data was estimated for 10 researched countries in two conditions (stereotype rating, self-report).

According to Lorenzo-Seva and ten Berge (2006) factors can be considered to replicate when congruence coefficients exceed .85. Among 100 factor congruence coefficients, 20 were below this criterion, nine of them were in the Openness to Experience factor. But even these 20 lower coefficients of congruence do not suggest total randomness as the means of the distribution of factor congruencies based on Procrustes rotations of randomly permuted data ranged from .32 to .34 for the five factors of personality (McCrae, Zonderman, Costa, Bond, & Paunonen, 1996). Among the 20 TCCs, German stereotype rating TCC (.84) and both Belarusian stereotype rating (.73) and Belarusian self-report TCCs (.82) did not reach .85. These results indicate that the structures of the NCS for stereotype ratings and self-reports replicated moderately well across the examined countries.[6]

Data Analysis

In the five Central European countries, where we had self-reports of personality both using the NCS and the NEO PI-R at our disposal, we tested the accuracy of national stereotypes against two accuracy criteria: (a) characteristics of real people rated on the NCS and (b) characteristics of real people rated on the NEO PI-R. The two types of ratings (real people and national stereotypes) on two measures (NEO PI-R and NCS) were provided by three *independent* groups of participants – a group that provided self-reports on the NEO PI-R, another group that provided self-reports on the NCS; and yet another group that rated national stereotypes using the NCS. In the five countries from the Baltic Sea region the *same*

[6] The factor and total congruencies are available at osf.io/2pgtj Factor and total congruence of the NCS structures.docx.

participants provided ratings of national stereotypes and self-reports always using the NCS. Due to these methodological differences we analyzed data from the two studies separately.

We were interested in agreement across the whole profiles of personality traits, not in agreement on particular traits. In order to eliminate the tendency called the "generalized other", all NCS stereotype scores were converted into *T*-scores (*M* = 50, *SD* = 10) using mean scores and *SDs* from two different normative samples – Russian and international or American norms. Russian norms were based on ratings of Russian national stereotypes from 3,296 Russian respondents (Allik et al., 2010). International norms were based on national stereotype ratings from 3,989 people in 49 different cultures (Terracciano et al., 2005). Self-reports on NCS were standardized using ratings from 10,308 Russian respondents (Allik et al., 2010) and using international norms based on 1,448 individuals from Finland, Estonia, Latvia, Lithuania, Poland and Belarus (Realo et al., 2009). Self-reports on the NEO PI-R were standardized using Russian norms from 1,080 participants (Martin et al., 2002) and using the American norms for the NEO PI-R self-reports (in Austria, the Czech Republic, Germany, and Poland) and observer-ratings (in Slovakia) taken from the professional manual (Costa & McCrae, 2002).[7] Observer-ratings from Slovakia on NEO PI-R were standardized using observer-ratings from 7,065 Russians (Allik et al., 2009) and 143 Americans (Costa & McCrae, 2002).

The profile agreement was calculated using intraclass correlation (ICC) across the 30 facets, using the double-entry method (Griffin & Gonzales, 1995) with the *p*-value based on the non-doubled *n* of 30 and Pearson's *r*.[8] We also used alternative to ICC and Pearson's *r*, the Spearman rank order correlations. To estimate Spearman rank order correlations, we calculated mean scores for 30 facets of self-ratings and ratings of national stereotypes

[7] For more details about internal consistency, factor structures and interjudge reliability of the NCS stereotypes and self-reports see Hřebíčková and Graf (2014) and Realo et al. (2009).

[8] The T scores of the 30 NCS and 30 NEO PI-R facets for ratings of national stereotypes and real people in 10 countries that were established using two different normative samples (international US and Russian norms) are available at osf.io/2pgtj Profile agreement_data_10_countries_two_norms.xls.

separately for the Central European and the Baltic Sea countries. Ranking of stereotypical and real people profiles across all 10 countries was not possible due to distinct methodology of data sampling: in Central Europe, different participants provided ratings of national stereotypes and real people in Central Europe whereas it was the same participants in the Baltic Sea region.

Next, we transformed the means on the 30 facets of self- and stereotypical ratings into ranks: The country with the lowest value on a given facet was assigned 1 and the country with the highest value on a given facet was assigned 5. Within each country, we computed Spearman rank order correlation between the rankings based on self-reports on one side and the rankings based on stereotype ratings on the other, arriving at a country-specific coefficient of agreement between self- and stereotype ratings.[9]

In order to estimate the significance of the rank order correlations, we used bootstrapping. Because of unbalanced sample sizes in the five Central European countries, we adjusted the procedure to an equal number of participants in each country. First, we randomly chose 119 participants from each country that corresponded to the size of the smallest subsample of self-ratings from Austria. Second, we randomly re-shuffled participants' nationality in both self- and stereotypical ratings. Third, we computed mean scores for each group. Fourth, mean scores were transformed into ranks and Spearman rank order correlations were computed between the self- and stereotypical ratings—exactly as we had done with non-shuffled data. These four steps were repeated 10,000 times. Based on the repeated pseudo estimates of stereotype accuracy, we obtained random variation intervals (RVI) specific for each country. Using the RVI, we evaluated the significance of agreement between self- and stereotypical ratings in the original data set. Specifically, if the Spearman correlation based on

[9] The data and R scripts for the rank order correlations are available at osf.io/2pgtj. The Central European data containing ratings of national stereotypes: Central_Europe_NCS_stereotypes_data.xls; and self-reports on NCS: Central_Europe_NCS_self_data.xls; R script: Code_Central_eur.R. The data containing ratings of national stereotypes and self-reports on NCS from Baltic Sea countries: Baltic_Sea_data.xls; R script: Code_baltic.R.

non-shuffled data did not fall between the 2.5% and 97.5% quantiles of the respective correlations based on shuffled data (i.e., RVI), then it was considered significant.

