Test of Measurement Invariance, and Evidence for Reliability and Validity of AMAS Scores in Dutch Secondary School and University Students

Schmitz, E.A.; Salemink, E.; Wiers, R.W.; Jansen, B.R.J.

DOI
10.1177/07342829221086141

Publication date
2022

Document Version
Final published version

Published in
Journal of Psychoeducational Assessment

Article

Journal of Psychoeducational Assessment
2022, Vol. 40(5) 663–677

DOI: 10.1177/07342829221086141
journals.sagepub.com/home/jpa


Test of Measurement Invariance, and Evidence for Reliability and Validity of AMAS Scores in Dutch Secondary School and University Students

Eva A. Schmitz[1], Elske Salemink[1,2], Reinout W. Wiers[1], and Brenda R. J. Jansen[1,3,4]

Abstract

The Abbreviated Math Anxiety Scale (AMAS) is commonly used to compare groups on math anxiety. Group comparisons should however be preceded by a demonstration of metric and scalar measurement invariance, which is currently only available for undergraduate students in the USA. This study tested for metric and scalar measurement invariance of AMAS scores across sexes and age groups and investigated reliability and validity evidence. Dutch secondary school students (N = 1504) and university students (N = 629) completed the AMAS and reported their math grades. Subsamples also completed questionnaires on related (e.g., trait anxiety) and unrelated constructs (e.g., anxiety for learning a foreign language). Results of multi-group Confirmatory Factor Analyses showed the 2-factor structure that was detected in earlier studies and supported partial metric and scalar invariance, although cross-loadings for one item were included for the female university students group to improve model fit. Reliability ranged from acceptable to good and validity was supported.

Keywords

measurement invariance, math anxiety, sex differences, age groups

[1]Department of Developmental Psychology, University of Amsterdam, the Netherlands
[2]Clinical Psychology, Utrecht University, The Netherlands
[3]Cognitive Science Centre Amsterdam, The Netherlands
[4]Yield, Research Institute of Child Development and Education, The Netherlands

Corresponding Author:
Brenda R. J. Jansen, University of Amsterdam, P.O. box 15916, 1001 NK, Amsterdam 1018 WB, The Netherlands.
Email: b.r.j.jansen@uva.nl

Math anxiety is a negative affective reaction to math-related situations that may disrupt math performance (Dowker et al., 2016; Suárez-Pellicioni et al., 2016). Of all math anxiety assessment instruments, the Abbreviated Math Anxiety Scale (AMAS; Hopko et al., 2003) is the most widely used internationally. The AMAS was first used in samples of undergraduates (e.g., Hopko et al.), followed by adaptations in school samples (e.g., Caviola et al., 2017; Hill et al., 2016), and translations to Persian (Vahedi & Farrokhi, 2011), Italian (Primi et al., 2014), Polish (Cipora et al., 2015), Spanish (Brown & Sifuentes, 2016), and German (Schillinger et al., 2018).

Often, females' AMAS scores are higher than males' (e.g., Hopko et al., 2003; but see Vahedi & Farrokhi, 2011) and scores seem highest among secondary school students (Hill et al., 2016; Primi et al., 2014). Although such group comparisons are informative for school practices and tailoring interventions, metric and scalar measurement invariance are required for these comparisons (explained below; Brown, 2015). The only study to date on metric and scalar invariance of AMAS scores involves a sample of undergraduates in the USA (Cho, 2022). Other invariance tests (Primi et al., 2014; Vahedi & Farrokhi, 2011) lack tests of metric and scalar invariance. The current study aims to investigate metric and scalar measurement invariance as well as evidence for reliability and validity of AMAS scores using its Dutch translation (Schmitz et al., 2019)[1] in a large sample of secondary school students and university students.

