The Meaning of Aggression Varies Across Culture: Testing the Measurement Invariance of the Refined Aggression Questionnaire in Samples From Spain, the United States, and Hong Kong

David Gallardo-Pujol, Eva Penelo, Cindy Sit, Montsant Jornet-Gibert, Carlos Suso, Macià Buades-Rotger, Alberto Maydeu-Olivares, Antonio Andrés-Pueyo & Fred B. Bryant

To cite this article: David Gallardo-Pujol, Eva Penelo, Cindy Sit, Montsant Jornet-Gibert, Carlos Suso, Macià Buades-Rotger, Alberto Maydeu-Olivares, Antonio Andrés-Pueyo & Fred B. Bryant (2019): The Meaning of Aggression Varies Across Culture: Testing the Measurement Invariance of the Refined Aggression Questionnaire in Samples From Spain, the United States, and Hong Kong, Journal of Personality Assessment

To link to this article: https://doi.org/10.1080/00223891.2019.1565572

Published online: 25 Mar 2019.
The Meaning of Aggression Varies Across Culture: Testing the Measurement Invariance of the Refined Aggression Questionnaire in Samples From Spain, the United States, and Hong Kong

David Gallardo-Pujol1, Eva Penelo2, Cindy Sit3, Montsant Jornet-Gibert1, Carlos Suso4, Macià Buades-Rotger5, Alberto Maydeu-Olivares1,6, Antonio Andrés-Pueyo1, and Fred B. Bryant7

1Department of Personality, Universitat de Barcelona, and Institute for Neurosciences (UBNeuro), Barcelona, Spain; 2Laboratori d’Estadística Aplicada, Departament de Psicobiologia i Metodologia de les Ciències de la Salut, Universitat Autònoma de Barcelona, Barcelona, Spain; 3The Chinese University of Hong Kong, Hong Kong, China; 4Universitat Jaume I, Castellón, Spain; 5Universität zu Lübeck, Lübeck, Germany; 6University of South Carolina; 7Loyola University Chicago

ABSTRACT
Cultural differences in aggression are still poorly understood. The purpose of this article is to assess whether a tool for measuring aggression has the same meaning across cultures. Analyzing samples from Spain (n = 262), the United States (n = 344), and Hong Kong (n = 645), we used confirmatory factor analysis to investigate measurement invariance of the refined version of the Aggression Questionnaire (Bryant & Smith, 2001). The measurement of aggression was more equivalent between the Chinese and Spanish versions than between these two and the U.S. version. Aggression does not show invariance at the cultural level. Cultural variables such as affective autonomy or individualism could influence the meaning of aggression. Aggressive behavior models can be improved by incorporating cultural variables.

Understanding factors that promote aggression is of crucial importance (Anderson & Bushman, 2002). Although aggression is inherent to human behavior (Anderson & Bushman, 2002; Berkowitz, 1993; Berkowitz & Harmon-Jones, 2004), little is known about how culture influences aggression (Fry, 1998).

Updating the Buss–Durkee Hostility Inventory (Buss & Durkee, 1957), Buss and Perry (1992) developed an empirically based questionnaire, the Aggression Questionnaire (AQ), which became the gold standard in the field. They obtained a four-factor structure that has been replicated across samples (e.g., Ramirez, Andreu, & Fujihara, 2001; Williams, Boyd, Cascardi, & Poythress, 1996) and languages (Andreu-Rodriguez, Peña-Fernández, & Graña-Gómez, 2002; Fossati, Maffei, Acquarini, & Di Ceglie, 2003; Maxwell, 2007; Nakano, 2001; von Collani & Werner, 2005).

To improve its structural stability, Bryant and Smith (2001) shortened the original AQ to 12 items (AQ–R). This version allows for efficient administration and maintains high standards of validity and reliability (Gallardo-Pujol et al., 2006). It has also been translated into Chinese (Maxwell, 2007) and Spanish (Gallardo-Pujol et al., 2006). Yet, its measurement invariance across culture remains unknown.

Measurement invariance or measurement equivalence consists of different levels (Kankaraş et al., 2010; Vijver & Leung, 1997). Structural or configural invariance exists when the given construct shows the same factor structure across different cultures. Metric invariance exists when factor loadings (which reflect the meaning of the construct) are equal across different cultures. Finally, scalar invariance exists when the intercepts of the indicators are the same across groups. This implies that mean differences across cultures might reflect actual mean differences in the latent constructs.