In the sample of five countries from the Baltic Sea region, the coefficients of agreement between self- and stereotypical ratings were determined the same way as in the Central European sample. Unlike the Central European sample, the sample sizes in the five countries of the Baltic Sea region were balanced and thus did not require the adjustment of the procedure for an equal number of participants like in the Central European data.

Results

Variability Indices in NCS Stereotype Ratings and Self-reports across Countries

To indicate the variability across countries in the rating of stereotypes and self-reports on the NCS items we performed an ANOVA using unstandardized scores with countries as independent variables for each 30 NCS items in the two different cases (stereotype rating, self-report) in Central European and Baltic See countries. To estimate the effect size, we used the partial Eta squared (η^2), computed as the ratio between the effect variance (variance between countries) and the sum of the effect and error variance (variance within countries). The proportion of variance that can be attributed to the effect of country in stereotype rating in Central European countries varied from 24.4% (C1: Competence) to 1.1% (N6: Vulnerability) with a median value of 8.1% and in Baltic See countries from 27.9% (E1: Warmth) to 2.1% (C1: Competence) with a median value of 12.3%. In the case of self-report the proportion of variance that is attributable to the effect of country in Central European countries varied from 11.6% (A4: Compliance) to 0.1% (C4: Achievement Striving) with a median value of 1,5% and in Baltic See countries the effect of country in self-report varied from 8.0% (E5: Excitement Seeking) to 0.9% (C1: Competence) with median value of 2.7%. Based on these results, it can be concluded that when national stereotypes are rated cross-country variance is greater than when self-reports are provided, suggesting that stereotypes rating may be sensitive to respondents' cultural background.[10]

Methodological Considerations of National Stereotype Accuracy

Next, we estimated stereotype accuracy in the five Central European countries using two accuracy criteria – ratings of real people on the NEO PI-R and the NCS. Thus, stereotype accuracy yielded two values in each of the five Central European countries (see the top of Table 1). The bottom of Table 1[11] shows stereotype accuracy in the five Baltic Sea countries with one accuracy criterion – ratings of real people on the NCS. In the columns of Table 1, stereotype accuracy based on two ways of transformation into T-scores (i.e., Russian, international, and American norms) and three correlation techniques (ICC, Pearson's r, and Spearman's rho on country-rankings) is plotted.[12]

Insert Table 1 about here

Methods for rating real people. In the five Central European countries, we could contrast stereotype accuracy based on the agreement between ratings of national stereotypes on the NCS and ratings of real people on the same (NCS) or different measure (NEO PI-R). The median of *NEO-based* correlations aggregated across ICC and Pearson correlations was -.09 and the median of *NCS-based* correlations was .24. Thus, employing the same

[10] More detailed information (such as the descriptive statistic and Partial Eta Squared on 30 NCS items are at osf.io/2pgtj Descriptive statistic and Partial Eta Squared on 30 NCS items.xlsx).

[11] All confidence intervals are to be found at osf.io/2pgtj (Table1_with_CI.docx).

[12] We tested the likelihood of replicating similar patterns of correlations (Sherman & Wood, 2014). The replicability indexes for correlations shown in Table 1 were counted using the multicon package in R for different methods and norms. The replicability indexes are as follows (the first value is based on international, US norms, the second on Russian norms): stereotype rating on NCS vs. self-report on NEO PI-R 0.856, 95%CI (0.757, 0.923); 0.781, 95%CI (0.630, 0.883); stereotype rating on NCS vs. self-report on NCS 0.477, 95%CI (0.152, 0.713); 0.421, 95%CI (0.061, 0.682). Thus replicability indexes showed that the patterns of correlations are relatively stable and expected to be replicated to substantial degree. R script for replicability indexes is to be found at osf.io/2pgtj Correlation_replicability.R.

measurement instrument (i.e., the NCS) for ratings of national stereotypes and the accuracy criterion – traits of real people – resulted in higher stereotype accuracy as compared to situation when different questionnaires for ratings of stereotypes and real people were used.

Norms for transformation of raw scores into T-scores. With respect to norms, we could use data from both Central European and Baltic Sea region. Russian norms yielded higher level of stereotype accuracy as compared to international norms in case of both ICC and Pearson based correlation in the Central European sample (see the median values in the middle of Table 1). The same applied in Baltic Sea countries, but only when Pearson correlations were used. The ICC provided comparable medians of stereotype accuracy for Russian and international/US norms. In sum, the comparison of distinct norms used for transformations of raw data into *T*-scores across the ten countries did not show consistent effect on the resulting level of stereotype accuracy, but the Russian-based standardization slightly overperformed the others.

