education sciences

MDPI

Article

Broadening the Assessment of Mathematics-Related Achievement Emotions: Structure and Differentiation of an Extended AEQ-ES Among Chilean Elementary School Children

Cristina Rodríguez [1,*], Roberto A. Ferreira [2], Bárbara Guzmán [1], Felipe Sepúlveda [1], Cristian A. Rojas-Barahona [3], Marcos H. Cárdenas-Mancilla [4,5], Andrea Casanova [6] and Isaac Marco [7]

1 Faculty of Education, Universidad Católica de la Santísima Concepción, Concepción 4090541, Chile; barbara.guzman@ucsc.cl (B.G.); fsepulveda@ucsc.cl (F.S.)
2 Faculty of Education Sciences, Science of Learning Lab, Universidad de Talca, Linares 3582979, Chile; roberto.ferreira@utalca.cl
3 Faculty of Psychology, Research Center on Cognitive Sciences, Universidad de Talca, Talca 3460000, Chile; c.rojas@utalca.cl
4 Center for the Well-Being and Development of Adolescents and Children in the Digital Age (BAND), Pontificia Universidad Católica de Chile, Santiago 8331150, Chile; marcos.cardenas@iie.cl
5 Institute of ICT in Education, Universidad de La Frontera, Temuco 4811230, Chile
6 Faculty of Psychology and Juan Ignacio Molina Institute of Humanistic Studies, Universidad de Talca, Talca 3460000, Chile; andrea.casanova@utalca.cl
7 Department of Computer Engineering and Systems, Universidad de La Laguna, 38200 San Cristóbal de La Laguna, Spain; imarcoro@ull.edu.es
* Correspondence: mrodriguez@ucsc.cl

Abstract

Achievement emotions play a central role in mathematics learning, yet instruments suitable for young children often cover only a narrow range of emotions and have rarely been validated in Latin American contexts. This study examined the structure, differentiation, and measurement properties of mathematics-related achievement emotions using an extended Spanish-language version of the Achievement Emotions Questionnaire–Elementary School (AEQ-ES). The instrument, referred to here as the AEQ-ES-M, was extended to include anger and hope, alongside the original dimensions of enjoyment, anxiety, and boredom. Items were translated and back-translated, and semantic equivalence was assessed using Sentence-BERT embeddings (cosine similarities = 0.70–0.96). A new set of facial icons depicting a boy or a girl was developed for all five emotions, and their recognizability was examined in an independent face-matching task with 399 second-graders (overall accuracy = 79.6%). The 46-item questionnaire was administered to 514 Chilean children in Grades 2 (n = 247) and 4 (n = 267) attending nine schools. At the full-sample level, confirmatory factor analyses supported a five-factor correlated-emotions model over an alternative two-factor valence model. Reliability was high across emotion scales (ω = 0.894–0.937), and convergent validity was adequate (average variance extracted, AVE = 0.538–0.726). However, substantial overlap emerged between enjoyment and hope, particularly in Grade 2 (r = 0.983), where the five-factor solution was inadmissible. A post hoc four-factor model combining these emotions showed similar approximate fit (change in comparative fit index, ΔCFI = 0.002). Scalar invariance across sex held for both models, whereas scalar invariance across grade was supported only for the four-factor model. Girls reported higher mathematics anxiety than boys, and, in this cross-sectional comparison, fourth graders showed lower levels of the combined enjoyment–hope factor and higher anger, anxiety, and boredom than second graders. Overall, the AEQ-ES-M provides a reliable and culturally adapted measure of a broader range of achievement emotions in mathematics among Chilean primary school

Check for updates

Academic Editors: Roberto Burro, Daniela Raccanello and Joana Pipa

Received: 20 September 2026
Revised: 11 October 2026
Accepted: 15 October 2026
Published: 20 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

children. Because enjoyment and hope were less clearly differentiated in Grade 2, the four-factor specification is recommended for this grade and for cross-grade comparisons.

Keywords: achievement emotions; mathematics; psychometric validation; measurement invariance; Control-Value Theory

1. Introduction

Research on mathematics learning has traditionally focused on cognitive processes, with comparatively less attention to the affective dimension (Forsblom et al., 2022; Schukajlow et al., 2023). Historically, research within this domain has focused largely on mathematics anxiety, one of the most extensively studied emotions in relation to mathematics performance (Pekrun, 2006). More recently, research has examined a broader spectrum of emotions, differentiated by valence and activation level (Camacho-Morles et al., 2021; Schoenherr et al., 2025; Xie et al., 2025). Mathematics seems to be a domain of particular relevance for the study of achievement emotions. Meta-analytic evidence indicates that some achievement emotions, particularly enjoyment and anger, are more strongly associated with academic achievement in mathematics than in other subject domains (Camacho-Morles et al., 2021).

According to the Control-Value Theory (CVT; Pekrun, 2006), achievement emotions emerge from cognitive appraisals of perceived control and value. The theory classifies them along three dimensions: valence (positive vs. negative), activation (activating vs. deactivating), and object focus (activity-related vs. outcome-related; Pekrun et al., 2023). Within this taxonomy, enjoyment and hope are positive activating emotions, anxiety and anger are negative activating emotions, and boredom is a negative deactivating emotion. Emotions that share valence and activation differ in object focus: enjoyment and anger relate to the ongoing activity, whereas hope and anxiety are prospective emotions directed at a future outcome: anticipated success in the case of hope and possible failure in the case of anxiety (Pekrun, 2006). CVT further posits that achievement emotions are context-specific: they vary across class, homework, and test settings because each context presents different demands and social dynamics that shape students' perceptions of control and value (Pekrun & Stephens, 2010).

Recent evidence highlights the important role of academic emotions in learning outcomes and the need to examine a broad range of emotions. In a meta-analysis of studies published between 2000 and 2024, Xie et al. (2025) reported statistically significant associations between academic emotions and performance (positive affect: $r = 0.24$; negative affect: $r = -0.25$). The authors also found that culture and location moderated the emotion–performance relationship, with stronger effects in Asian and European samples than in North American samples. Complementing these findings, a systematic review and meta-analysis of mathematics achievement reported that mathematics achievement is positively related to enjoyment ($\beta = 0.247$) and hope ($\beta = 0.224$) and negatively related to anger ($\beta = -0.322$) and boredom ($\beta = -0.187$). These findings support the inclusion of a broader range of positive and negative emotions beyond the traditional focus on mathematics anxiety (Schoenherr et al., 2025).

Research on achievement emotions has relied heavily on the Achievement Emotions Questionnaire (AEQ; Pekrun et al., 2011) and its variants, which have been primarily validated with university students, limiting their use with younger populations. To address this gap, the AEQ-Elementary School (AEQ-ES; Lichtenfeld et al., 2012) was developed for primary school children with simplified item wording and a 5-point pictorial Likert scale

featuring drawings of boys' and girls' faces. However, the AEQ-ES assesses only enjoyment, anxiety, and boredom, leaving out theoretically and empirically relevant activating emotions such as hope and anger (Camacho-Morles et al., 2021; Schoenherr et al., 2025). Domain-specific adaptations have extended the AEQ framework to older students. For example, the AEQ-Mathematics (AEQ-M; Bieleke et al., 2022) covers seven emotions across three contexts in secondary school students, while the AEQ-Pre-Adolescents (AEQ-PA; Peixoto et al., 2015) adapted the instrument for students aged 10 to 13.

Collectively, these instruments present a persistent trade-off: the AEQ-ES is developmentally appropriate for primary school but restricted to three emotions, while broader adaptations were validated with older populations. These limitations leave open how clearly a broader range of achievement emotions can be distinguished in younger children's self-reports. Evidence addressing this question in Latin American primary-school populations remains limited.

