Early Education and Development

ISSN: 1040-9289 (Print) 1556-6935 (Online) Journal homepage: http://www.tandfonline.com/loi/heed20

The Spanish Version of the Empathy Questionnaire (EmQue): Evidence for Longitudinal Measurement Invariance and Relationship With Emotional Regulation

Beatriz Lucas-Molina, Renata Sarmento, Laura Quintanilla & Marta Giménez-Dasí

To cite this article: Beatriz Lucas-Molina, Renata Sarmento, Laura Quintanilla & Marta Giménez-Dasí (2018) The Spanish Version of the Empathy Questionnaire (EmQue): Evidence for Longitudinal Measurement Invariance and Relationship With Emotional Regulation, Early Education and Development, 29:4, 467-476, DOI: 10.1080/10409289.2018.1427929

To link to this article: https://doi.org/10.1080/10409289.2018.1427929

EARLY EDUCATION AND DEVELOPMENT
2018, VOL. 29, NO. 4, 467–476
https://doi.org/10.1080/10409289.2018.1427929

The Spanish Version of the Empathy Questionnaire (EmQue): Evidence for Longitudinal Measurement Invariance and Relationship With Emotional Regulation

Beatriz Lucas-Molina[a], Renata Sarmento[b], Laura Quintanilla[b], and Marta Giménez-Dasí[a]

[a]Department of Developmental and Educational Psychology, University of Valencia; [b]Department of Methodology of Behavioral Sciences, National University of Distance Education

ABSTRACT
Research Findings: Empathy, or the ability to understand what others are thinking or feeling, can be observed in early developmental stages. The purpose of this study was to validate the Spanish version of the Empathy Questionnaire (EmQue) and examine its longitudinal measurement invariance (LMI) at 2 time points. Parents of 103 children completed the EmQue when their children were 3 (M = 41.76) and 4 (M = 51.65) years old. Confirmatory factor analyses revealed that Grazzani, Ornaghi, Pepe, Brazzelli and Rieffe's (2016) 3-factor model of emotional contagion (EC), attention to others' feelings, and prosocial actions (PA) presented the best fit indices at both time points (Time 1: CFI = .931, TLI = .914, and RMSEA = .070; Time 2: CFI = .941, TLI = .935, and RMSEA = .064). Moreover, preliminary evidence was obtained for the LMI of this model. PA scores increased over time. Score reliability ranged from .60 (EC) to .83 (PA). Positive correlations were found between PA and emotional regulation at each time point and across time.*Practice or Policy*: The great relevance of empathy and prosocial behavior in academic achievement and psychological adjustment justifies the development of reliable instruments to evaluate these constructs from early ages.

Empathy has been defined as an affective response to others' emotions that produces adaptive reactions focused on the needs of others (Hoffman, 2000). Empathy involves affective and cognitive processes, and it is considered the basis of many prosocial behaviors, such as helping or consolation behaviors (Jolliffe & Farrington, 2006; Zahn-Waxler, Radke-Yarrow, Wagner, & Chapman, 1992). Most authors agree that affective empathy is the capacity to share emotional states with others, and cognitive empathy is the ability to adopt the mental perspectives of others (Shamay-Tsoory, 2011). Both components are essential in the development of empathy. In fact, when they are dissociated, as occurs in some pathologies, the individual's social functioning is altered. For example, in autism spectrum disorders, impairments are found in the cognitive component but not the affective component. However, in schizophrenia or depersonalization, impairments are found in the affective component but not the cognitive component (Cox et al., 2012). The balance between affective and cognitive empathy is unique to each person and defines his or her empathic experience (Kerem, Fishman, & Josselson, 2001).

Empathy is an important skill in social relations that can be observed from early childhood and reaches different levels in each person. Some aspects such as temperament, socialization, and cultural norms seem to be involved in the differential development of empathy and can explain some

CONTACT Beatriz Lucas-Molina beatriz.lucas@uv.es Department of Educational and Developmental Psychology, University of Valencia, Avenida Blasco Ibáñez 21, 46010, Valencia, Spain.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/heed.

individual differences (Eisenberg, 2000). Given empathy's relationship with other variables (e.g., externalizing behavior problems) and its relevance in socialization, it is important to discover its developmental path, understand individual differences, and have the tools to assess it from the earliest stages of life.