Pearson's r and ICC. Medians of stereotype accuracy based on Pearson's *r* in Central European countries for Russian and international norms were .36, .24 and .34, .18 for Baltic Sea samples. Medians of stereotype accuracy based on ICC in Central European countries for Russian and international norms were .13, .07 and .12, .07 for Baltic Sea samples.Thus medians of Pearson's *r* were consistently higher than medians of stereotype accuracy based on ICC in both and Baltic Sea samples (see the first four columns of Table 1).

Spearman's rank-order correlations. In order to test an alternative approach, we ranked countries according to 30 NCS scales and then compared the profiles of rankings using Spearman's rank-order correlations. The last column of Table 1 shows stereotype accuracy in Central European and Baltic Sea countries using Spearman's rank-order correlations. The highest level of agreement in the Central European region was found in the Polish sample, ρ =

.48, 95% RVI [-.43, .43].[13] In the other four countries, the coefficients of agreement between self-reports and stereotypical ratings were not outside of the given RVI (Slovak sample: *ρ* = .36, [-.43, .44]; Czech sample: *ρ* = .18, [-.43, .44]; Austrian sample: *ρ* = .16, [-.44, .44]; German sample: *ρ* = -.07, [-.43, .42]). Despite the fact that the agreement between self-report and stereotypical ratings was significant only in the Polish sample, the size of agreement in the other samples was non-negligible, except for the German sample.

In the *Baltic Sea region*, where the same group of participants rated both national stereotypes and real people, were the coefficients of agreement between national stereotypes and ratings of real people generally higher than the coefficients in Central Europe. We found the highest level of agreement in Finland, *ρ* = .68, 95% RVI [-.29, .54]. In Belarus and Lithuania, the coefficients were also outside of the confidence intervals, *ρ* = .67, [-.29, .54] and *ρ* = .57, [-.30, .54]. We found the lowest coefficients of agreement in Estonia, *ρ* = .45, [-.31, .54], and in Latvia, *ρ* = .44, [-.30, .53]. These correlations were all positive and sizable with respect to typical effect sizes in the area.

Rank order correlations on the domain level were also run. Six facets were ranked separately for five personality domains in Central European and Baltic See countries. Since the bootstrapping technique is sensitive to the number of items (30 vs 6) the estimation of statistical significance was not interpreted in the analysis. Creating ranks on six facets only (instead of 30) leads to relatively broad random variation intervals and even high correlation coefficients lie inside these intervals. The results of separate rankings for the six facets showed *ρ* higher than .40 in all personality domains; 10 coefficients reach this criterion in Central European countries and 13 in Baltic see countries. In six countries the *ρ* coefficients were higher than .40 in Agreeableness (CZ, PL, AU, GE, EST, LAT); Extraversion (FIN, LIT, BLR, LAT, CZ, AU) and Openness to Experience (EST, FIN, LIT, BLR, SK, PL). In

[13] The random variance intervals (RVI) specify the range of stereotype accuracy based on chance. The value of Spearman’s rho outside of this interval indicates agreement between stereotypical and real people ratings beyond chance.

addition Neuroticism was relatively accurately judged in Latvia and Slovakia and Conscientiousness in Estonia, Finland and the Czech Republic. The rank order correlations at the trait level therefore demonstrated accuracy in some traits in some countries.[14]

Discussion

Research on accuracy of group stereotypes in personality psychology has recently gained increased attention. While stereotypes on biologically rooted characteristics such as gender and age were found to agree with characteristics of women and men of certain age, studies on the accuracy of national stereotypes returned less clear-cut picture. In order to address the unequivocal conclusions of past research concerning the accuracy of national stereotypes, we compared several methodological approaches that can influence agreement between national stereotypes and personality characteristics of real people living in the given countries. Our thorough analysis of methods, norms for standardization and correlation techniques showed that each of the methodological approaches can shape the resulting stereotype accuracy.

We found that using the same method for ratings of national stereotypes and real people (i.e., the NCS) considerably increased the agreement between both types of rating even if independent groups of participants rated national stereotypes and real people. Furthermore, we found that in the past studies usually employed correlation technique, the ICCs), underestimated the level of agreement between national stereotypes and characteristics of real people. Our results showed that the previously suggested alternative, Pearson’s r, has limitations in estimating the convergence between group stereotypes and real people due to the associated statistical significance.

We proposed a novel approach to stereotype accuracy that uses ranks of the studied countries and non-parametrical Spearman’s correlation with bootstrapped confidence intervals

[14] More detailed information about the Spearman correlations on the domain level are provided at osf.io/2pgtj Spearman correlations on domain level in 10 countries.xlsx.

rather than external reference norms employed in past studies Furthermore, this approach is the most robust as compared to past research since it can account for dependencies in the data (facets belonging to the same domain) and is not biased by unrealistically high *a priori* expectations for what would be a significant level of accuracy. All in all, the novel approach provided evidence that national stereotypes are generally accurate.