1.1. Previous Evidence on the Factor Structure of the AEQ in Mathematics

The AEQ-ES assesses enjoyment, anxiety, and boredom across three academic settings: class, homework, and tests, with the exception of test-related boredom. The instrument has a hierarchical factor structure in which eight first-order factors (emotion × setting) are nested under three second-order emotion factors. The original validation by Lichtenfeld et al. (2012) provided support for this hierarchical structure in German samples of second (N = 594) and third grade (N = 595) students and in an American sample of third-grade students (N = 163). Model fit was adequate in the German samples and somewhat weaker in the American sample. In all samples, the hierarchical model showed a better fit than a two-factor valence model and a three-factor correlated-emotion model. Importantly, for German second-graders, test-related items were not included because they were not considered relevant at that developmental stage. Students had limited experience with formal testing and did not yet receive grades. This adaptation shows that the AEQ-ES may need to reflect the academic situations that are actually relevant to children at a given age and in a particular educational context.

Subsequent studies have provided further support for the hierarchical structure across samples and cultural contexts. In an Italian sample of 767 second- and fourth-grade students, Raccanello et al. (2019) replicated this structure. Using additional data from the German and American samples included in the original validation, they also found evidence of measurement invariance across countries. Within the Italian sample, measurement invariance was further supported across grades, sex, and academic domains. Raccanello et al. (2022) also provided additional support for the hierarchical measurement structure in Italian second- and fourth graders. Nevertheless, despite overall support for the hierarchical structure, some studies have found difficulties in distinguishing between theoretically distinct emotions. A Portuguese replication by Sanches et al. (2020; N = 350, Grades 3–4) confirmed the hierarchical structure but reported compromised discriminant validity between enjoyment and boredom. Specifically, their maximum shared variance was 0.90, implying an association of approximately r = 0.95, despite the two emotions being retained as theoretically distinct constructs. Taken together, existing studies generally support the hierarchical organization of the original three-emotion AEQ-ES, although overall model fit has tended to be acceptable rather than consistently strong, with CFI values typically remaining below the more stringent 0.95 heuristic. More importantly, some evidence suggests that theoretically distinct emotions may not always be clearly differentiated, particularly enjoyment and boredom, which have shown substantial overlap in some studies.

Evidence from studies extending the AEQ beyond the original three-emotion framework suggests that distinguishing among theoretically distinct emotions may become more

challenging when a broader range of emotions is assessed. Peixoto et al. (2015) initially assessed a broader set of eight achievement emotions in Portuguese students in Grades 5 and 7 (N = 1515). Preliminary CFAs yielded poor fit or identification problems and revealed latent correlations exceeding 0.90 between hope and enjoyment, hope and pride, and anxiety and shame. Hope and shame were consequently excluded because they could not be adequately differentiated from closely related emotions. Following these exclusions, six correlated emotion factors were examined separately within class and test contexts and provided better fit than a broad positive-versus-negative valence model. Importantly, however, Peixoto et al. did not test a hierarchical model integrating emotion-specific factors across contexts. Bieleke et al. (2022) provided complementary evidence on the structure of a broader set of achievement emotions in German secondary-school students. In Study 1 (N = 781, Grades 5–10), seven discrete emotions were assessed across class, learning, and test contexts. The best-fitting model represented the seven emotions as separate factors while also accounting for shared variance within class, learning, and test contexts. In Study 2 (N = 699, Grades 7 and 9), a similar structure was examined for five emotions and again performed better than simpler alternatives, although the fit indices were mixed. Together, these findings suggest that instruments assessing a broader range of emotions may not necessarily reproduce the hierarchical organization found for the original three-emotion AEQ-ES.

1.2. The Present Study

The literature reviewed above points to two unresolved issues. First, it remains unclear how well theoretical distinctions among a broader set of achievement emotions are reflected in younger children's self-reports. The broader AEQ adaptations reviewed here have focused on students in Grade 5 and above (Peixoto et al., 2015; Bieleke et al., 2022). Extending assessment beyond enjoyment, anxiety, and boredom therefore raises questions about whether younger children's reports distinguish hope and anger from other emotions. It also remains uncertain whether these reports follow the hierarchical organization of emotions across academic contexts proposed for the original AEQ-ES. Second, evidence on the structural validity of AEQ-based measures in Latin American elementary-school populations remains limited. Given that achievement emotions are embedded in sociocultural and educational contexts, findings from other populations cannot be assumed to generalize directly to Chilean children (Xie et al., 2025).

The Chilean context also gives these questions educational relevance. In PISA 2022, Chile ranked 7th among 78 education systems in mathematics anxiety (OECD, 2023). This places Chilean students among those reporting the highest levels of mathematics-related anxiety internationally. Mathematics anxiety, in turn, is consistently associated with lower mathematics performance across PISA-participating systems. Because this relationship can already be detected in the early years of schooling (Barroso et al., 2021; Cargnelutti et al., 2016), early identification of mathematics-related achievement emotions may help inform timely support. In Chile, studies have assessed mathematics anxiety as early as kindergarten and second grade (del Río et al., 2023; Guzmán et al., 2021), while other achievement emotions remain less explored at these ages. Given the relevance of emotions beyond anxiety to mathematics achievement (Camacho-Morles et al., 2021; Schoenherr et al., 2025), extending their assessment to younger children would help broaden our understanding of early emotional experiences in mathematics. Developmentally appropriate measures would also provide a basis for future longitudinal studies examining how these emotions and their relationships with achievement develop across schooling.

The present study examined the differentiation and organization of five mathematics-related achievement emotions in Chilean children in Grades 2 and 4 from different socioe-

conomic backgrounds. To this end, we evaluated a five-emotion adaptation of the AEQ-ES. The adaptation retained enjoyment, anxiety, and boredom from the original instrument (Lichtenfeld et al., 2012) and added anger and hope, which are assessed in broader AEQ instruments. Previous research has identified differences in achievement emotions across school grades (Raccanello et al., 2013), while findings on sex differences have varied across samples and contexts (Lichtenfeld et al., 2012). We therefore also examined grade and sex differences in children's reported emotions.

Specifically, we aimed to (1) examine the extent to which enjoyment, hope, anger, anxiety, and boredom are distinguishable in children's self-reports and whether their organization across academic contexts is consistent with the hierarchical structure proposed for the original AEQ-ES; (2) evaluate internal consistency and convergent and discriminant validity; (3) test measurement invariance across sex and grade levels; and (4) examine whether children's reported levels of achievement emotions differ between Grades 2 and 4 and between boys and girls, where sufficient measurement invariance supports these comparisons.

2. Materials and Methods

2.1. Participants

Participants were 514 students (262 girls, 252 boys) enrolled in Grade 2 (n = 247; 48.1%) and Grade 4 (n = 267; 51.9%). Mean age was 8.11 years (SD = 0.32) in Grade 2 and 10.10 years (SD = 0.33) in Grade 4. Students were recruited from nine schools in the Biobío and Maule regions (public: n = 130, 25.3%; government-subsidized private: n = 324, 63.0%; private: n = 60, 11.7%). School-level socioeconomic context, indexed by the School Vulnerability Index (IVE) of the National School Aid and Scholarship Board (JUNAEB), ranged from very low to very high vulnerability (Sepúlveda et al., 2020).

2.2. Materials

The AEQ-ES-M (Achievement Emotions Questionnaire–Elementary School–Mathematics) is a Spanish-language adaptation and extension of the AEQ-ES for Chilean elementary school students. The adaptation is based primarily on the AEQ-ES (Lichtenfeld et al., 2012), designed for primary school students, and the AEQ-M (Bieleke et al., 2022), which assesses a broader range of mathematics-related emotions. Enjoyment, anxiety, and boredom were adapted from the AEQ-ES, while anger and hope were primarily adapted from the AEQ-M. The Hope dimension required additional construction. The basic AEQ-M does not include a hope scale; it includes a hopelessness scale, which is assessed only in the test context. Within Control-Value Theory, hope and hopelessness are prospective outcome emotions of opposite valence that share the same object (anticipated success or failure) and are grounded in the same appraisal dimension, namely expectancies of success (Pekrun, 2006; Pekrun et al., 2011). On this basis, and to retain item content already adapted to mathematics, candidate hope items were developed by semantically adapting six of the seven AEQ-M hopelessness items into positively valenced counterparts. The resulting items are positively worded hope items scored in the direction of hope; they are not reverse-keyed hopelessness items. We did not assume, however, that hope and hopelessness are psychometric opposites, as the two emotions can co-occur. Because most of these items are worded without reference to a specific setting, with context established by the instructions (Bieleke et al., 2022), some were assigned to class or homework rather than test. To anchor the scale in established hope content, two class-related items were adapted from the class-related hope scale of the original AEQ (Pekrun et al., 2011), and one homework item was newly developed, so that hope was represented in class, homework, and test contexts. The complete item set is provided in Supplementary Table S2.