According to Hoffman (1987), empathy develops sequentially in four levels between the first year of life and late childhood. In this case, emotional contagion (EC), appears in the first year of life. It is not a very adaptive level because it involves an emotional response that echoes what other people express but without the capacity to regulate it. This level corresponds to crying behavior observed when babies see another person or another baby cry. It is assumed that in this state, the lack of differentiation between the self and the other favors this contagion or complete identification with the other. The second level, attention to others' feelings (AOF), begins after 1 year of life. In this case, the child does not experience contagion but instead is able to observe emotion in others as something separate from himself or herself and realize that he or she does not have the same feeling. At this level, children can show concern for others, but they do not have the resources to act. Therefore, there is no contagion of the emotional state; instead, children usually stop and observe the other with a serious expression. Children with more developed skills already perform some types of gestures, such as hugging. The third level, prosocial actions (PA), develops during the second year of life and involves the initiation of helping and consolation behaviors. At this level, children already identify and react to the emotional states of others, and they are capable of adaptively performing helping behaviors. The fourth and final level, empathy for others' living conditions, involves a great change and appears in late childhood. At this level, children feel affected by the negative living conditions of people with whom they have no direct contact. This level is more complex because it involves going beyond the child's direct reality. Likewise, the behaviors corresponding to this level (e.g., sending goods or making donations) do not have an immediate effect that the child can observe.

The empathy levels described by Hoffman have often been confirmed from a theoretical perspective, but until very recently there were no instruments available to measure them empirically (Rieffe, Ketelaar, & Wiefferink, 2010). Thus, despite the importance of this construct, there are almost no specific measurement instruments for early childhood. The only exception is the Empathy Questionnaire (EmQue) developed by Rieffe et al. (2010). Using this questionnaire, the parents of preschool-age children evaluate behaviors related to empathy. The questionnaire includes the empathy levels described by Hoffman in three dimensions that coincide with the first three developmental levels of empathy (EC, AOF, and PA).

The authors' validation of the questionnaire in The Netherlands found good reliability of the three subscales as well as significant negative correlations between the EC subscale and a measure of emotional regulation (ER). Recently, a validation of the questionnaire was carried out in an Italian population (Grazzani, Ornaghi, Pepe, Brazzelli, & Rieffe, 2016). This validation confirmed the factorial structure of the three subscales, although the number of items was reduced from 20 to 13, and their gender invariance was demonstrated. Both validations found positive correlations between children's age and PA.

Much of the research on the development of empathic behaviors is based on the assumption that the measurement instrument and the underlying psychological construct behave similarly and have the same significance over time (e.g., Eisenberg & Strayer, 1987; Sallquist, Eisenberg, Spinrad, Eggum, & Gaertner, 2009). From a methodological standpoint, however, this assertion is untenable if longitudinal measurement invariance (LMI) is not tested (Millsap & Cham, 2012). Indeed, in many cases, LMI cannot be established because the same instrument can measure a different construct at different ages (especially given the rapid transitions that take place in the empathy domain in early childhood). If the data do not support LMI, or if this invariance has not been tested, the validity of the inferences and interpretations drawn from the data could be erroneous or unfounded (Byrne, 2012). Therefore, an essential prerequisite for studying constructs across time is to establish measurement invariance (Coertjens, Donche, De Maeyer, Vanthournout, & Van Petegem, 2012; Little, 2013; Newsom, 2015).

The aim of this study was to validate the EmQue in a sample of Spanish children and examine its longitudinal invariance. To do so, we took two longitudinal measures in a group of 3-year-old children. Categorical confirmatory factor analysis and LMI were used to confirm the tridimensional structure of the instrument and its longitudinal invariance. We also examined the internal consistency of the EmQue over time. In addition, children's parents filled out an ER scale at the same time points. Thus, we evaluated possible relationships between the EmQue and the ER measure over time. Related to this latter point, and based on the results from Rieffe et al. (2010), we expected to find positive correlations between children's capacity for ER and the AOF and PA subscales.

Method

Participants

The participants in the present study were parents of 103 children (50 girls, 53 boys) in the first (T1) and second (T2) years of preschool. Children had a mean age of 41.76 months (range = 36–47) and 51.65 months (range = 46–57) at T1 and T2, respectively. Most of the children were of Spanish origin (97.09% Spanish, 1.94% German, and 0.97% Ecuadorian). They attended two nursery schools from two different boroughs in northern and eastern Madrid (Spain). The schools were private. The populations attending these schools belong to the middle or upper middle social class. The families that reside in both towns have an average income at the mean Spanish income levels. Attrition rates were 16.5% at T1 (n = 86, M_{age} = 41.71 months, range = 36–47) and 9.03% at T2 (n = 93, M_{age} = 51.58, range = 46–57).