Methods for National Stereotypes and Characteristics of Real People

Studies on accuracy of group stereotypes that employ Five Factor Model for operationalization of stereotype content offer an advantage of elaborated methods for rating of stereotypical and real people characteristics. The NEO PI-R is a method with a long tradition, employed throughout different contexts to capture individuals' personality traits. The recently developed NCS for measuring traits typical of countries' representatives contains the same characteristics as the NEO PI-R, which makes the comparison between stereotypical and real people characteristics easier. Although theoretically both instruments operate with the same theoretical background, their form is quite different (i.e., number of items and their wording). Konstabel, Lönnqvist, and Walkowitz (2012) found agreement between self-reports provided on the NCS and the NEO PI-R, ranging between .19 and .74, *Mdn* = 0.55. If NCS and NEO PI-R correlate only moderately when identical targets (i.e., real people) are rated, then even lower level of agreement can be expected when using them for ratings of different targets (i.e., real people vs. stereotypes).

Reviewing past research on accuracy of national stereotypes, one can notice that studies employing different measures for stereotypes and characteristics of real people found lower level of stereotype accuracy (Terracciano et al., 2005) than studies that employed the same measurement instrument (Realo et al., 2009). However, in research by Realo et al. (2009) it was not only the same instrument that was used for ratings of national stereotypes and real people but also just one group of participants that provided ratings of both targets, which could have further enhanced the level of stereotypical accuracy. This is exactly what we addressed in our current study where different groups of participants from Central Europe provided ratings of national stereotypes and real people using the same measurement instrument. We showed that despite different people who rated stereotypes and real people, stereotypical accuracy was higher when using the same method for both ratings. Thus using the same method for stereotype rating and self-report is crucial in accuracy research.

Norms for Standardization and Correlation Techniques

Unlike the methods for self-ratings of real people, no research has previously examined the effect of standardization on the accuracy of national stereotypes. The second methodological consideration that we focused on in our study was whether the same norms for standardization of raw scores (i.e., Russian norms) or different norms for standardization (i.e., international and U.S. norms) influence the resulting level of stereotype accuracy. The results on standardization cannot be interpreted independently of related methodological issues. Standardization is the first step to comparison of stereotypical and real people profiles with a chosen correlation technique. Based on the different combinations of methodological choices examined in our research, the results showed that the use of the same method for ratings of both targets in combination with Russian norms for standardization and Pearson's r yielded the highest level of stereotypical accuracy. Following these methodological choices, national stereotypes significantly agreed with characteristics of real people in six out of ten European countries. When we used international norms for standardization of NCS-based ratings of real people and US norms for standardization of NCS-based stereotypical profiles, stereotypes were accurate in four out of ten countries.

The effect of the two most widely used correlation techniques in research on stereotype accuracy was also compared by McCrae and colleagues (2013). Their results showed that the more conservative ICC were consistently smaller than the Pearson's correlations. The median Pearson correlation between stereotypical and real people profiles across 26 countries was .50; whereas ICC based median was .38. Our results support their findings in that the median based on Pearson's correlations across the 10 countries was .32; while the ICCs-based median was .11.

Methodological Advancements in the Estimation of Stereotype Accuracy

Our study brought evidence that different methodological approaches to data sampling and data analysis can decisively influence the accuracy of national stereotypes as simulated on the very same data. Thus, the key aim of our research was to suggest and test a novel approach that would overcome the methodological issues complicating interpretation of results from previous studies. The employed Spearman's rank order correlations address the problematic choice of normative samples for standardization of raw scores and of correlation techniques.

In the Central European countries, the median agreement between Spearman rank order correlation of national stereotypes and characteristics of real people was .18. In Poland, we found a significant convergence between national stereotypes and real people, corroborating the findings of accurate perception of Polish national stereotypes in previous studies (Realo et al., 2009; Terracciano et al., 2005). In the Baltic Sea region, the median agreement between national stereotypes and characteristics of real people was considerably higher than in Central Europe, median Spearman rank order correlation .57. The reason for the higher level of stereotype accuracy estimated with the rank order correlations in the Baltic Sea region is probably the difference in the employed methodology. All other procedures being equal, in the Baltic Sea region the same participants provided both stereotypical and real people ratings, whereas in Central Europe, different participants rated national stereotypes and real people. Social psychological literature documents that different types of biases occur when more social targets are rated at the same time. The tendency to expect similarities between oneself and others was termed social projection (Robbins & Kruger, 2005) or alternatively false consensus effect (Ross, Greene, & House, 1997). A meta-analytic review showed that social projection is a robust phenomenon with a medium to large effect size (Mullen et al., 1985; Robbins & Kruger, 2005). If the same participants rate themselves and national stereotypes, then the rank order correlation between the two profiles can reflect mutual interdependence caused by such within-subject design where social projection enhances agreement between ratings of different targets. The non-parametric rank order correlations and bootstrapping can partially account for the effect of social projection as shown by more positive values of random variance intervals, the RVI, in the Baltic Sea region as compared to mid-point symmetrical RVI in Central Europe. All in all, the higher agreement between stereotypical and real people profiles in the Baltic Sea region testifies the fact that people from the Baltic Sea region are able to more accurately extrapolate characteristics of national stereotype from characteristics of real people than people in Central Europe.