The adaptation followed a four-phase sequential procedure. In Phase 1, the items were translated and back-translated by a native English speaker and a native Chilean Spanish speaker, both proficient in the other language, following established recommendations for test adaptation (International Test Commission, 2017). In Phase 2, semantic similarity was examined using Sentence-BERT (SBERT; Reimers & Gurevych, 2019). A multilingual Sentence-Transformers model, distiluse-base-multilingual-cased-v1 (Reimers & Gurevych, 2020), was used to compare the original English and Spanish items, whereas the English-language all-MiniLM-L6-v2 model, implemented within the Sentence-Transformers framework (Reimers & Gurevych, 2019), was used to compare the original and back-translated English versions. Cosine similarity was calculated between the resulting sentence embeddings for each item. Additional psycholinguistic features were considered when reviewing item equivalence, including concreteness (Brysbaert et al., 2014), Zipf word frequency derived from SUBTLEX-US norms (Brysbaert & New, 2009), and item length. In Phase 3, a co-investigator with expertise in psycholinguistics and professional translation reviewed items with lower similarity scores and made minor grammatical and syntactic adjustments without changing their meaning. In Phase 4, the revised items were reviewed, and the final version was established. Item-level cosine similarity scores between the original English and Chilean Spanish versions ranged from 0.70 to 0.96, supporting overall semantic similarity across versions.

The final AEQ-ES-M comprises 46 items assessing five emotions: enjoyment (9 items), hope (9 items), anger (9 items), anxiety (12 items), and boredom (7 items), across class, homework, and test contexts (see Supplementary Table S2 for the complete item set in Spanish with English translations, and Supplementary Table S1 for item-level descriptive statistics). Responses were given on a pictorial 5-point scale, described in the next section.

Pictorial Response Scale and Icon Recognition

Responses were recorded on a 5-point Likert scale anchored with facial icons depicting five levels of increasing emotional intensity, following the pictorial response format of the AEQ-ES (Lichtenfeld et al., 2012). Two parallel versions of each icon set were developed, one depicting a boy's face and one depicting a girl's face. Because the AEQ-ES-M assesses five emotions, a complete set of facial icons was designed for each emotion by a professional graphic designer. The AEQ-ES faces served as the primary stylistic reference to ensure visual consistency across emotion sets. For the anger icons, the five-level intensity gradient was developed with reference to the Karolinska Directed Emotional Faces (KDEF; Lundqvist et al., 1998), which provided validated exemplars of anger expression at varying intensity levels. For the hope icons, the achievement emotion drawings from Raccanello and Burro (2024) served as the primary visual reference. The research team provided the designer with written specifications for each emotion and intensity level, and the illustrations were refined iteratively until the final sets were established. The complete icon sets are shown in Supplementary Figure S2.

To examine whether children identified the intended emotional meaning of the newly developed pictorial anchors, an independent icon-recognition check was conducted with 399 second-grade students who were not included in the main psychometric validation. Children completed a touchscreen one-to-one matching task in which they matched the five emotion labels to the corresponding facial icons. Responses were coded as correct or incorrect for each emotion, yielding a total accuracy score ranging from 0 to 5. Overall accuracy was moderate to high, $M = 3.98$, $SD = 1.22$, corresponding to 79.6% correct matches. Recognition was highest for Anger (97.7%) and Boredom (94.2%), followed by Anxiety (77.9%), Enjoyment (68.4%), and Hope (59.6%). Among children with exactly three correct matches, the most frequent error pair was Enjoyment–Hope, observed in 79 of 132 cases

(59.8%). These findings indicate that the pictorial anchors were recognizable for most emotions, whereas enjoyment and hope were less clearly discriminated. Because this check was conducted only with second-grade children, it provides no direct evidence on icon recognition among fourth graders, and its results are therefore used only as contextual information when interpreting the psychometric findings (see Section 4.1).

2.3. Procedure

Data collection was conducted by trained examiners who received standardized instructions prior to the study. Each child was assessed individually in a quiet location within the school (e.g., a separate room or low-distraction area) to minimize interruptions and ensure privacy. The AEQ-ES-M was administered on touchscreen laptops. Each item was presented in written form and accompanied by an audio recording, and children responded by touching the facial icon that best represented how they felt. The audio presentation was included to minimize reading demands, particularly for second-grade students who were still developing their reading skills. All procedures were approved by the Scientific Ethics Committee of the Universidad de Talca (Folio 19/2024) and subsequently endorsed by the Scientific Ethics Committee of the Universidad Católica de la Santísima Concepción. Written parental informed consent was obtained prior to participation, and children provided assent before completing the assessment.

2.4. Data Analysis

All analyses were conducted in R (Version 4.1; R Core Team, 2021). Given that item responses were measured on a 5-point Likert scale, indicators were treated as ordered-categorical. Prior to model estimation, item-level descriptive statistics were examined, including response distributions, category use, and missingness. Polychoric correlations were also inspected to characterize the preliminary association structure among items. Confirmatory factor analyses (CFAs) and measurement invariance analyses were conducted using the lavaan package (Rosseel, 2012). Students were nested within nine schools. Multilevel approaches (multilevel confirmatory factor analysis or cluster-robust standard errors) were not applied because nine clusters fall well below the commonly cited minimum of 30 to 50 higher-level units needed for unbiased estimation of between-level parameters and their standard errors (Maas & Hox, 2005; McNeish & Stapleton, 2016). Analyses were therefore conducted at the student level; the implications of this decision are considered in Section 4.4.

2.4.1. Construct Validity: Confirmatory Factor Analysis and Competing Measurement Models

The internal structure of the instrument was evaluated using confirmatory factor analysis (CFA) for ordered-categorical indicators. Models were estimated using diagonally weighted least squares with robust standard errors and a mean and variance-adjusted test statistic (WLSMV), as implemented in lavaan (Li, 2015; Rosseel, 2012). Three theoretically specified representations of the AEQ-ES-M structure were specified a priori: (a) a hierarchical emotion-by-context model reflecting the organization of the original AEQ-ES; (b) a five-factor model comprising correlated Enjoyment, Hope, Anger, Anxiety, and Boredom factors; and (c) a two-factor valence model distinguishing positive and negative emotions. Because the five-factor solution indicated substantial overlap between Enjoyment and Hope, an additional four-factor model combining these two emotions into a single Positive factor was examined as a post hoc analysis. Model fit was evaluated using the comparative fit index (CFI), Tucker–Lewis index (TLI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR). Models yielding improper solutions (e.g., a non-positive definite latent covariance matrix) were not interpreted, and

their fit indices are not reported because they cannot be meaningfully compared with those of admissible models.

2.4.2. Reliability and Construct Validity

Internal consistency was evaluated using McDonald's omega (ω), with Cronbach's alpha and ordinal alpha reported as complementary indices. Given the ordered-categorical nature of the items, ordinal alpha was based on polychoric correlations. Convergent validity was assessed using average variance extracted (AVE), computed from the standardized CFA solution, with values ≥ 0.50 considered indicative of adequate convergent validity (Fornell & Larcker, 1981).

Discriminant validity was evaluated using interfactor correlations, the Fornell–Larcker criterion (Fornell & Larcker, 1981), and the heterotrait–monotrait ratio of correlations (HTMT) (Henseler et al., 2015). Given the ordered-categorical nature of the indicators, HTMT was computed using polychoric correlations (Roemer et al., 2021). The assessment of discriminant validity was considered alongside comparisons of theoretically competing measurement models (Rönkkö & Cho, 2022). Discriminant validity was considered supported when the square root of each factor's AVE exceeded its correlations with the other factors (Fornell–Larcker criterion) and when HTMT values were below 0.85, with 0.90 as a more liberal upper bound (Henseler et al., 2015). HTMT was computed in its HTMT2 form, based on geometric means of the item correlations, with the semTools package.