Measures

EmQue for children (Rieffe et al., 2010)

The EmQue is a parent-report questionnaire that measures empathy-related behaviors in preschool children. "It consists of 20 items representing three dimensions: Emotion contagion (EC, e.g., "When another child is upset, my child needs to be comforted too"); Attention to others' feelings (AOF, e.g., "My child looks up when another child laughs"); and Prosocial actions (PA, e.g., "When another child gets frightened, my child tries to help him/her")." In the original Rieffe version of the EmQue, the EC (1, 4, 7, 10, 13, 16, and 19) and AOF (3, 6, 9, 12, 15, 18, and 20) subscales contain seven items each, and the PA subscale (2, 5, 8, 11, 14, and 17) contains six items. These three subscales displayed satisfactory internal consistency: EC: α = .58, AOF: α = .71, PA: α = .80. The Italian version of the EmQue (Grazzani et al., 2016) omitted seven items from the original questionnaire (1, 2, 11, 16, 18, 19, and 20). These items were removed because of poor performance (i.e., factor loading less than .30, excessive mean standardized residual covariances or high correlated error covariances). Thus, the Italian version consists of only 13 items: The EC and PA subscales have four items each, whereas the AOF subscale has five items. Cronbach's alphas for these subscales were close to Rieffe's: EC: α = .73, AOF: α = .74, PA: α = .80. Further information about the validity of the Dutch and Italian versions of the EmQue is available in Rieffe et al. (2010) and Grazzani et al. (2016).

Emotion Regulation Checklist (ERC; Shields & Cicchetti, 1997)

The ERC is a 24-item other-report measure that can be completed by parents and/or teachers. Raters use a 4-point scale to indicate how often a child displays affective behaviors (1 = never, 2 = sometimes, 3 = often, 4 = almost always). The Emotion Regulation (ER) subscale captures key processes in adaptive regulation, including socially appropriate emotional displays and empathy. This subscale contains eight items, such as "is a cheerful child" or "responds positively to neutral or friendly overtures by peers." High scores indicate a greater capacity for ER. We obtained each child's score by adding the scores on all of the items together, making allowances for the items that required reverse scoring. The maximum score was 32, and the minimum was 8. The Lability-negativity subscale assesses arousal, reactivity, anger dysregulation, and mood lability. Internal consistencies, assessed with Cronbach's alpha, were .96 for Lability-negativity and .83 for ER (see Shields & Cicchetti, 1997, for further information about the psychometric properties of the ERC). In the present study, only the ER subscale was used, with alpha values of .55 and .59 at T1 and T2, respectively.

Procedure

Participants were taking part in a longitudinal study of socioemotional development during the preschool years. The research team informed the principals of the two schools about the study objective, and the parents signed the appropriate consent forms. Parents completed questionnaires at home after receiving information from the research team. Parents had 10–15 days to return all of the questionnaires at each wave. Both the university and the schools followed the protocols for applying the ethical procedures that regulate research in Spain.

Data analysis

First, we established a baseline model of the EmQue. We used categorical confirmatory factor analysis to assess the fit of two different three-factor models: the Rieffe et al. (2010) and Grazzani et al. (2016) models, with 20 and 13 items, respectively. Factor covariances were added in both three-factor solutions.

Second, we tested the LMI of the resulting baseline model to find out whether the EmQue behaves consistently at two measurement time points in childhood (T1 and T2). To test for LMI, we used a single-group confirmatory factor analysis framework, in which two waves of data were combined into a single input. In this model, the same indicators were allowed to correlate over time. We tested three types of invariance hypotheses by successively setting the equality of the parameters of the measurement model at the two time points: configural, weak, and strong invariance (Coertjens et al., 2012; Liu, Millsap, West, Tein, & Tanaka, 2017). For the least restrictive model (configural invariance), manifest indicators (items) were constrained to load on the same factor over time. Next, factor loadings were constrained to be equal over time (weak or loading invariance). Finally, equality constraints on the loadings and thresholds were set to evaluate whether the items' thresholds remained the same across time (strong or threshold invariance). Confirmation of strong invariance allowed the latent means to be compared at the two time points. Because the items on the EmQue were measured using ordinal response categories, thresholds rather than intercepts were tested in this study (Millsap & Yun-Tein, 2004). The number of thresholds was equal to the number of categories minus 1, which resulted in two thresholds to be estimated.

Third, Cronbach's alpha was calculated as an estimation of the reliability of the EmQue subscales. Finally, we investigated the associations between the EmQue subscales and the ER subscale of the ERC across time using Pearson correlations.

All analyses were performed with IBM SPSS Statistics Version 22.0 (SPSS, 2013) and Mplus 7.0 (Muthén & Muthén, 2012). We used theta parameterization and the robust weighted least squares mean and variance-adjusted (WLSMV) estimator for our analyses (Newsom, 2015). As criteria for model fit, we report the comparative fit index (CFI), the Tucker–Lewis index (TLI), and root mean square error of approximation (RMSEA). Hu and Bentler (1999) suggested that RMSEA should be ≤.06 for good model fit, and CFI and TLI should be greater than .95, although any value ≥.90 tends to be considered acceptable.