The key issue in research on stereotype accuracy is the interpretation of the resulting correlations between two profiles. The question of what actually is a reasonable standard for characterizing a stereotype as accurate is still under discussion among researchers. In other words, how much of a congruence between two profiles should be considered accurate (see also Allik, Borkenau, Hřebíčková, Kuppens, & Realo, 2015). Cohen (1988) encourages social scientists to examine not only the statistical significance, but the size of the effects they obtained in their studies as well. He suggests that effect sizes above .8 corresponding to correlations of approximately .40 could be considered accurate because they represent a large congruence between a stereotype and reality. According to Rosenthal (1991) a correlation of at least .40 means that people’s judgements are correct at least 70% of the time. Also Jussim et al., (2016) suggest the correlation of .40 as an appropriate cutoff point for considering a stereotype to be reasonably accurate. They characterize the correlations between the stereotypes and accuracy criteria ranging from .25 to .40 as moderately accurate and correlations below .25 to be inaccurate. Empirical research based on large international data sets showed that stereotype accuracy differs across different social groups. Whereas the median of accuracy scores for age stereotypes is .74 (Chan et al., 2012) and .67 for gender stereotypes (Löckenhoff et al., 2013), the median accuracy of national stereotypes is only .12 (McCrae, et al., 2013). Thus it is hard to establish objective gold standard for accuracy of stereotypes. According to the suggested margins that delineate correlations higher than .4 as high and correlations ranging from .25 to .40 as moderate, the accuracy of national stereotypes based on the rank order correlations is large in six out of ten countries and moderate in one country. Moreover our approach to determining significance did not rely on a priori cutoff points, but rather on empirical estimates. In effect, our criteria for "significance" appeared even more conservative than in other studies.

Limitations and Future Directions

McCrae et al. (2013) assumed that the concept of national stereotypes only exists through the contrast of real or imagined differences across nations. Past research shows that people tend to rate ingroup stereotypes higher and outgroup stereotypes lower on the same dimension—or vice versa—resulting in mirroring of stereotypical profiles (Realo et al., 2009; Terracciano et al., 2005; Hřebíčková et al., (2017). People from geographically close countries therefore sometimes contrast their ingroup stereotypes against the outgroup stereotypes of neighboring countries. They exaggerate the real differences between the ingroup and the relevant outgroups because they probably strive for group distinctiveness (Brewer, 1999; Plichtová, Lášticová, & Petrjanošová, 2009), meaning that national stereotypes are relative to the perception of other countries. Because we ranked countries that are geographically close and where mirroring was detected, the next step is to apply our country ranking-based approach in geographically distant countries.

The statistical procedure proposed in this study is sensitive to the number of ranking countries. In our study we applied ranking in five countries. However a smaller sample of researched countries for ranking could be problematic. Future studies should also address the issue of whether our approach is beneficial assuming high number of countries. One option would be to reanalyze the real data from the PPOC project with 49 countries (Terracciano et al., 2005) and to check whether results based on rank order correlations will lead to the same conclusions about the inaccuracy of national stereotypes. Conducting a simulation study would be another future research option.

Other potential pitfalls of the rank order correlation-based method lie in the number of ranked variables. The bootstrapping technique for statistical significance estimation is applicable when profiles contain multiple variables. A small number of variables in profiles cause low variability and even high correlation coefficients lie inside RVIs. We discovered this fact when we ranked six variables (facets) separately for each Big Five dimension. Our results show that ratings of national stereotypes and real people converge on three personality domains (Extraversion, Openness to Experience and Agreeableness) in the majority of the researched countries. Even though the whole profiles of stereotypical and real people traits do not correspond, the convergence found on certain traits can provide interesting insights into the given cultural context. The next studies should therefore test national stereotype accuracy on the domain level, profile similarity approach however is not the optimal way to conduct such research – alternative approaches that focus on single traits should be used instead.

The accuracy of national stereotypes is studied predominantly using the Big Five personality inventories. Studies measuring other variables (e.g. behavior, values, socioeconomic characteristic) would therefore be a great help in indicating whether low accuracy is limited to personality traits. There is a need for research directly assessing the role of the reference group effect in the context of national stereotype accuracy as well.

The results of this study indicate that in addition to methods, norms and correlation techniques, the levels of stereotype accuracy could also be influenced by participants themselves. In Central Europe, different groups of participants rated national stereotypes and real people, while in the Baltic Sea region the same participants provided both ratings. Future studies should include ratings of both national stereotypes and real people by the same versus different participants in each of the researched countries in order to empirically test our preliminary findings. Such complex design would allow to further investigate the effect of participants on stereotype accuracy. Since the research on stereotype accuracy is often criticized because of the haphazard samples criteria (Jussim at al., 2016), more representative samples are therefore needed in future studies.

In our previous study (Hřebíčková & Graf, 2014) we found a high level of congruence between national stereotypes rated by university students and working adults. However, the levels of congruence between students and working adults depend on the target of their rating. The two age groups agreed more on the outgroup stereotypes concerning the four neighboring countries in comparison to the Czech ingroup stereotypes. Czech working adults perceived a typical country representative in a more socially desirable way than students; the stereotypical profile therefore corresponds more with the self-report measures and is more accurate. In Estonia and Latvia (Realo et al., 2009), where the working adults samples were also available, the results showed a minor role of age differences in ingroup stereotype ratings. Considering the inconsistent findings regarding the effect of participants' age on the results, the national stereotypes should be rated by more representative, age balanced samples.