2.4.3. Measurement Invariance Across Grade and Sex

Measurement invariance across grade and sex was evaluated to determine whether the latent constructs were measured equivalently across groups before latent mean comparisons were conducted (Putnick & Bornstein, 2016; Vandenberg & Lance, 2000). Given the ordered-categorical nature of the indicators, invariance testing followed the identification approach proposed by Wu and Estabrook (2016). A sequence of increasingly constrained models was estimated, including configural invariance, threshold invariance, metric invariance (equal thresholds and factor loadings), and scalar invariance (equal thresholds, factor loadings, and intercepts). Changes in approximate fit indices were evaluated using ΔCFI, ΔRMSEA, and ΔSRMR criteria (Chen, 2007). Because these criteria were developed for continuous indicators and maximum likelihood estimation, they were used as heuristics rather than strict cut-offs. Invariance steps were interpreted only when the configural model yielded admissible solutions in all groups. Where scalar invariance was supported, latent mean differences between groups were subsequently examined.

3. Results

3.1. Descriptive Statistics and Polychoric Correlations

Across the full sample (N = 514), all items used the complete 1–5 response range (observed minima = 1; maxima = 5) with negligible missingness. Item means varied widely (1.33–4.36) with moderate dispersion (SDs = 0.77–1.46); medians ranged from 1 to 5 and interquartile ranges from 0 to 3. Item distributions were non-normal and often asymmetric, as expected for 5-point Likert indicators of discrete emotional experiences; accordingly, items were analyzed as ordinal using polychoric correlations and WLSMV estimation. Full item-level descriptives are reported in Supplementary Table S1.

Polychoric correlations were consistent with the theorized structure of academic emotions. Items intended to measure the same emotion showed clearly positive associations (e.g., Enjoyment item pairs r = 0.49–0.80; Hope item pairs r = 0.49–0.71; Boredom item pairs r = 0.56–0.83), whereas correlations were generally negative between positive- and negative-valence indicators (e.g., Enjoyment with Boredom items $r = -0.30$ to -0.65).

Correlations among negative emotions (e.g., Anger–Anxiety item pairs up to $r = 0.78$) and positive emotions (Enjoyment with Hope item pairs up to $r = 0.77$) were positive, consistent with their shared unpleasant and pleasant affective tone. A heatmap summarizing the polychoric correlation matrix is provided in Supplementary Figure S1.

3.2. Measurement Model Testing

Competing CFA models were evaluated using WLSMV for ordered categorical indicators (Table 1). The two-factor valence model (Positive vs. Negative) showed inadequate absolute fit, indicating that a broad valence structure alone did not reproduce the observed item associations. The five-factor correlated emotions model (Enjoyment, Hope, Anger, Anxiety, Boredom) demonstrated acceptable fit and was retained as the primary measurement model at the full-sample level (see Figure 1). Given the high latent correlation between Enjoyment and Hope ($r = 0.95$), a four-factor alternative combining these two emotions into a single factor was also tested post hoc (see Figure 2); however, it fit slightly worse and was significantly inferior to the five-factor solution (robust $\Delta\chi^2(4) = 43.67$, $p < 0.001$). The a priori hierarchical Context–Emotion specification, in which emotion-by-context first-order factors loaded on second-order emotion factors, did not yield an admissible solution: the latent covariance matrix was not positive definite because of extreme collinearity between the Enjoyment and Hope second-order factors. Because fit indices from an improper solution cannot be meaningfully interpreted or compared with those of admissible models, this model is not included in Table 1 and was not considered further.

Table 1. Fit Indices for Competing Confirmatory Factor Analysis Models.

Model	χ^2 Scaled	df	CFI Scaled	TLI Scaled	RMSEA Scaled	SRMR Scaled
Valence two-factor (Positive vs. Negative)	4563.392	988	0.885	0.879	0.084	0.113
Correlated five-factor emotions (ENJ, HOP, ANG, ANX, BOR)	2544.659	979	0.950	0.947	0.056	0.075
Correlated four-factor emotions (POS, ANG, ANX, BOR)	2595.617	983	0.948	0.945	0.057	0.076

Note. Models were estimated with WLSMV using ordered-categorical indicators. WLSMV = weighted least squares mean and variance adjusted; χ^2 = chi-square; df = degrees of freedom; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; ENJ = Enjoyment; HOP = Hope; ANG = Anger; ANX = Anxiety; BOR = Boredom; POS = Positive emotions (ENJ + HOP). The a priori hierarchical Context–Emotion model is not reported because it did not yield an admissible solution (non-positive definite latent covariance matrix; see Section 3.2).

3.3. Reliability, Convergent Validity, and Discriminant Validity

Internal consistency was evaluated for each emotion factor using McDonald's omega (ω), alongside Cronbach's alpha and ordinal alpha (Table 2). All subscales demonstrated high reliability, with ω ranging from 0.894 to 0.937 and ordinal alpha values ranging from 0.927 to 0.954. Evidence for convergent validity was also supported by the average variance extracted (AVE), which exceeded the conventional 0.50 benchmark for all factors (AVE range = 0.538–0.726), indicating that each latent construct explained more than 50% of the variance in its indicators, on average.

Using the criteria described in Section 2.4.2, discriminant validity was supported for most construct pairs. The clearest exception was Enjoyment and Hope: their HTMT value (0.938) exceeded both the 0.85 and the 0.90 thresholds, and their latent correlation ($r = 0.95$) exceeded the square root of the AVE of either factor (0.827 for Enjoyment and 0.771 for Hope), so the Fornell–Larcker criterion was not met. Anger and Anxiety showed a smaller overlap: their latent correlation ($r = 0.85$) exceeded the square root of the AVE of both factors (0.844 for Anger and 0.734 for Anxiety), but their HTMT value (0.830) was below

the 0.85 threshold, supporting their discriminant validity under this criterion. All other pairs met both criteria.

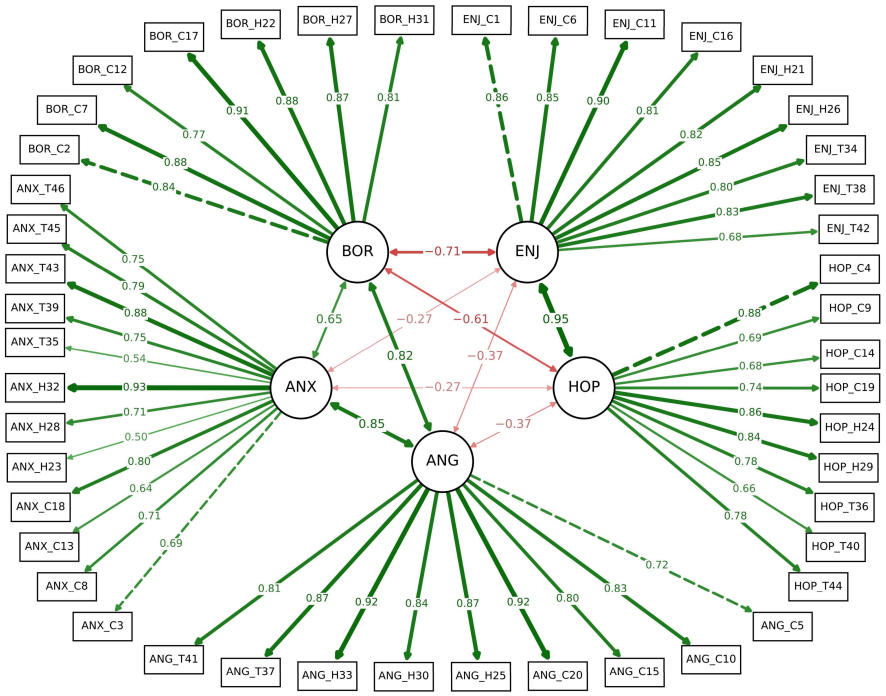


Figure 1. Standardized Five-Factor Measurement Model of the AEQ-ES-M. Green and red paths indicate positive and negative standardized estimates, respectively; line width and color intensity are proportional to the magnitude of the estimate. Dashed arrows indicate the marker indicator of each factor, whose unstandardized loading was fixed to 1 for model identification.