In testing the invariance hypothesis, we considered several fit indices to determine fit equivalence. First, we tested a scaled chi-square difference (i.e., $\Delta WLSMV\chi^2$) between two nested models using the DIFFTEST option in Mplus. Second, changes in the model fit of CFI (ΔCFI) and RMSEA (ΔRMSEA) were used. A decrease (≥.01) in CFI and an increase (≥.015) in RMSEA revealed an unacceptable decrease in model fit (Chen, 2007; Little, 2013).

Missing data imputation

Given the small number of participants who completed the assessments at both time points (n = 76, 73.79%), and taking into account the fact that both attrition rates were less than 20% (16.5% and 9.03% at T1 and T2, respectively; Arbuckle, 1996), we followed Little's (2013) recommendations for handling missing values in longitudinal structural equation modeling and performed multiple imputation of missing data at both T1 and T2 before examining LMI. To do so, we first used Little's test to check that the data were missing completely at random (MCAR) at both time points. This first step is essential in selecting the proper imputation method. Thus, missing value patterns were analyzed, and Little's MCAR test was performed using the IBM SPSS Statistics Version 22.0 Missing Data Module (SPSS, 2013). We concluded that the data were MCAR because the test was nonsignificant (χ^2 = 36.42, df = 40, p = .636). Under the MCAR assumption, unbiased estimates of all model parameters and tests of model fit can be achieved through complete case analysis (listwise deletion) or available case analysis (pairwise deletion; i.e., the default missing treatment for WLSMV in Mplus). However, the main disadvantage of these strategies is that the statistical power of the nested model tests (e.g., LMI tests) is low (Liu et al., 2017), especially in small sample sizes, and multiple imputation of missing data is recommended.

Next we conducted multiple imputation of missing values for both waves with the IBM SPSS Statistics Version 22.0 Missing Data Module (SPSS, 2013) using the fully conditional specification method. Five imputed data sets were created and analyzed for each wave using different types of variables: (a) sociodemographic variables (i.e., children's gender, chronological and developmental ages), (b) children's language development (provided by the children through an individual test battery), (c) children's ER (provided by teachers), and (d) children's social competence (provided by parents and teachers).

Results

Descriptive statistics for EmQue scores for the two time periods

Table 1 shows the means and standard deviations for the EmQue items at the two time points for the complete and imputed data. At both time points, Item 3 ("When my child sees other children laughing, he/she starts laughing too") had the highest mean score, whereas Item 19 ("When another child cries, my child looks away") had the lowest mean score.

Baseline model of the EmQue

Two possible solutions (the Rieffe et al. and Grazzani et al. models) were fit to the two waves of data. The upper section of Table 2 shows the resulting fit indices. Both models performed better at T1 in terms of CFI, TLI, and RMSEA. Only the three-factor, 13-item model (Grazzani et al., 2016) showed an adequate fit. Accordingly, this model was chosen as the baseline model for further assessment of LMI. In addition, as the note to Table 2 states, the results for both models at the two time points were very similar for the complete and imputed data.

LMI

To establish LMI, we used the imputed data. We first started with the least restrictive model (configural invariance), which constrained manifest indicators (items) to load on the same factor across time. As Table 2 reveals, this model achieved a good model fit according to RMSEA. However, CFI and TLI were close to standard cutoffs, which suggested that caution should be used. Then, factor loadings were constrained to be equal over time (weak or loading invariance). Again, this more restrictive model also achieved an acceptable model fit according to RMSEA, although the CFI and TLI values were slightly below .90 (see Table 2). In addition, ΔWLSMVχ^2 was not significant.

Table 1. Descriptive statistics for the EmQue at T1 and T2, complete and imputed data.