Finally, some authors argue that aggregate personality traits are not appropriate criteria for evaluating the accuracy of national stereotypes, because different translations, response styles or reference group effects (RGE) may limit scalar equivalence of the scores (Church, et al., 2011; Heine, Butchel, & Norenzayan, 2008; Zecca et al., 2013). McCrae with co-authors

(2013) cited several studies documenting that response styles and translations may have some impact on country level scores, but that it is likely to be relatively small. They also react in detail to the critique relating to the reference group effect. RGE assumes that members of different cultures implicitly use different standards of evaluation from one another, because they implicitly arrive at ratings by comparison with the typical person in their own culture (Heine at al., 2008). They argue that responses to personality items are not absolute judgments, but are made relative to some implicit normative group, therefore they recommend behavioral criteria. McCrae et al. (2013) points out that if this problem was pervasive; nearly all cultures would, on average, rate themselves as average on personality traits – something which data do not demonstrate. Furthermore, Heine et al. (2008) conducted a single study focusing on only one trait (conscientiousness). The extent to which the reference group effect explains the lack of apparent accuracy in studies of national stereotypes is therefore unclear (Jussim et al., 2016).

Guidelines for Future Studies on Accuracy of National Stereotypes

This study addressed different methodological issues in research on stereotype accuracy with the aim of suggesting a more optimal procedure. Based on our findings we can articulate several recommendations for future studies on stereotype accuracy:

1) Employing the same method for ratings of stereotypes and real people results in higher stereotype accuracy in comparison to employing different methods. Thus, we recommend using the same method for ratings of all studied targets.

2) The rank order correlations are an optimal way to assess the accuracy of national stereotypes. This approach ranks a given set of (national) groups in any characteristic and then correlates the profiles of ranks pertaining to the groups (i.e., any profile consists of the respective group's ranks across all personality traits). The main advantage of this approach is that it does not require any additional normative data for transformation of the raw data into

T-scores and decisions about the significance of resulting correlations. However, using the rank order correlations implies a sufficient number of groups (e.g., countries) for creating ranks and also a sufficient number of variables that make up a profile. In this study we successfully used this approach in 5 countries and 30 facets in profile. In principle, the number of countries could be anything higher than one, but with a very low number of groups the ranks are not very informative. More research is needed to test and specify the exact assumptions of our suggested analytical approach.

3) Since the method of bootstrapping avoids the issues of statistical significance associated with Pearson's and ICC correlations, it is a more convenient way of deciding whether stereotypes are accurate or inaccurate. Based on the repeated pseudo estimates of stereotype accuracy, one can determine random variation intervals (RVI) specific to each group (e.g., country). The RVI then enable evaluation of the significance of agreement between self- and stereotypical ratings in the original data set. This approach eliminates the potential effect of social projection on stereotype accuracy, which occurs when the same participants provide both self-report and stereotype rating. This approach also fixes the potentially unrealistically high a priori expectations for accuracy to appear significant.[15]

Even though our study primarily focuses on the accuracy of national stereotypes, we strongly believe that our recommendations are applicable to a wider range of social groups, such as typical representatives of states, regions or cities.

[15] The standard way of calculating the agreement between two profiles of 30 characteristics requires a correlation > .35 for it to be significant, whereas typical effect sizes are much lower in social sciences).

References

Allik, J., Mõttus, R., Realo, A., Pullmann, H., Trifonova, A., McCrae, R. R. & 56 Members of the Russian Character and Personality Survey (2011). Personality profiles and the "Russian Soul:" Literary and scholarly views evaluated. *Journal of Cross-Cultural Psychology, 42,* 372–389.

Allik, J., Borkenau, P., Hřebíčková, M., Kuppens, P., & Realo, A. (2015). How are personality trait and profile agreement related? *Frontiers in Psychology, 6*, ArtID: 785.

Allik, J., Realo, A., Mõttus, R., Pullmann, H., Trifonova, A., McCrae, R. R., & 56 Members of the Russian Character and Personality Survey (2009). Personality traits of Russians from the observer's perspective. *European Journal of Personality, 23*, 567–588.

Allik, J., Mõttus, R., & Realo, A. (2010). Does national character reflect mean personality traits when both are measured by the same instrument? *Journal of Research in Personality, 44*, 62-69.

Allport, G.W. (1954). *The nature of prejudice*. Reading: Massachusetts. Addison-Wesley Publishing Company.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate? *Journal of Social Issues, 55*, 429–444.

Campbell, D.T. (1967). Stereotypes and the perception of group differences. *American Psychologists, 22*, 817-829.

Chan, W., McCrae, R.R., De Fruyt, F., Jussim, L., Löckenhoff, C.E., De Bolle, M. ... & Terracciano, A. (2012). Stereotypes of age differences in personality traits: Universal and accurate? *Journal of Personality and Social Psychology, 103*, 1050–1066.

Church, A.T., Alvarez, J.M., Mai, N.T.Q., French, B. F., Katigbak, M.S. & Ortiz, F.A. (2011). Are cross-cultural comparisons of personality profiles meaningful? Differential item and facet functioning in the Revised NEO Personality Inventory. *Journal of Personality and Social Psychology,10*, 1068-1089.

Cohen, J (1988). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Costa, P. T., & McCrae, R. R. (1992): *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J. (1955). Process affecting scores on “understanding others” and “assumed similarity”. *Psychological Bulletin, 52*, 177-193.