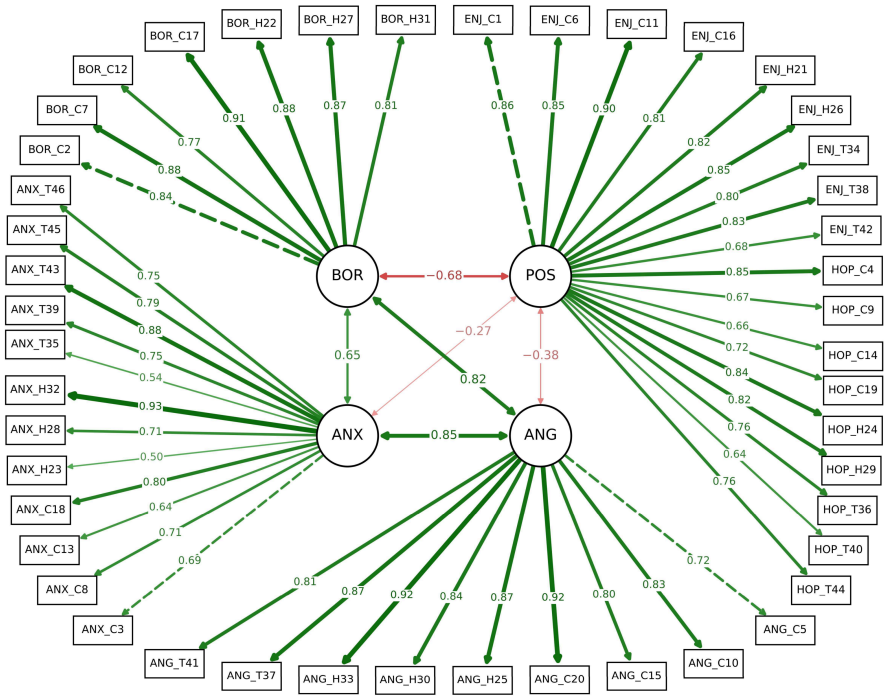


Figure 2. Standardized Four-Factor Model of the AEQ-ES-M, With Enjoyment and Hope Combined Into a Positive Emotion Factor. Green and red paths indicate positive and negative standardized estimates, respectively; line width and color intensity are proportional to the magnitude of the estimate. Dashed arrows indicate the marker indicator of each factor, whose unstandardized loading was fixed to 1 for model identification.

Table 2. Internal Consistency and Convergent Validity Indices for the Five-Factor Model.

Construct	α	α (Ordinal)	ω	AVE
ENJ (Enjoyment)	0.930	0.948	0.937	0.684
HOP (Hope)	0.904	0.929	0.910	0.594
ANG (Anger)	0.914	0.954	0.921	0.713
ANX (Anxiety)	0.886	0.927	0.894	0.538
BOR (Boredom)	0.925	0.945	0.925	0.726

Note. α = Cronbach's alpha (items treated as numeric); α (ordinal) = ordinal alpha based on polychoric correlations; ω = McDonald's omega; AVE = average variance extracted. Values are for the five-factor correlated-emotions model (WLSMV, ordered-categorical indicators). Discriminant validity was evaluated with the Fornell–Larcker criterion ($\sqrt{}$AVE greater than interfactor correlations) and HTMT2 computed from polychoric correlations (threshold 0.85; Henseler et al., 2015; Roemer et al., 2021); results are reported in Section 3.3.

3.4. Measurement Invariance Across Sex and Grade

Measurement invariance was examined for both the five- and four-factor models across sex and grade level using WLSMV estimation for ordered-categorical indicators and the Wu–Estabrook identification approach (Table 3). Sequential tests evaluated configural invariance, threshold invariance, metric invariance (equal thresholds and factor loadings), and scalar invariance (equal thresholds, factor loadings, and intercepts). Invariance was evaluated primarily from changes in approximate fit indices across increasingly constrained models (Chen, 2007).

Table 3. Fit Indices for Measurement Invariance Models Across Sex and Grade.

Model	χ^2	*df*	CFI	TLI	RMSEA	SRMR	ΔCFI	ΔRMSEA	ΔSRMR
					Sex				
Five-factor model									
Configural	3372.93	1958	0.955	0.952	0.053	0.084			
Threshold	3479.33	2050	0.954	0.954	0.052	0.084	−0.0005	−0.0009	0.0000
Metric	3495.20	2091	0.955	0.956	0.051	0.084	+0.0008	−0.0010	0.0000
Scalar	3500.29	2132	0.956	0.958	0.050	0.084	+0.0011	−0.0011	+0.0002
Four-factor model									
Configural	3419.67	1966	0.954	0.951	0.054	0.085			
Threshold	3526.70	2058	0.953	0.953	0.053	0.085	−0.0005	−0.0009	0.0000
Metric	3542.66	2100	0.954	0.955	0.052	0.085	+0.0008	−0.0010	0.0000
Scalar	3547.77	2142	0.955	0.957	0.051	0.085	+0.0012	−0.0012	+0.0002
					Grade				
Four-factor model									
Configural	3352.12	1966	0.953	0.951	0.052	0.083			
Threshold	3460.03	2058	0.953	0.952	0.052	0.083	−0.0005	−0.0009	0.0000
Metric	3496.83	2100	0.953	0.954	0.051	0.083	+0.0002	−0.0006	+0.0001
Scalar	3569.84	2142	0.952	0.953	0.051	0.084	−0.0010	+0.0001	+0.0004

Note. χ^2 = scaled chi-square; df = degrees of freedom; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual. ΔCFI, ΔRMSEA, and ΔSRMR represent changes relative to the immediately preceding invariance model. Grade invariance is reported only for the four-factor model because the five-factor configural solution yielded a non-positive-definite latent covariance matrix in Grade 2.

Across sex, the five-factor model showed evidence of invariance at all tested levels. The four-factor model showed a comparable pattern, with similarly small changes in CFI, RMSEA, and SRMR across successive levels of invariance. Thus, both specifications demonstrated scalar measurement invariance across boys and girls. Across grade levels, however, the five-factor model did not yield an admissible configural solution. In Grade 2, the five-factor solution produced a non-positive-definite latent covariance matrix, primarily as-

sociated with an extremely high latent correlation between Enjoyment and Hope (r = 0.983). The corresponding correlation remained high, although lower, in Grade 4 (r = 0.913), where the latent covariance matrix was positive definite. Consequently, invariance of the five-factor model across grades could not be established. In contrast, the four-factor model, in which Enjoyment and Hope were combined, yielded admissible solutions in both grades, and changes in approximate fit indices were minimal across successive levels of invariance, supporting scalar measurement invariance across grades for this specification. Thus, scalar invariance across grades was established only for the four-factor model, which therefore provided the only admissible basis for latent mean comparisons across grade levels.

3.5. Latent Mean Differences

Latent mean comparisons showed few sex differences but clear grade-related differences (Table 4). Across both factor specifications, boys and girls did not differ significantly in the positive-emotion factor(s), anger, or boredom; however, girls showed significantly higher anxiety than boys. Grade-related differences were more pronounced. Grade comparisons were based exclusively on the four-factor model, the only specification for which scalar invariance across grades was established. Grade 4 students showed significantly lower levels of the combined Enjoyment–Hope factor than Grade 2 students, but significantly higher levels of anger, anxiety, and boredom.

Table 4. Latent Mean Differences by Sex and Grade.