Item	T1 (N = 86)	T2 (N = 93)	T1[a] (N = 103)	T2[a] (N = 103)
1	1.70 (.69)	1.81 (.63)	1.71 (.69)	1.83 (.63)
2	1.87 (.73)	2.09 (.62)	1.87 (.73)	2.10 (.64)
3	2.64 (.59)	2.68 (.51)	2.61 (.58)	2.69 (.50)
4	1.26 (.46)	1.29 (.54)	1.27 (.46)	1.29 (.54)
5	2.22 (.60)	2.25 (.60)	2.24 (.61)	2.24 (.61)
6	2.66 (.57)	2.57 (.52)	2.66 (.54)	2.55 (.52)
7	1.27 (.54)	1.24 (.43)	1.29 (.53)	1.24 (.42)
8	2.07 (.66)	2.17 (.54)	2.05 (.67)	2.16 (.56)
9	2.47 (.65)	2.31 (.66)	2.42 (.67)	2.31 (.66)
10	1.20 (.40)	1.16 (.42)	1.20 (.38)	1.18 (.43)
11	1.87 (.67)	2.10 (.63)	1.86 (.66)	2.12 (.63)
12	1.92 (.75)	1.99 (.70)	1.92 (.75)	2.03 (.69)
13	1.33 (.60)	1.16 (.37)	1.38 (.61)	1.18 (.38)
14	1.28 (.50)	1.42 (.56)	1.28 (.48)	1.42 (.56)
15	2.45 (.52)	2.34 (.58)	2.45 (.52)	2.34 (.58)
16	1.71 (.76)	1.75 (.69)	1.69 (.74)	1.75 (.69)
17	1.98 (.65)	2.17 (.64)	1.97 (.66)	2.17 (.64)
18	1.83 (.69)	1.80 (.65)	1.82 (.67)	1.80 (.65)
19	1.16 (.37)	1.13 (.34)	1.21 (.39)	1.13 (.34)
20	2.07 (.72)	2.28 (.63)	2.09 (.72)	2.28 (.63)

Note. Data are *M* (*SD*). EmQue = Empathy Questionnaire; T1 = Time 1; T2 = Time 2.
[a]Imputed data.

Table 2. Model fit indices for the baseline model and longitudinal measurement invariance of the Grazzani et al. model.

Model	WSLMVχ^2	df	CFI	TLI	RMSEA (90% CI)	ΔWSLMVχ^2(Δdf)	ΔCFI	ΔRMSEA
T1 (*N* = 86)[a]								
Three factors, 20 items	301.7	167	.786	.756	.097 (.079, .114)			
Three factors, 13 items	96.7	62	.931	.914	.070 (.036, .099)			
T2 (*N* = 93)[b]								
Three factors, 20 items	277.0	167	.743	.708	.084 (.066, .101)			
Three factors, 13 items	103.0	62	.941	.935	.064 (.025, .087)			
Model (*N* = 103)								
Configural invariance	332.4	271	.885	.862	.047 (.026, .063)			
Weak invariance	337.4	281	.894	.878	.044 (.022, .061)	8.1 (10)	.009	.003
Strong invariance	358.4	301	.894	.886	.043 (.021, .059)	22.8 (20)	.000	.001

Note. WSLMVχ^2 = weighted least squares mean and variance-adjusted chi-square; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; CI = confidence interval; T1 = Time 1; three factors, 20 items = Rieffe et al. model; three factors, 13 items = Grazzani et al. model; T2 = Time 2.
[a]Solution for the imputed data: three factors, 20 items: WSLMVχ^2 (*df* = 167) = 273.6, CFI = .769, TLI = .737, RMSEA (90% CI) = .079 (.062, .095); three factors, 13 items: WSLMVχ^2 (*df* = 62) = 95.4, CFI = .930, TLI = .911, RMSEA (90% CI) = .063 (.030, .087).
[b]Solution for the imputed data: three factors, 20 items: WSLMVχ^2 (*df* = 167) = 282.2, CFI = .716, TLI = .697, RMSEA (90% CI) = .082 (.065, .198); three factors, 13 items: WSLMVχ^2 (*df* = 62) = 86.6, CFI = .917, TLI = .905, RMSEA (90% CI) = .064 (.026, .093).

Furthermore, ΔCFI and ΔRMSEA did not exceed the cutoff values, which suggested that a weak invariance model was obtained. Finally, the most restrictive model, in which thresholds of items were also constrained to be equal across the two time points, also yielded an adequate fit according to the RMSEA value. Compared to the weak invariance model, the strong invariance model did not represent a statistically significant reduction in model fit (ΔWLSMVχ^2 was not significant, and both ΔCFI and ΔRMSEA were below the cutoff values). Based on these results, one might conclude that strong invariance would be maintained over time with regard to empathy. However, this conclusion must be considered with caution, as CFI and TLI were under the .90 cutoff criterion.

Table 3 presents correlations between the EmQue dimensions. Examination of these cross-scale associations indicates stronger correlations between the same factors than between different factors.

Table 3. Reliability and correlations between EmQue factors and the ER subscale at two time points.