Many studies have explored the configural invariance of the AQ–R (Fossati et al., 2003; Gallardo-Pujol et al., 2006; Maxwell, 2007; Nakano, 2001; Vigil-Colet, Lorenzo-Seva, Codorniu-Raga, & Morales, 2005), confirming the same set of factors in all adaptations so far. Yet, its full measurement invariance (configural, metric, and scalar) across cultures has not been investigated. Establishing metric invariance is the first step in showing that cross-cultural differences in mean aggression scores reflect differences in aggression levels rather than unknown factors. Indeed, directly comparing mean scores (scalar invariance) without establishing metric invariance could produce distorted conclusions. Hence, the aim of this study was to evaluate measurement invariance across three different versions of the AQ–R: Spanish, U.S. English, and Chinese (Hong Kong).

The reasons for choosing these three cultures are not trivial. Benet-Martínez (Aaker, Benet-Martínez, & Garolera, 2001; Benet-Martínez, 2007) proposed an approach for evaluating cultural differences based on a triangulation of three
cultures that vary with respect to at least two explanatory constructs (Benet-Martínez, 2007). Hence, we selected samples from these three cultures because they vary on two sociocultural dimensions (Schwartz & Bilsky, 1987). These dimensions describe preferences for one state of affairs over another that distinguish countries (Hofstede, 2001; Hofstede & McCrae, 2004). In this case, we evaluated individualism (the United States vs. Spain and Hong Kong) and affective autonomy (Hong Kong vs. Spain and the United States). Individualism (vs. collectivism), defined as the preference for a framework in which individuals are expected to take care of themselves (Hofstede, 2001; Hofstede & McCrae, 2004), has been linked to violence and aggression in Western societies (Menzer & Torney-Purta, 2012). Affective autonomy refers to the independent pursuit of affectively positive experiences (Schwartz & Bilsky, 1987); high affective autonomy is related to leading a pleasant, happy, and exciting life. Hence, low affective autonomy might be related to unhappiness, poor emotion regulation, frustration, and therefore proneness to exhibit aggressive behaviors (Matsumoto, Yoo, & Nakagawa, 2008).

This analysis differs from earlier work in two ways: (a) it is the first study to test the measurement invariance of the AQ–R across Eastern and Western cultures; and (b) it systematically selected three cultures that differ in terms of the possible explanatory or mediating variables responsible for observed structural differences.

Method

Participants and procedure

The Spanish sample, taken from Gallardo-Pujol et al. (2006), consisted of 262 students from Catalonia (154 females, 99 males, and 9 who did not report gender). Mean age was 21.68 (SD = 2.84). Further details are available in Gallardo-Pujol et al. (2006).

The U.S. sample, taken from Bryant and Smith (2001), consisted of 344 U.S. undergraduates (250 females and 94 males) at a private Midwestern metropolitan university. Mean age was 18.49 (SD = 1.26). Further details are available in Bryant and Smith (2001).

The Hong Kong sample, taken from Maxwell (2007), consisted of 645 undergraduate Hong Kong Chinese students (372 females, 272 males, and 1 who did not report gender) at the University of Hong Kong. Mean age was 19.71 (SD = 1.26). Further details are available in Maxwell (2007).

For all samples, participation was voluntary and anonymous, and all participants provided informed consent for the inclusion of their data. The analyses conducted in this study are secondary to already existing data. Secondary analyses involve reanalyzing data collected with different purposes to pursue a new research question not addressed by the original study.

Measures

The AQ–R (Bryant & Smith, 2001) is a short self-report questionnaire that consists of 12 Likert-type items rated on a 5-point scale ranging from 1 (never) to 5 (always). The AQ–R is organized in four scales of three items each: Physical Aggression (PA), Verbal Aggression (VA), Anger (ANG), and Hostility (HO). All versions showed good psychometric properties (Bryant & Smith, 2001; Gallardo-Pujol et al., 2006; Maxwell, 2007).