Math anxiety manifests itself in cognition, behavior, affect, and physiology (Hopko et al., 2003). From the very first math anxiety instrument (Math Anxiety Rating Scale—MARS; Richardson & Suinn, 1972), factors of math anxiety instruments were based on the situations in which the anxiety occurs. Factor analysis of AMAS scores resulted in two moderately correlated factors: Learning Math Anxiety (LMA: anxiety about the process of learning math), and Math Evaluation Anxiety (MEA: anxiety in math testing situations; Hopko et al.). The factor structure was replicated in both sexes (Cho, 2022) and several age groups (e.g., Primi et al., 2014: high school students and university students), but for some university samples loading item 5 (on "homework") on both factors resulted in superior model fit (Cipora et al., 2015; Schillinger et al., 2018).

In the case of metric measurement invariance, factor loadings are equal across groups. Scalar invariance additionally requires equality of intercepts across groups. Equal factor loadings and intercepts imply that the underlying factor in the groups has the same unit of measurement (factor loading) and same origin (intercept), and thus, that the measurement scales have the same operational definition across groups (Brown, 2015; Chen, 2007; Cheung & Rensvold, 2002; van de Schoot et al., 2012). If intercepts are unequal across groups, mean differences may be invalid and cannot be interpreted in terms of the underlying latent variables (Putnick & Bornstein, 2016; Wicherts & Dolan, 2010). Therefore, the first aim of the current study is to test for metric and scalar measurement invariance of the AMAS and if tenable, also residuals invariance, across sexes and age-groups (i.e., secondary school and university students).

The second aim of this study is to assess reliability of the scores, to provide norms for the AMAS scales, and to compare groups on math anxiety. Research with the original AMAS and translations supports reliability/precision of AMAS scores, as internal consistency ranged from $\alpha = .82$ (Vahedi & Farrokhi, 2011) to $\alpha = .90$ (Hopko et al., 2003), and test–retest correlations were as high as $r = .85$ after 2 weeks (Hopko et al.), declining to $r = .82$ after 6 weeks (Schillinger et al., 2018), and $r = .71$ after 4 months (Cipora et al., 2015). Further, it will be tested whether current females' AMAS scores are higher than males' (e.g., Cipora et al., 2015), and lower in university than secondary school students (Primi et al., 2014).

The third aim is to collect validity evidence of AMAS scores. Next to the evidence based on the AMAS′ internal structure that may follow from the factor analyses required for testing measurement invariance, relations of AMAS scores with other variables may provide convergent validity evidence (AERA et al., 2014). Supporting the idea that math anxiety is a related though

distinct type of anxiety, trait anxiety was positively related to math anxiety (Dowker et al., 2016; Suárez-Pellicioni et al., 2016). Also for AMAS scores, the relation with trait anxiety scores is weakly to moderately positive (e.g., Hopko et al., 2003). Math anxiety may worsen performance through avoidance, worries, physiological arousal, and anxiety feelings. Conversely, weak math performance may lead to math anxiety (Carey et al., 2016; Suárez-Pellicioni et al., 2016). The relation tends to be weaker for standardized achievement tests than for researcher-made instruments or teacher-reported grades (Ma, 1999). A weak to moderate negative relation between math anxiety scores and math performance is expected, similar to previous studies with the AMAS (e.g., Cipora et al., 2015) or other instruments (Dowker et al.). Finally, a negative relation between math anxiety and math self-concept and math self-efficacy has been observed (e.g., Lee, 2009). In sum, positive relations between AMAS scores and scores on trait anxiety as well as negative relations between AMAS scores and math performance, math self-concept, and math self-efficacy would support convergent validity evidence. Additionally, discriminant validity is supported if AMAS scores have very weak relations with anxiety and performance in another school subject: learning a foreign language.

Method

Participants

Secondary school sample. A total of 1549 Dutch secondary school students participated in three studies. The original study purposes were the introduction of a math anxiety questionnaire (Study 1: $n = 998$, Schmitz, 2020), comparing implicit and explicit math anxiety measures (Study 2: $n = 189$, Schmitz et al., 2019), and relating math anxiety to other motivational and affective factors (Study 3: $n = 362$, Sachisthal et al., 2021).[2] Combining the data of the studies allowed for a powerful test of measurement invariance.[3]

In all studies, participants were recruited by contacting schools in the vicinity and slightly wider surroundings of the university. After the school's acceptance, students and parents were approached. Participants received no reward. Parents received an information letter and could exempt their child from participating. All studies were approved by the local ethical review board. Data were excluded for participants for whom sex was not reported ($n = 8$), AMAS scores were missing ($n = 30$), or who had a high error percentage and missing data on other tasks in the study, suggesting that participation was not taken seriously ($n = 7$).