Time and Variable	1	2	3	4	5	6	7	8
1. T1 EC	(.58)							
2. T1 AOF	.29**	(.64)						
3. T1 PA	.08	.16	(.83)					
4. T1 ER	.10	.17	.41***	(.55)				
5. T2 EC	.39***	.27*	.06	.06	(.60)			
6. T2 AOF	.32**	.44***	.16	.12	.27*	(.62)		
7. T2 PA	.14	.07	.38***	.33**	.23*	.24*	(.75)	
8. T2 ER	.07	.29**	.03	.47**	.08	.32**	.37**	(.59)

Note. Values in parentheses are reliabilities. EmQue = Empathy Questionnaire; ER = emotional regulation; T1 = Time 1; EC = emotion contagion; AOF = attention to others' feelings; PA = prosocial actions; T2 = Time 2.
*p < .05. **p < .01. ***p < .001.

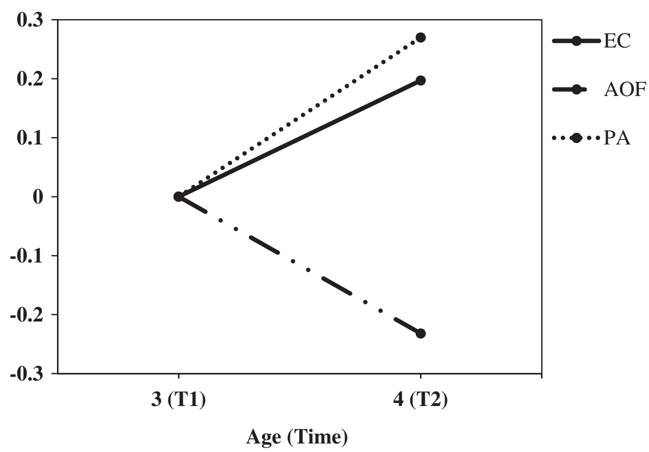

Figure 1. Empathy Questionnaire three-factor means across time. Mean estimates are from the strong invariance model. The initial measurement point at age 3 (T1) was used as a reference, with a factor mean of zero. EC = emotion contagion ; AOF = attention to others' feelings; PA = prosocial actions; T1 = Time 1; T2 = Time 2.

Latent mean differences

Latent mean differences across the two time points were estimated from the strong invariance model. Figure 1 presents the mean-level estimates of the three latent factors (EC, AOF, and PA) with the first measurement time (T1) as the reference; that is, T2 was scaled relative to T1. The results suggested that AOF scores decreased slightly with time, whereas EC and PA scores increased with age. These differences were statistically significant only for PA (d = .27, p < .05). However, these differences must be viewed with caution because of the marginal fit of the invariance models.

Reliability estimate of the EmQue subscales at the two time points

To examine the reliability of the three subscales at the two different points in time, we computed Cronbach's alpha coefficients. Scale score reliabilities of the EmQue subscales were acceptable, ranging from α = .60 (EC) to α = .83 (PA). These values were similar to those obtained by Grazzani et al. (2016) and Rieffe et al. (2010).

Relationship with ER

Correlation analyses were conducted to examine the relationship between the three EmQue subscales and the ER subscale from the ERC. Table 3 provides the Pearson correlations. At each point in time, ER was significantly positively associated with the PA factor. In addition, ER at T1 was significantly positively related to AOF at both time points.

Discussion

The main objective of this study was to validate the tridimensional structure and examine the longitudinal invariance of the EmQue in a sample of Spanish children through a longitudinal study. Furthermore, we wanted to explore whether the AOF and PA dimensions have any relationship with ER and whether these relationships remain over time. Our results confirm the tridimensional structure of the scale, in agreement with the results of Grazzani et al. (2016), and they provide similar acceptable reliability indices to those found by Rieffe et al. (2010) and Grazzani et al. Regarding correlations with the ER measure, the results show significant positive correlations between ER and the PA subscale at each moment in time. A positive correlation was also found between ER at T2 and the AOF subscale at both T1 and T2.

As Grazzani et al. (2016) found with the Italian version, the Spanish version provides evidence for the three-factor scale structure based on the same 13 items. In spite of the age differences between the Italian and Spanish participants, the confirmation of Grazzani et al.'s 13-item, three-factor model reveals the robust factorial structure of the EmQue with these items. Moreover, although the CFI and TLI values were slightly below the .90 cutoff criterion, the RMSEA values seemed to support the measurement invariance of this model across time. Given that the few studies that have evaluated longitudinal invariance have not always evaluated the same attribute (De Frias & Dixon, 2005; Millsap & Cham, 2012; Molt, Dishman, Birnbaum, & Lytle, 2005), this result is especially important for the scale's validation. In addition to the Grazzani et al. results on the invariance of the Italian version of the EmQue based on gender, our results contribute new evidence of the validity of the EmQue. After testing the LMI, we examined latent means of the three subscales across time and found that PA scores increased with age. Although these trends should be better determined using latent growth curve analysis (Little, 2013; Newsom, 2015), this result coincides with the positive correlations between PA and age found by Rieffe et al. (2010) and Grazzani et al.