Statistical analysis

Multigroup confirmatory factor analysis was conducted using polychoric correlations with diagonally weighted least squares (WLSMV) with a mean- and variance-adjusted chi-square test as implemented in Mplus 7.2 (Muthén & Muthén, 2016). For model identification, factor loadings of the first item for each factor were freely estimated, but all factor variances were fixed at 1 to avoid the use of a marker item (Kim & Yoon, 2011). Factors were allowed to intercorrelate. Factorial invariance across the three samples was tested with the chi-square test (Asparouhov & Muthén, 2006) for nested models (Byrne, 2011; Vandenberg & Lance, 2000) estimated using mean- and variance-corrected statistics. This is the procedure DIFFTEST implemented in Mplus. We started with a configural model (Model 1), in which all parameters were freely estimated across samples but the same theoretical model was specified across populations. Then, full metric invariance was tested (Model 2) by equating factor loadings across populations, and freeing factor variances in the second and third groups (which had been fixed at 1 in the first group for model identification, as in Ezepeleta & Penelo, 2015). The metric invariance model across the three populations was rejected. Then, we tested full metric invariance across two of the populations (Models 3–5). Then, we examined partially invariant models (Models 6–9) in which the parameters of one item were relaxed sequentially using a backward procedure (Kim & Yoon, 2011). Finally, should metric invariance have been met, scalar invariance would have been explored. Goodness of fit was assessed using (Jackson, Gillaspy, & Purc-Stephenson, 2009) $\chi^2$, comparative fit index (CFI), Tucker–Lewis Index (TLI), and root mean square error of approximation (RMSEA) using conventional thresholds (Marsh, Hau, & Wen, 2004).

To compare all three questionnaires, we decided to recode values of 6 into values of 5 for the U.S. sample, given that the frequencies of 6 responses were extremely low, median frequency = 4%, compared with the total sample. Converting the 6-point scale to a 5-point scale by recoding 6s as 5s produced item scores that were virtually identical ($r > .988$) to those produced by subtracting 1 from 6-point-scale scores, multiplying the result by 0.8, and adding 1 to the product to obtain a 5-point scale. Additionally, the Spain and Hong Kong samples retained the original AQ 5-point rating scale that was modified in the U.S. AQ–R. To make sure that recoding category data did not affect the results obtained, we repeated all measurement invariance analyses using the original coding (6-point scale for U.S. sample and 5-point scale for the China and Spain samples). The results obtained are consistent with those reported here, only partial measurement invariance holds, and for the same items and combinations of countries reported in this brief report.
Freed parameters) for the Spanish sample, items involved were one from the VA scale (My friends say that I am somewhat argumentative). In both cases, females had larger factor loadings than males.

Gender invariance was also tested within each country, given the asymmetry between males and females in terms of aggression. We found absolute gender invariance in the United States, \(\chi^2(18) = 7.123, p = .5234\), and Hong Kong, \(\chi^2(8) = 3.887, p = .8672\). There was partial gender invariance (20% freed parameters) for the Spanish sample, \(\chi^2(6) = 6.938, p = .3266\). The two items involved were one from the VA scale (My friends say that I am somewhat argumentative/Mis amigos/as dicen que soy discutidor/ra), and another one from the HOST scale (My friends say that I am somewhat argumentative/Mis amigos/as dicen que soy discutidor/ra). In both cases, females had larger factor loadings than males.

Partial metric invariance could not be rejected, \(\Delta\chi^2(6) = 12.3, p = .05\). Four items involved were one from the VA scale (Mis amigos/as dicen que soy discutidor/ra), and another one from the HOST scale (My friends say that I am somewhat argumentative.).

Table 1. Item and scale descriptive for Spanish, U.S., and Hong Kong samples.