Model	Δ*M*	SE	*z*	*p*	95% CI
			Sex		
Five-factor model					
ENJ	0.070	0.097	0.73	0.468	[−0.120, 0.261]
HOP	0.019	0.100	0.19	0.850	[−0.176, 0.214]
ANG	0.068	0.110	0.62	0.537	[−0.147, 0.282]
ANX	−0.337	0.120	−2.81	0.005	[−0.572, −0.102]
BOR	−0.023	0.097	−0.24	0.811	[−0.214, 0.167]
Four-factor model					
POS	0.051	0.096	0.53	0.597	[−0.138, 0.240]
ANG	0.068	0.110	0.62	0.537	[−0.147, 0.282]
ANX	−0.337	0.120	−2.81	0.005	[−0.572, −0.102]
BOR	−0.023	0.097	−0.24	0.811	[−0.214, 0.167]
			Grade		
Four-factor model					
POS	−0.695	0.086	−8.10	<0.001	[−0.863, −0.527]
ANG	0.289	0.096	3.00	0.003	[0.100, 0.478]
ANX	0.332	0.088	3.76	<0.001	[0.159, 0.504]
BOR	0.365	0.082	4.44	<0.001	[0.204, 0.527]

Note. ENJ = Enjoyment; HOP = Hope; ANG = Anger; ANX = Anxiety; BOR = Boredom; POS = Positive emotions (ENJ + HOP). For sex, Δ*M* represents boys minus girls (0 − 1); therefore, negative values indicate higher latent means among girls. For grade, Δ*M* represents Grade 4 minus Grade 2; therefore, negative values indicate higher latent means in Grade 2 and positive values indicate higher latent means in Grade 4. Grade comparisons are reported only for the four-factor model because invariance across grades could not be established for the five-factor specification. *p* values are two-tailed (Wald *z* tests).

4. Discussion

The present study examined the differentiation and organization of five achievement emotions in mathematics among Chilean children in Grades 2 and 4, alongside differences in reported emotions across sex and grade. Three broad findings emerged. First, a five-factor correlated-emotions model fit the data well and outperformed a two-factor valence model,

supporting, at the full-sample level, the empirical differentiation of discrete emotions beyond simple positive–negative valence. Second, although the five factors were generally differentiated, enjoyment and hope showed substantial overlap, and this overlap was severe enough in Grade 2 to render the five-factor solution inadmissible in that grade. Third, scalar measurement invariance was supported across sex for both the five- and four-factor models, but across grades only for the four-factor model. Latent mean comparisons based on these models revealed higher anxiety among girls and, in Grade 4 relative to Grade 2, lower levels of the combined enjoyment–hope factor and higher levels of anger, anxiety, and boredom.

4.1. Construct Validity: A Broader Emotion Taxonomy in Early Elementary School

The superiority of the five-factor model over the two-factor valence model indicates that, consistent with the Control-Value Theory (Pekrun, 2006; Pekrun et al., 2023), discrete achievement emotions are empirically distinguishable from one another rather than reducible to a single positive–negative dimension. This conclusion applies to the full sample, however; in Grade 2, enjoyment and hope could not be separated empirically. This finding extends prior evidence from older samples (e.g., Bieleke et al., 2022) to a developmental period in which discrete-emotion differentiation had not previously been tested using an AEQ-based instrument. However, the hierarchical Context-Emotion model intended to reproduce the structure reported for the original three-emotion AEQ-ES (Lichtenfeld et al., 2012) could not be interpreted because it produced an improper solution, mainly because of the very strong association between enjoyment and hope. Extending the AEQ-ES to a broader range of emotions may therefore make that original hierarchical structure harder to maintain. Similar difficulties have been reported with broader AEQ-based versions in older students (Bieleke et al., 2022; Peixoto et al., 2015).

The overlap between enjoyment and hope was the main exception to an otherwise favorable pattern. Across the full sample, the two factors exceeded the recommended HTMT threshold (HTMT = 0.938). In Grade 2, their latent correlation approached unity (r = 0.983), resulting in an inadmissible five-factor solution for this subgroup. The correlation remained high, although somewhat lower, in Grade 4 (r = 0.913). Internal consistency and convergent validity were nevertheless strong for both subscales when considered separately (ω = 0.894–0.937; AVE = 0.538–0.726). The overlap is therefore unlikely to reflect unreliable measurement of either subscale. One interpretation is conceptual: within Control-Value Theory, enjoyment and hope are both positive, activating emotions grounded in similar control and value appraisals (Pekrun et al., 2023). This appraisal kinship may help explain why the two factors are hard to separate in early elementary school. The lower latent correlation in Grade 4 than in Grade 2 is compatible with greater differentiation among older children, but because the two grades comprised different cohorts, it cannot be interpreted as developmental change. A second, measurement-related possibility should also be considered: because six hope items were derived from hopelessness items, their content may partly overlap with general positive affect, which could strengthen their association with enjoyment (see Section 2.2). Adjacent AEQ emotions have shown the same problem in older samples and with verbal rather than pictorial items (Peixoto et al., 2015; Sanches et al., 2020). These explanations are not mutually exclusive: item content may set a common level of overlap, while younger children's less differentiated understanding of hope may further increase it. The enjoyment–hope distinction was therefore supported at the full-sample level but not in Grade 2.

Icon recognition may also have contributed to the overlap, although it is unlikely to be its main source. In an independent Grade 2 sample, anger and boredom sets were identified near ceiling, whereas hope (59.6%) and enjoyment (68.4%) were identified less

accurately; among children with exactly three correct matches, enjoyment–hope was the most frequent error pair (59.8%). Raccanello and Burro (2024) likewise found hope to be among the hardest achievement emotions for young children to recognize pictorially. That parallel concerns hope only. In their study, enjoyment was recognized near ceiling; in ours, it was not, so the recognition errors cannot be reduced to the secondary-emotion status of hope, nor can they be treated as the source of the HTMT result. Recognition was assessed only in Grade 2 and only as dichotomous accuracy. The icons may therefore have added some method variance to a pair that is already close on theoretical grounds and in item content.

Given this pattern, the four-factor model that combines enjoyment and hope, although showing slightly poorer overall fit, was the only specification that was admissible in both grades and invariant across them. We therefore recommend it for use with Grade 2 children and for comparisons across grade levels. The five-factor model may be used with older children, or at the full-sample level, when distinguishing enjoyment from hope is theoretically important; however, the high latent correlation between these factors even in Grade 4 ($r = 0.913$) indicates that separate enjoyment and hope scores should be interpreted with caution.

4.2. Measurement Invariance and Group Differences

The AEQ-ES-M demonstrated scalar invariance across sex for both the five- and four-factor specifications. Across grade, scalar invariance was supported only for the four-factor specification, because the five-factor model was inadmissible in Grade 2. Latent means could therefore be compared across sex for both specifications, but across grade only for the combined enjoyment–hope factor, anger, anxiety, and boredom. This pattern is consistent with prior AEQ-based evidence, although previous studies have examined somewhat different factor structures from those tested in the present study. Raccanello et al. (2019) found that a joint hierarchical AEQ-ES model comprising three emotions (enjoyment, boredom, and anxiety) showed scalar invariance across country, grade, sex, and academic domain (native language vs. mathematics) in Italian second- and fourth-graders, the same developmental range examined here. Raccanello et al. (2022), also focusing on this age range in the mathematics domain, examined invariance separately for each of the three emotions rather than jointly, and likewise reported scalar invariance across grade and sex for all three emotions. Moreira et al. (2019), using a correlated-uniqueness Emotion × Setting approach with the AEQ-M, similarly established full scalar invariance across school year and sex for a joint structure encompassing seven mathematics-related emotions in Portuguese adolescents. Taken together, these findings suggest that the measurement of achievement emotions tends to remain comparable across sex and educational level, despite differences in the specific AEQ version and model structure used. The present study adds to this evidence by showing that scalar invariance across sex and grade can also be achieved in young elementary-school students for a joint multidimensional structure capturing several positive and negative mathematics-related emotions, provided that enjoyment and hope are combined into a single factor.

Girls reported significantly higher anxiety than boys, with no significant sex differences for the remaining emotions. This finding is consistent with the broader literature on mathematics-related emotions. Higher anxiety among girls has been linked to less favorable control and value appraisals of mathematics, even when achievement is comparable (Frenzel et al., 2007). Consistent with this pattern in the AEQ-ES literature, Lichtenfeld et al. (2012) found a comparable sex gap favoring boys in enjoyment and anxiety among German second- and third-graders, whereas in their American sample, the only significant sex difference was in boredom, and in the opposite direction (boys reporting more boredom

than girls), with no differences in enjoyment or anxiety. The authors attributed the pattern to the absence of sex differences in perceived control and value in that sample. Raccanello et al. (2019) similarly found no reliable sex main effect on boredom, enjoyment, or anxiety among Italian second- and fourth-graders in mathematics. Together, these findings suggest that higher mathematics anxiety among girls is a commonly reported pattern, but it is not universal. Sex differences in other achievement emotions also vary across cultural contexts and measurement approaches. This variability underscores the importance of establishing measurement invariance before comparing group means.