Another interesting result has to do with the correlations found between ER and the AOF and PA subscales. The significant positive correlations found at the two moments in time show the clear relationship between these variables. When considering the longitudinal correlations, we observed a special relationship between ER and PA that was maintained over time. However, this relationship is not as clear between ER and AOF at T1, and we found no significant relationships between ER and EC. Based on the developmental model by Hoffman (1987), and given the ages of the children in the sample, it is likely that prosocial behavior is the element of empathy that is more clearly expressed across time and also in relation to ER. The Russian doll model (De Waal, 2008), which proposes that the earliest components of empathy are necessary prerequisites for the development of the next ones, could explain this result. From our point of view, this model is compatible with the levels described by Hoffman. Thus, children develop empathy through the different components, with each component being a prerequisite for the next one. However, prosocial behavior is special because it provides a clear sign of empathy. Children who act prosocially are undoubtedly empathic, but those who just pay attention to others' feelings are not necessarily empathic. Therefore, the prosocial level could be understood as true empathy (i.e., measurable, observable, and trainable), which requires other skills related to emotional competence. Of the three empathy factors, prosocial behavior could be the doll that covers, includes, and is more visible; therefore, it is more clearly related to other elements of emotional competence, such as ER. From this point of view, we can suggest that prosocial behavior could be the key element in empathy assessment and intervention. The relevance of this variable was revealed in previous studies showing that prosocial behavior at the beginning of childhood is one of the best predictors of academic and social trajectory in adolescence (Caprara, Barbaranelli, Pastorelli, Bandura, & Zimbardo, 2000).

Finally, two results were expected that were not found. ER and EC should have shown a negative correlation at both time points. Likewise, ER and AOF should have been significantly positively related at T1. Although we cannot explain why we did not find these two results, it might be necessary to use other complementary measures (i.e., observation of the child or an interview with the parents) in order to more precisely establish the nature of these relationships. The use of other measures is especially important if we take into account the low internal consistency values of the ER subscale obtained in the present study compared to those obtained in other studies (Reis et al., 2016; Shields & Cicchetti, 1997).

From an applied point of view, as mentioned previously, empathy and PA are essential in the development and psychological adjustment of children and adults. Therefore, reliable and validated measurement instruments are needed to evaluate these skills from the beginning of childhood. These instruments would allow experts to identify children with difficulties in this area and rate the efficacy of interventions designed to improve levels of empathy and prosocial behavior during early development.

The results of this study must be interpreted with some limitations in mind. On the one hand, it is necessary to increase the sample size and overcome the usual limitations of sample attrition in longitudinal studies. In fact, in this study we had to carry out multiple imputation of missing values in order to perform the LMI analyses. In addition, the internal consistency of some of the study measures did not meet the minimum standards for reliability research, and this aspect should be taken into account. On the other hand, future studies should extend the follow-up time and, given the lack of measurement instruments, adapt the questionnaire for use with elementary school children. From an applied point of view, it would be desirable to have reliable measurement instruments throughout preschool and elementary school. Given the relevance of empathy and prosocial behavior in children's social and academic lives, these types of measures could contribute to designing specific interventions or prevention programs for bullying behaviors. Another future line of research would involve assessing other variables related to social and emotional development (i.e., emotional knowledge, theory of mind, etc.) in order to examine their possible relationships with empathy and prosocial behavior. Finally, it would be interesting to evaluate the instrument's validity in other cultural contexts. In this regard, further studies should test the LMI of the EmQue in larger and different samples in order to find out whether sample size, participant age, or cultural and language characteristics could have had an influence on the low goodness-of-fit indices of the resulting LMI models.

Despite its limitations, the present study provides evidence for the construct validity and equivalence across time of the 13-item, three-factor structure of the EmQue for preschool-age children, making this instrument a useful tool for studying this population. It is worth noting that, to our knowledge, no previous studies have examined the LMI of the EmQue. In addition, this study also provides important insights into the development of the relationships among empathy, prosocial behavior, and ER in childhood. Nonetheless, further studies are needed to explore and confirm the dimensionality and equivalence of the EmQue across time in larger samples and other cultural contexts.

Funding

This study received funding through research projects EDU2013-45181-R and GV2016-136.

References

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243–277). Mahwah, NJ: Erlbaum.