<table>
<thead>
<tr>
<th>Scale item</th>
<th>Spain</th>
<th>U.S.</th>
<th>Hong Kong</th>
<th>Spain</th>
<th>U.S.</th>
<th>Hong Kong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical Aggression</td>
<td>4.74</td>
<td>5.72</td>
<td>4.21</td>
<td>.76</td>
<td>.78</td>
<td>.68</td>
</tr>
<tr>
<td>3</td>
<td>1.46</td>
<td>1.76</td>
<td>1.53</td>
<td>.51</td>
<td>.75</td>
<td>.43</td>
</tr>
<tr>
<td>8</td>
<td>1.58</td>
<td>2.06</td>
<td>1.29</td>
<td>.81</td>
<td>.64</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1.69</td>
<td>1.89</td>
<td>1.40</td>
<td>.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal Aggression</td>
<td>8.42</td>
<td>8.52</td>
<td>6.56</td>
<td>.52</td>
<td>.81</td>
<td>.40</td>
</tr>
<tr>
<td>1</td>
<td>2.94</td>
<td>3.21</td>
<td>2.27</td>
<td>.67</td>
<td>.49</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2.55</td>
<td>2.71</td>
<td>2.30</td>
<td>.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2.93</td>
<td>2.60</td>
<td>1.99</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anger</td>
<td>7.21</td>
<td>7.25</td>
<td>6.42</td>
<td>.60</td>
<td>.52</td>
<td>.41</td>
</tr>
<tr>
<td>5</td>
<td>2.38</td>
<td>2.40</td>
<td>2.48</td>
<td>.74</td>
<td>.49</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2.98</td>
<td>2.85</td>
<td>1.79</td>
<td>.71</td>
<td>.46</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1.85</td>
<td>2.00</td>
<td>2.16</td>
<td>.67</td>
<td>.40</td>
<td></td>
</tr>
<tr>
<td>Hostility</td>
<td>6.37</td>
<td>7.96</td>
<td>6.56</td>
<td>.55</td>
<td>.30</td>
<td>.67</td>
</tr>
<tr>
<td>2</td>
<td>1.96</td>
<td>2.70</td>
<td>1.88</td>
<td>.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.95</td>
<td>2.69</td>
<td>2.11</td>
<td>.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2.43</td>
<td>2.57</td>
<td>2.56</td>
<td>.49</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Multigroup confirmatory factor analysis: Measurement equivalence across Spanish, U.S., and Hong Kong samples.

<table>
<thead>
<tr>
<th>Model</th>
<th>Goodness-of-fit indexes</th>
<th>Comparisona</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\chi^2(df))</td>
<td>CFI</td>
</tr>
<tr>
<td>1. Same configuration all</td>
<td>570.733 (144)</td>
<td>.952</td>
</tr>
<tr>
<td>2. Total metric invariance</td>
<td>700.681 (160)</td>
<td>.939</td>
</tr>
<tr>
<td>3. Total metric invariance</td>
<td>593.367 (152)</td>
<td>.950</td>
</tr>
<tr>
<td>4. Total metric invariance</td>
<td>622.746 (152)</td>
<td>.947</td>
</tr>
<tr>
<td>5. Total metric invariance</td>
<td>669.677 (152)</td>
<td>.942</td>
</tr>
<tr>
<td>6. Partial metric invariance</td>
<td>560.125 (149)</td>
<td>.954</td>
</tr>
<tr>
<td>7. Partial metric invariance</td>
<td>551.105 (152)</td>
<td>.955</td>
</tr>
<tr>
<td>8. Partial metric invariance</td>
<td>554.436 (149)</td>
<td>.954</td>
</tr>
<tr>
<td>9. Partial metric invariance</td>
<td>554.672 (150)</td>
<td>.955</td>
</tr>
</tbody>
</table>

Table. CF = comparative fit index; TLI = Tucker–Lewis Index; RMSEA = root mean square error of approximation.
aBased on difference chi-square test for mean- and variance-adjusted chi-squares (Asparouhov & Muthén, 2006).

Results

Table 1 reports descriptive statistics for each item and sub-scale, and internal consistency for each dimension in each of the three samples. Table 2 summarizes the results for the tests of measurement invariance across the three samples. Full metric invariance did not hold across all three samples (Model 2) or between any of the three pairs of samples (Models 3–5). Partial metric invariance held across pairs of samples as follows: six factor loadings equivalent for Spanish and U.S. samples (Model 6), eight factor loadings equivalent for Spanish and Hong Kong samples (Model 7), and six factor loadings equivalent for U.S. and Hong Kong samples (Model 8). Finally, analysis of partial metric invariance across the three samples was conducted simultaneously. Partial metric invariance could not be rejected, \(\Delta\chi^2(6) = 12.3, p = .05\). Fit statistics for the final, partially invariant model (Model 9) were \(\chi^2(157) = 554.7, CFI = .96, TLI = .94, \) and RMSEA = .080. Each model always included a multigroup approach, assessing all three groups, but just fixing parameters across two of the samples and freeing the third not involved (detailed results of sequential analyses are available on request).

Table 1 shows standardized (unstandardized factor loadings are available on request) factor loadings and factor correlations for the final partially invariant model (Model 9). Equivalent factor loadings between samples were as follows: five items (two for PA and one for each of the other factors) across Spanish and U.S. samples, five items (all three for PA and one for VA and HO) across U.S. and Hong Kong samples, and seven items (two for PA, VA and AN, and one for HO) across Spanish and Hong Kong samples. Of these, four items showed equivalent factor loadings across the three samples: two for PA, one for VA, and one for AN. In contrast, two items did not have equivalent factor loadings.
across any of the three samples: one for VA and one for AN, being items showing lower loadings in the Spanish sample. That HO was the only AQ factor with no equivalent loadings across all three samples suggests that culture influences the meaning of hostility more than the meaning of physical or verbal aggression or of anger—a conclusion consistent with cross-cultural research using the 29-item AQ (Vigil-Colet et al., 2005, p. 607).