The pattern of grade-related differences was more pronounced. Grade 4 students reported lower levels of the combined enjoyment–hope factor and higher levels of anger, anxiety, and boredom than Grade 2 students. The less favorable emotional profile observed in Grade 4 is broadly consistent with Control-Value Theory. Formal assessment practices and social comparison tend to become more salient across the elementary school years, which may influence perceived control and the value attached to academic outcomes (Pekrun et al., 2011, 2023). This interpretation is consistent with previous findings in similar age groups. For example, Raccanello et al. (2019) found a similar overall pattern in Italian elementary school children: second-graders reported higher enjoyment and lower boredom and anxiety than fourth-graders. Similarly, in a longitudinal study of Norwegian children followed from Grade 1 to Grade 3, Rawlings et al. (2023) observed a decrease in mathematics enjoyment and an increase in mathematics boredom, consistent with the lower positive emotions and higher boredom observed in Grade 4 in the present study, although mean anxiety levels remained low. Longitudinal findings by Lichtenfeld et al. (2023) for Grades 2 to 4 showed a somewhat different pattern. Enjoyment declined, particularly between Grades 3 and 4, but boredom remained stable, and anxiety unexpectedly decreased between Grades 2 and 3 and then remained stable. This pattern contrasts with the higher boredom and anxiety observed in Grade 4 in the present cross-sectional comparison. The discrepancies with Lichtenfeld et al. (2023), particularly concerning anxiety, may reflect cultural or educational context. They may also be related to study design: because Grades 2 and 4 comprised different cohorts, the present findings describe cross-sectional, grade-related group differences, and developmental change and cohort effects cannot be disentangled (see Section 4.4).

4.3. Theoretical and Practical Implications

These findings extend the AEQ-ES framework in the two directions identified in the introduction. First, the present study shows that a broader set of achievement emotions can be assessed in children as young as Grade 2 and that a five-factor structure is supported at the full-sample level. However, not all emotions were equally well differentiated: enjoyment and hope overlapped substantially and could not be separated empirically in Grade 2. Second, this study provides some of the first evidence for the structural validity, reliability, and measurement invariance of an AEQ-based instrument in a Latin American elementary-school sample. It therefore extends the cross-cultural evidence base beyond the German, American, Italian, and Portuguese samples in which the AEQ-ES and related adaptations have primarily been examined (Lichtenfeld et al., 2012; Peixoto et al., 2015; Raccanello et al., 2019, 2022; Sanches et al., 2020). From a practical standpoint, the AEQ-ES-M offers researchers and educators in Chile, and potentially other Latin American contexts, a reliable tool for assessing a broad range of mathematics-related emotions from the early years of primary school, with the four-factor scoring recommended for Grade 2 and for comparisons across grades. The less favorable emotional profile observed in Grade 4 than in Grade 2, with lower enjoyment–hope and higher anger, anxiety, and boredom, also highlights the value of assessing mathematics-related emotions across different stages

of primary education. Such assessment may help identify students or groups with less favorable emotional profiles and inform targeted support.

4.4. Limitations and Future Research

Several limitations should be considered. First, the cross-sectional design compared two non-adjacent grade cohorts (Grade 2 and Grade 4) rather than following the same children over time, so the observed differences reflect grade-related group comparisons rather than confirmed developmental trajectories and cannot rule out cohort effects. Second, although the AEQ-ES-M items were written to distinguish class, homework, and test contexts, the retained five-factor model treats each emotion as a single, context-collapsed factor. This is not necessarily at odds with more recent theorizing. Pekrun et al. (2023) conceptualize achievement settings as situational antecedents of emotions. In their taxonomy, context is not a structural dimension of the emotion construct itself. Nonetheless, the present study cannot speak to whether the five emotions operate differently across class, homework, and test settings. Testing this dimension may require a larger sample, a reduced emotion set, or an alternative modeling strategy (e.g., exploratory structural equation modeling). Third, hope was the most extensively constructed component of the instrument, and several of its items were derived from hopelessness items. Because hope and hopelessness are not necessarily psychometric opposites, these items may partly overlap with general positive affect, which may have contributed to the overlap with enjoyment; cognitive interviews with children and item-level analyses would help refine this scale. Fourth, students were nested within nine schools, and this clustering was not modeled. With so few clusters, multilevel or cluster-robust estimation would have produced biased or unstable estimates (McNeish & Stapleton, 2016). However, ignoring clustering may underestimate standard errors, so the significance tests for latent mean differences should be interpreted with some caution. Finally, the lower recognition of the hope and enjoyment icons in Grade 2 may have contributed to the limited discriminant validity between these emotions, alongside item content and their conceptual closeness, although this contribution could not be estimated with the present data. Future research should examine these questions longitudinally, ideally spanning a wider and more continuous range of grades, to determine whether the differentiation between enjoyment and hope strengthens across development, as the lower latent correlation in Grade 4 might suggest. Further validation of the newly developed facial icon set, including recognition tasks with older children and intensity-rating tasks, would help clarify the extent to which the icons contribute to the enjoyment-hope overlap.

5. Conclusions

Overall, the present study extends the AEQ-ES to five mathematics-related achievement emotions and provides evidence of reliability and of a five-factor structure at the full-sample level in Chilean elementary school children, thus bringing the AEQ framework to a Latin American elementary-school context. Enjoyment and hope, however, overlapped substantially, and the five-factor solution was inadmissible in Grade 2, where the four-factor model was appropriate. Scalar invariance was established across sex for both the five- and four-factor models, but across grade only for the four-factor model combining enjoyment and hope, which is therefore recommended for Grade 2 and for comparisons across grades. In addition, the hierarchical emotion-by-context structure of the original AEQ-ES could not be retained with five emotions.

Girls reported higher mathematics anxiety than boys, while no sex differences emerged for the remaining emotions. In this cross-sectional comparison, Grade 4 students showed a less favorable emotional profile than Grade 2 students, with lower levels of the combined enjoyment–hope factor and higher anger, anxiety, and boredom. These findings highlight

the value of examining both sex and grade differences in mathematics-related emotions from the early years of primary education. In sum, the AEQ-ES-M offers a reliable tool for assessing a broad range of mathematics-related emotions in Chilean primary-school students, with the four-factor specification recommended for younger children.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/educsci16101760/s1: Table S1: Item-Level Descriptive Statistics for the AEQ-ES-M; Figure S1: Heatmap of Polychoric Correlations Among AEQ-ES-M Items; Table S2: Final AEQ-ES-M Items in Spanish With English Translations; Figure S2: Facial Icon Sets (Boy and Girl Versions) Used as Response Anchors for the Five Emotions.

Author Contributions: Conceptualization, C.R.; Methodology, C.R.; Software, I.M.; Formal Analysis, C.R. and M.H.C.-M.; Investigation, C.R., B.G., F.S., R.A.F., C.A.R.-B. and A.C.; Resources, C.R. and M.H.C.-M.; Data Curation, C.R., M.H.C.-M. and I.M.; Writing—Original Draft Preparation, C.R.; Writing—Review and Editing, C.R., B.G., F.S., R.A.F., C.A.R.-B., M.H.C.-M., A.C. and I.M.; Project Administration, C.R.; Funding Acquisition, C.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Chilean National Research and Development Agency (ANID), FONDECYT Regular, grant number 1240552.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Scientific Ethics Committee of the Universidad de Talca (Folio 19/2024, 29 May 2024). This approval was subsequently endorsed by the Scientific Ethics Committee of the Universidad Católica de la Santísima Concepción (CEC 01, 29 January 2026).