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York, NY: Routledge.
Caprara, G. V., Barbaranelli, C., Pastorelli, C., Bandura, A., & Zimbardo, P. G. (2000). Prosocial foundations of children's academic achievement. *Psychological Science*, *11*, 302–306.
Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*, 464–504.
Coertjens, L., Donche, V., De Maeyer, S., Vanthournout, G., & Van Petegem, P. (2012). Longitudinal measurement invariance of Likert-type learning strategy scales: Are we using the same ruler at each wave? *Journal of Psychoeducational Assessment*, *30*, 577–587.
Cox, C. L, Uddin, L. Q, Di Martino, A, Castellanos, F. X, Milham, M. P, & Kelly, C. (2012). The balance between feeling and knowing: affective and cognitive empathy are reflected in the brain's intrinsic functional dynamics. *Social Cognitive And Affective Neurosciences*, *7*, 727–737.
De Frias, C. M., & Dixon, R. A. (2005). Confirmatory factor structure and measurement invariance of the memory compensation questionnaire. *Psychological Assessment*, *17*, 168–178.
De Waal, F. B. (2008). Putting the altruism back intro altruism: The evolution of empathy. *Annual Review of Psychology*, *59*, 279–300.
Eisenberg, N. (2000). Emotion, regulation and moral development. *Annual Review of Psychology*, *51*, 665–697.
Eisenberg, N., & Strayer, J. (1987). *Empathy and its development. Cambridge studies in social and emotional development*. New York, NY: Cambridge University Press.
Grazzani, I., Ornaghi, V., Pepe, A., Brazzelli, E., & Rieffe, C. (2016). The Italian version of the Empathy Questionnaire for 18- to 36-months-old children: Psychometric properties and measurement invariance across gender of the EmQue-I13. *European Journal of Developmental Psychology*. Advance online publication. doi:10.1080/17405629.2016.1140640
Hoffman, M. L. (2000). *Empathy and moral development*. Cambridge, UK: Cambridge University Press.
Hoffman, M. L. (1987). The contribution of empathy to justice and moral judgment. In N. Eisenberg & J. Strayer (Eds.), *Empathy and its development* (pp. 47–80). Cambridge, UK: Cambridge University Press.
Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.
Jolliffe, D., & Farrington, D. P. (2006). Examining the relationship between low empathy and bullying. *Aggressive Behaviour*, *32*, 540–550.
Kerem, E., Fishman, N., & Josselson, R. (2001). The experience of empathy in everyday relationships: Cognitive and affective elements. *Journal of Social and Personal Relationships*, *18*, 709–729.
Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford Press.
Liu, Y, Millsap, R. E, West, S. G, Tein, J. Y, Tanaka, R, & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods, 22*, 486.
Millsap, R. E., & Cham, H. (2012). Investigating factorial invariance in longitudinal data. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 109–126). New York, NY: Guilford Press.
Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial-invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*, 479–515.
Molt, R. W., Dishman, R. K., Birnbaum, A. S., & Lytle, L. E. (2005). Longitudinal invariance of the Center for Epidemiological Studies–Depression scale among girls and boys in middle school. *Educational and Psychological Measurement*, *65*, 90–108.
Muthén, L. K., & Muthén, B. O. (2012). *Mplus: Statistical analysis with latent variables: User's guide* (7th ed.). Los Angeles, CA: Authors.
Newsom, J. T. (2015). *Longitudinal structural equation modeling: A comprehensive introduction*. New York, NY: Routledge.
Reis, A. H., Oliveira, S. E. S. D., Bandeira, D. R., Andrade, N. C., Abreu, N., & Sperb, T. M. (2016). Emotion Regulation Checklist (ERC): Estudos preliminares da adaptação e validação para a cultura brasileira. *Temas Em Psicología*, *24*, 77–96.
Rieffe, C., Ketelaar, L., & Wiefferink, C. H. (2010). Assessing empathy in young children: Construction and validation of an Empathy Questionnaire (EmQue). *Personality and Individual Differences*, *49*, 362–367.
Sallquist, J., Eisenberg, N., Spinrad, T. L., Eggum, N. D., & Gaertner, B. M. (2009). Assessment of preschoolers' positive empathy: Concurrent and longitudinal relations with positive emotion, social competence, and sympathy. *Journal of Positive Psychology*, *4*(3), 223–233.
Shamay-Tsoory, S. G. (2011). The neural bases for empathy. *Neuroscientist*, *17*, 18–24.
Shields, A., & Cicchetti, D. (1997). Emotion regulation among school-age children: The development and validation of a new criterion Q-sort scale. *Developmental Psychology*, *33*, 906–916.
SPSS. (2013). *IBM SPSS statistics for Windows: Version 22.0*. Armonk, NY: IBM.
Zahn-Waxler, C., Radke-Yarrow, M., Wagner, E., & Chapman, M. (1992). Development of concern for others. *Developmental Psychology*, *28*, 126–136.