**Discussion**

Our aim was to assess metric invariance across three versions of the AQ–R. Of the 12 AQ–R items, 7 (58.3%) were metric invariant for the Spanish and Chinese samples, whereas only 5 (41.6%) were invariant for the Spanish and Chinese samples, and 5 (41.6%) for the U.S. and Chinese samples. This pattern of results suggests that aggression is closer in meaning between the Chinese and Spanish versions than between each of these two versions and the U.S. version.

One potential explanation for discrepancies is the use of an imposed-etic approach. This approach refers to the generalized practice of translating and adapting items originally adapted within one culture to another one, in contrast to an emic approach, which relies on items originally developed from within that culture (Berry, 1980; Berry, Poortinga, Breugelmans, Chasiotis, & Sam, 2011). Although imposed-etic instruments allow for quick comparisons across cultures, measurement was not metric equivalent in all three countries, suggesting that the meaning of aggression differs across culture. However, this does not explain the similarities between the two adaptations from English into Spanish and Chinese. These cross-cultural similarities could be attributed to certain values present in each of these societies (Schwartz, 1992). In particular, the similarity between Spain and Hong Kong with respect to the PA and VA subscales might be explained by the similarity between both cultures with respect to individualism (Menzer & Torney-Purta, 2012). Collectivistic societies report fewer episodes of violence at schools (Menzer & Torney-Purta, 2012). Spanish and U.S. adaptations are closer when considering the AN and HO subscales. Thus, it might be reasonable to think that these societies conceive and promote both aspects of aggression in a similar way, given that Spain and the United States show similar levels of affective autonomy compared to Hong Kong society (Aaker et al., 2001, p. 494). However, the variables studied here cannot explain the high degree of variation that remains across all three cultures with respect to the self-reported manifestations of aggression.

Future research should include comparisons among cultures differing on other cultural dimensions (Schwartz, 1992). Such research would complement current aggression models (e.g., the general aggression model) that do not go beyond proximal causes of aggression (Anderson & Bushman, 2002). Moreover, because contemporary models of aggression are culturally centered within the perspective of Western societies (Henrich, Heine, & Norenzayan, 2010), it is important to develop cross-cultural models. Additionally, an important avenue of research could be using item-response theory analyses to study cross-cultural differences in AQ–R (and other measures) with respect to differential item functioning or differential test functioning. It is likely that in the future it would enable fine-grained comparisons (e.g., Hambrick et al., 2010). This work is not
exempt from limitations that should be addressed in further studies. We found that mean age was different in all three samples. This could actually be affecting the composition of the sample and thus hampering the robustness of our findings. However, there is evidence that by the age of our subjects, aggression has already peaked in late adolescence and is actually slowly steadily declining at similar levels (Liu, Lewis, & Evans, 2013; Moffitt, 1993). With respect to gender, we conducted separate analyses to explore gender invariance (see footnote 2) within each country. We only found partial invariance in Spain, but at the threshold for accepting it in practical applications (Dimitrov, 2010), as it is intended this questionnaire (Gallardo-Pujol et al., 2006).

All in all, our results have shown that (a) metric invariance should be tested before proceeding to direct comparisons of national and cultural mean levels of aggression, and (b) certain cultural variables, such as individualism and affective autonomy, could influence the meaning of aggression across culture (Schwartz, 1992). As has typically been the case in previous comparative cross-cultural research on the AQ, this study did not assess criterion measures as correlates of AQ–R subscales across multiple countries. However, because such criterion measures are crucial for establishing cross-cultural construct validity, future international work on the AQ–R should include criterion measures. Our results suggest that this future research should be careful to address potential cross-cultural differences in factor structure, which could otherwise produce misleading evidence about the generalizability of construct validity across culture.

**Funding**

This research was supported in part by Grant PSI2012-33601 from the Spanish Ministry of Economy and Competitiveness, Catalan Government Grant 2017SGR1237, and by the National Science Foundation Grant No. SES-1659936.

**References**