Furr, R.M. (2010). The double-entry intraclass correlation as an index of profile similarity: Meaning, limitations, and alternatives. *Journal of Personality Assessment*, 92, 1-15.

Griffin, D., & Gonzales, R. (1995). Correlational analysis of dyad-level data in the exchangeable case. *Psychological Bulletin, 118*, 430-439.

Heine, S.J., Buchtel, E.E., Norenzayan, A. (2008). What do cross-national comparisons of personality traits tell us? The case of conscientiousness. *Psychological Science, 19*, 309-313.

Hřebíčková, M. (2004). *NEO osobnostní inventář podle NEO PI-R P.T. Costy a R.R. McCraee* [NEO Personality Inventory according to P.T. Costa and R.R. McCrae]. Praha, Testcentrum.

Hřebíčková, M., Graf, S., Tegdes, T., Brezina, I. (2017) : We are the opposite of you! Mirroring of national, regional and ethnic stereotypes. *Journal of Social Psychology,157*, 703-717.

Hřebíčková, M., & Graf, S. (2014). Accuracy of national stereotypes in Central Europe: Outgroups are not better than ingroup in considering personality traits of real people. *European Journal of Personality, 28*, 60–72.

Hřebíčková, M., & Kouřilová, S. (2012): Jak se vidíme, jak nás vidí a jací jsme: porovnání českého národního auto- a heterostereotypu s posuzováním reálných lidí v kontextu pětifaktorového modelu osobnosti [How do we see ourselves, how are we seen and how we are: Comparison of Czech national auto- and heterostereotypes with ratings of real people in context of five-factor model of personality]. *Československá psychologie, 56*, 1-18.

Konstabel, K., Lönnqvist, J-E., & Walkowitz, G. (2012): The 'Short Five' (S5): Measuring personality traits using comprehensive single items. *European Journal of Personality*, 26, 13–29.

Krueger, J. I. & Wright, J. C. (2006). How to measure national stereotypes? *Science, 311*, 776-779.

Jussim, L., Cain, T.R., Crawford, J.T., Harber, K., & Cohen, F. (2009). The unbearable accuracy of stereotypes. In T.D. Nelson (Ed.), *Handbook of prejudice, stereotyping and discrimination* (pp. 199-227). New York: Taylor & Francis Group.

Jussim, L., Crawford, J.T., Anglin, S.M., Chambers, J.R., Stevens, S.T. & Cohen, F. (2016). Stereotype accuracy: One of the largest and most replicable effects in all socail psychology. In T.D. Nelson (Ed.), *Handbook of prejudice, stereotyping and discrimination, 2nd edition* (pp. 31-63). New York, NY; London: Psychology Press, Taylor & Francis Group.

Löckenhoff, C.E., Chan, W., McCrae, R.R., De Fruyt, F., Jussim, L….& Terracciano, A. (2014). Gender stereotypes of personality: Universal and accurate? *Journal of Cross-Cultural Psychology, 45*, 675–694.

Lönnqvist, J-E., Konstabel, K., Lönnqvist, N. & Verkasalo, M. (2014). Accuracy, consensus, in-group bias, and cultural frame shifting in the context of national character stereotypes. *Journal of Social Psychology, 154*, 40–58.

Lorenzo-Seva, U., & ten Berge, J.M.F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *European Journal of Research Methods for the Behavioral & Social Sciences, 2,* 57-64.

Mackie, M. (1973). Arriving at "truth" by definition: The case of stereotype inaccuracy. *Social Problems, 20*, 489-499.

Martin, T.A., Draguns, J.G., Oryol, V.E., Senin, I.G., A. A. Rukavishnikov, & Klotz, M.L. (2002). The Russian NEO PI-R and NEO-FFI, Yaroslavl, Russia: Psychodiagnostica.

McCrae, R. R. (2002). NEO PI-R data from 36 cultures. In A. J. Marsella (Series Ed.) & R. R. McCrae & J. Allik (Eds.), *The five-factor model across cultures* (pp. 53-78), New York, NY: Kluwer Academic Publishers/Plenum.

McCrae, R. R. (2008). A note on some measures of profile agreement. *Journal of Personality Assessment, 90*, 105-109.

McCrae, R. R., Zonderman, A. B., Costa, P. T. Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology, 70*, 552-566.

McCrae, R.R., Terracciano, A., Realo, A. & Allik, J. (2007). Climatic warmth and national wealth: some culture-level determinants of national character stereotypes. *European Journal of Personality, 21*, 953-976.

McCrae, R. R., Chan, W., Jussim, L., De Fruyt, F., Löckenhoff, C. E., De Bolle, M., ..., & Terracciano, A. (2013). The inaccuracy of national character stereotypes. *Journal of Research in Personality, 47*, 831–842.

Mullen, B., Atkins, J.L., Champion, D.S., Edwards, C., Hardy, D., Story, J.E., & Vanderklok, M., (1985). The false consensus effect: A meta-analysis of 115 hypothesis tests. *Journal of Experimental Social Psychology, 21*, 262-283.

Ostendorf, F., & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae. Revidierte Fassung* [NEO Personality Inventory according to Costa and McCrae. Revised], Göttingen, Germany, Hogrefe.