Informed Consent Statement: Informed consent was obtained from the parents/legal guardians of all children who participated in the study, and assent was obtained from the children themselves prior to participation, in accordance with the protocol approved by the Ethics Committee.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Acknowledgments: The authors thank the schools, teachers, and children who participated in this study.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

Barroso, C., Ganley, C. M., McGraw, A. L., Geer, E. A., Hart, S. A., & Daucourt, M. C. (2021). A meta-analysis of the relation between math anxiety and math achievement. *Psychological Bulletin*, *147*(2), 134–168. [CrossRef] [PubMed]

Bieleke, M., Goetz, T., Yanagida, T., Botes, E., Frenzel, A. C., & Pekrun, R. (2022). Measuring emotions in mathematics: The Achievement Emotions Questionnaire—Mathematics (AEQ-M). *ZDM—Mathematics Education*, *55*(2), 269–284. [CrossRef] [PubMed]

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. [CrossRef] [PubMed]

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. [CrossRef] [PubMed]

Camacho-Morles, J., Slemp, G. R., Pekrun, R., Loderer, K., Hou, H., & Oades, L. G. (2021). Activity achievement emotions and academic performance: A meta-analysis. *Educational Psychology Review*, *33*(3), 1051–1095. [CrossRef]

Cargnelutti, E., Tomasetto, C., & Passolunghi, M. C. (2016). How is anxiety related to math performance in young students? A longitudinal study of grade 2 to grade 3 children. *Cognition & Emotion*, *31*(4), 755–764. [CrossRef] [PubMed]

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. [CrossRef]

del Río, M. F., Susperreguy, M.-I., & Morales, M.-F. (2023). Kindergarten children's math anxiety and its relationship with mathematical performance. *Studies in Psychology/Estudios de Psicología*, *44*(2–3), 542–561. [CrossRef]

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50. [CrossRef]

Forsblom, L., Pekrun, R., Loderer, K., & Peixoto, F. (2022). Cognitive appraisals, achievement emotions, and students' math achievement: A longitudinal analysis. *Journal of Educational Psychology*, *114*(2), 346–367. [CrossRef]

Frenzel, A. C., Pekrun, R., & Goetz, T. (2007). Girls and mathematics—A "hopeless" issue? A control-value approach to gender differences in emotions towards mathematics. *European Journal of Psychology of Education*, *22*(4), 497–514. [CrossRef]

Guzmán, B., Rodríguez, C., & Ferreira, R. A. (2021). Longitudinal performance in basic numerical skills mediates the relationship between socio-economic status and mathematics anxiety: Evidence from Chile. *Frontiers in Psychology*, *11*, 611395. [CrossRef] [PubMed]

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, *43*(1), 115–135. [CrossRef]

International Test Commission. (2017). *The ITC guidelines for translating and adapting tests* (2nd ed.). International Test Commission. Available online: https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf (accessed on 14 October 2026).

Li, C. H. (2015). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949. [CrossRef] [PubMed]

Lichtenfeld, S., Pekrun, R., Marsh, H. W., Nett, U. E., & Reiss, K. (2023). Achievement emotions and elementary school children's academic performance: Longitudinal models of developmental ordering. *Journal of Educational Psychology*, *115*(4), 552–570. [CrossRef]

Lichtenfeld, S., Pekrun, R., Stupnisky, R. H., Reiss, K., & Murayama, K. (2012). Measuring students' emotions in the early years: The Achievement Emotions Questionnaire-Elementary School (AEQ-ES). *Learning and Individual Differences*, *22*(2), 190–201. [CrossRef]

Lundqvist, D., Flykt, A., & Öhman, A. (1998). *The Karolinska directed emotional faces—KDEF* [CD-ROM]. Department of Clinical Neuroscience, Psychology Section, Karolinska Institutet.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. [CrossRef]

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, *28*(2), 295–314. [CrossRef]

Moreira, P., Cunha, D., & Inman, R. A. (2019). Achievement Emotions Questionnaire-Mathematics (AEQ-M) in adolescents: Factorial structure, measurement invariance and convergent validity with personality. *European Journal of Developmental Psychology*, *16*(6), 750–762. [CrossRef]

OECD. (2023). *PISA 2022 results (Volume I): The state of learning and equity in education*. OECD Publishing. [CrossRef]

Peixoto, F., Mata, L., Monteiro, V., Sanches, C., & Pekrun, R. (2015). The achievement emotions questionnaire: Validation for pre-adolescent students. *European Journal of Developmental Psychology*, *12*(4), 472–481. [CrossRef]

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, *18*(4), 315–341. [CrossRef]

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology*, *36*(1), 36–48. [CrossRef]

Pekrun, R., Marsh, H. W., Elliot, A. J., Stockinger, K., Perry, R. P., Vogl, E., Goetz, T., van Tilburg, W. A. P., Lüdtke, O., & Vispoel, W. P. (2023). A three-dimensional taxonomy of achievement emotions. *Journal of Personality and Social Psychology*, *124*(1), 145–178. [CrossRef] [PubMed]

Pekrun, R., & Stephens, E. J. (2010). Achievement emotions: A control-value approach. *Social and Personality Psychology Compass*, *4*(4), 238–255. [CrossRef]

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. [CrossRef] [PubMed]

Raccanello, D., Brondino, M., & De Bernardi, B. (2013). Achievement emotions in elementary, middle, and high school: How do students feel about specific contexts in terms of settings and subject-domains? *Scandinavian Journal of Psychology*, *54*(6), 477–484. [CrossRef] [PubMed]

Raccanello, D., Brondino, M., & Moè, A. (2022). Malleability beliefs shape mathematics-related achievement emotions: The mediating role of emotion regulation in primary school children. *Learning and Individual Differences*, *97*, 102177. [CrossRef]

Raccanello, D., Brondino, M., Moè, A., Stupnisky, R., & Lichtenfeld, S. (2019). Enjoyment, boredom, anxiety in elementary schools in two domains: Relations with achievement. *The Journal of Experimental Education*, *87*(3), 449–469. [CrossRef]

Raccanello, D., & Burro, R. (2024). Development of a Drawing Set for the Achievement Emotions Adjective List (DS-AEAL): Preliminary data on a pictorial instrument for children. *Education Sciences*, *14*(7), 756. [CrossRef]

Rawlings, A. M., Niemivirta, M., & Korhonen, J. (2023). Achievement emotions and arithmetic fluency—Development and parallel processes during the early school years. *Learning and Instruction*, *86*, 101776. [CrossRef]

R Core Team. (2021). *R: A language and environment for statistical computing* (Version 4.1). R Foundation for Statistical Computing. Available online: https://www.R-project.org/ (accessed on 14 October 2026).

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Association for Computational Linguistics. [CrossRef]

Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4512–4525). Association for Computational Linguistics. [CrossRef]

Roemer, E., Schuberth, F., & Henseler, J. (2021). HTMT2–an improved criterion for assessing discriminant validity in structural equation modeling. *Industrial Management & Data Systems*, *121*(12), 2637–2650. [CrossRef]

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. [CrossRef]

Rönkkö, M., & Cho, E. (2022). An updated guideline for assessing discriminant validity. *Organizational Research Methods*, *25*(1), 6–14. [CrossRef]

Sanches, C., Monteiro, V., Mata, L., Santos, N., & Gomes, M. (2020). Psychometric properties of the Portuguese version of the achievement emotions questionnaire for elementary school. *Analise Psicologica*, *38*(1), 127–139. [CrossRef]

Schoenherr, J., Schukajlow, S., & Pekrun, R. (2025). Emotions in mathematics learning: A systematic review and meta-analysis. *ZDM–Mathematics Education*, *57*(4), 603–620. [CrossRef]

Schukajlow, S., Rakoczy, K., & Pekrun, R. (2023). Emotions and motivation in mathematics education: Where we are today and where we need to go. *ZDM—Mathematics Education*, *55*(2), 249–267. [CrossRef] [PubMed]

Sepúlveda, F., Rodríguez, C., & Peake, C. (2020). Differences and associations in symbolic and non-symbolic early numeracy competencies of Chilean kinder grade children, considering socioeconomic status of schools. *Early Education and Development*, *31*(1), 137–151. [CrossRef]

Vandenberg, R. J., & Lance, C. E. (2000). A Review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–69. [CrossRef]

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika*, *81*(4), 1014–1045. [CrossRef] [PubMed]

Xie, J., Cho, K. W., Wei, T., Xu, J., & Fan, M. (2025). The effects of academic emotions on learning outcomes: A three-level meta-analysis of research conducted between 2000 and 2024. *Learning and Motivation*, *90*, 102109. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.