Plichtová, J., Lášticová, B., Pertjánošová, M., (2009). *Konštuovanie slovenskosti vo verejnom priestore*. [Constructing Slovakness in Public Space.] Bratislava: Kabinet výskumu sociálnej a biologickej komunikácie SAV.

Realo, A., Allik, J., Lönnqvist, J. E., Verkasalo, M., Kwiatkowska, A., Kööts, L.,... &Renge, V. (2009). Mechanisms of the national character stereotype: How people in six neighboring countries of Russia describe themselves and the typical Russian. *European Journal of Personality, 23*, 229-249.

Richard, F.D., Bond, C.F., Jr., & Stokes-Zoota, J.J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7,* 331-363.

Robbins, J.M. & Kruger, J.I. (2005). Social projection to ingroups and outgroups: A review and meta-analysis. *Personality and Social Psychology Review, 9*, 32-47.

Rosenthal, R., (1991). Effect sizes: Pearson´s correlation, its display via the BESD, and alternative indices. *American Psychologists, 46*, 1086-1087.

Ross, L., Greene, D. & House, P. (1977). The false consensus effect: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology, 13*, 279-301.

Rossier, J., Aluja, A., Blanch, A., Barry, O., Hansenne, M., Carvalho, André F.,…& Karagonlar, G. (2016). Cross- cultural generalizability of the alternative five- factor model using the Zuckerman–Kuhlman–Aluja personality questionnaire. *European Journal of Personality, 30*, 139-156.

Sherman, R.A., & Wood, D. (2014). Estimating the expected replicability of a pattern of correlations and other measures of association. *Multivariate Behavioral Research*, *49*,1, 17-40, DOI: 10.1080/00273171.2013.822785.

Siuta, J. (2007). *Inwentarz osobowości NEO PI-R Paula T. Costy i Roberta McCrae. Aplikacija polska. Podręcznik* [NEO Personality Inventory NEO PI-R of Paul Costa and Robert McCrae. Polish version. Questionnaire]. Warszawa, Pracownia Testów Psychologicznych.

Terracciano, A., Abdel-Khalek, A. M., Ádám, N., Adamovová, L., Ahn, C. K., Ahn, H.-N.,...& McCrae, R. R. (2005). National character does not reflect mean personality trait levels in 49 cultures. *Science, 310*, 96-100.

Zecca, G., Verardi, S., Antonietti, J-P., Dahourou, D., Adjahouisso, M. Ah-Kion, J. …& Rossier, J. (2013). African cultures and the Five-Factor Model of personality: Evidence for a specific pan-African structure and profile? *Journal of Cross-Cultural Psychology, 44*, 684-700.

Zhong, Ch.-B. & Leonardelli, G.J. (2008). Cold and lonely: Does social exclusion literally feel cold? Psychological Science,19, 838-842.

Table 1

Stereotype accuracy in ten European countries based on two accuracy criteria (NCS vs NEO PI-R); US, Russian and international, norms; intraclass correlations (ICCs), Person's r and Spearman rank-order correlations (Rho)

	Accuracy criterion	National stereotypes rated on NCS					
		ICC		*Pearson's r*		*Median*	*Rho*
		RusN	IntUSN	RusN	IntUSN		
Austrians	NEO (S)	-.23	-.17	-.09	-.12	*-.15*	
	NCS (S)	.17	.01	$.49^{**}$	.16	*.17*	.16
Czechs	NEO (S)	$-.36^{*}$	.31	-.20	$.43^{*}$	*.06*	
	NCS (S)	.20	.06	$.64^{***}$	.32	*.26*	.18
Germans	NEO (S)	-.32	-.28	-.28	-.28	*-.28*	
	NCS (S)	.22	.07	$.39^{*}$	.09	*.16*	-.07
Poles	NEO (S)	.03	$.61^{***}$	.26	$.64^{***}$	*.44*	
	NCS (S)	.09	.25	.32	$.35^{*}$	*.29*	$.48^{*}$
Slovaks	NEO (R)	.30	-.08	$.59^{***}$	.05	*.18*	
	NCS (S)	.21	$.48^{**}$	$.42^{*}$	$.51^{**}$	*.45*	.36
Median		*.13*	*.07*	*.36*	*.24*		*.18*
Belarusians	NCS (S)	.20	.34	$.51^{**}$	$.70^{***}$	*.43*	$.67^{*}$
Estonians	NCS (S)	.05	.07	.31	.16	*.12*	.45
Finns	NCS (S)	.33	$.44^{*}$	$.71^{***}$	$.69^{***}$	*.57*	$.68^{*}$
Latvians	NCS (S)	.12	.07	.34	.09	*.11*	$.57^{*}$
Lithuanians	NCS (S)	-.04	.13	.07	.18	*.10*	.44
Median		*.12*	*.13*	*.34*	*.18*		*.57*
Overall median		***.12***	***.07***	***.34***	***.18***		***.45***

Notes. ICC = Intraclass correlations; Rho = Spearman rank-order correlations. RusN = Russian norms; IntUSN = international norms in case of ratings of national stereotypes or American norms in case of ratings of real people. NCS = National Character Survey; NEO PI-R = NEO Personality Inventory. (S) self-report; (R) observer rating.
$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.