The final sample consisted of 1504 students ($M_{age} = 13.4$ years, $SD = 1.02$; 699 males), attending junior classes (41.8% first, 32.0% second, and 26.3% third year) of all Dutch educational levels (38.4% lower; 21.3% middle; 40.2% higher), except special education. All secondary school students followed Dutch education and were capable of understanding the Dutch questionnaires. Females' average grades for math and English were significantly but not meaningfully higher than males' (math: $M_{males} = 6.6$, $SD_{males} = 1.27$; $M_{females} = 6.8$, $SD_{females} = 1.28$; $t\,(1495) = -2.73$, $p = .006$, $d = 0.14$; English: $M_{males} = 7.0$, $SD_{males} = 1.23$; $M_{females} = 7.2$, $SD_{females} = 1.21$; $t\,(1495) = -2.73$, $p = .006$, $d = 0.16$). Sex differences were not significant for any other measure.

University Sample

A total of 637 first-year Psychology students participated in three semi-mandatory annual assessments for course credits (2015: $n = 105$; 2017: $n = 362$; 2019: $n = 170$). Recruitment took place through a central student registration website. In 2015 and 2017, Dutch was the program's language of instruction and participants were thus capable of understanding the Dutch

questionnaires. In 2019, the program became bilingual and participants chose between completing the Dutch or the English questionnaire. Only responses to the Dutch questionnaire were included here. Informed consent was obtained, and procedures were approved by the local ethical review board. Data for participants were excluded because sex was unreported (n = 4) or reported as "other" (n = 3), or because of reported lack of math experience (n = 1).

The final sample consisted of 629 students (M_{age} = 20.0 years, SD = 2.73; 176 males). Women were slightly younger (M_{age} = 19.7 years, SD = 2.19) than men (M_{age} = 20.8 years, SD = 3.68; t (221.95) = 3.74, $p < .001$, d = 0.37). A total of 84.7% were Dutch native, 12.6% bilingual, and 2.7% non-Dutch native speakers. Most participants had finished Dutch pre-university compulsory math education, whereas a minority finished a different math program (13.2%). The average math grade did not significantly differ between sexes, t (598) = −1.21, p = .229, d = 0.11.

Materials

Math Anxiety

The Dutch version of the AMAS (Schmitz et al., 2019) was used to assess math anxiety (see for details on its development: Schmitz et al., 2019). Participants rated how anxious they would feel in a given math situation (9 items) on a 5-point rating scale (1: *almost not anxious*; 5: *very anxious*). The AMAS contained the subscales *Learning Math Anxiety* (LMA) and *Math Evaluation Anxiety* (MEA).

Trait Anxiety

The trait scale of the Dutch State-Trait-Anxiety Inventory for Children (STAI-C; Spielberger et al., 1973; Dutch translation: Bakker et al., 1989) was used to measure trait anxiety (20 items). Participants indicated for each anxiety-related statement how often it applied to them in general using a 3-point rating scale (1: *hardly ever*; 3: *often*). Higher scores indicated higher trait anxiety. Bakker et al. (1989) reported content validity evidence. Cronbach's alpha in the current sample was .89 (95%-confidence interval: .85–.91).

Math Performance

As an indicator of math performance, secondary school students reported the average math grade on their most recent school report (Studies 1 and 2) or current average math grade (Study 3). University students reported on the final math grade and the type of math at secondary school or pre-university level. Grades represent the degree to which a participant masters the curriculum offered. Although self-reported grades may be higher than school-reported grades, their strong correlation and the clear instruction of which grade was requested allow using self-reported grades to investigate the relation between AMAS scores and math performance (Kuncel et al., 2005; Sticca et al., 2017; see Cipora et al., 2015 for a similar approach). Data for math grades were deleted if participants reported international education (n = 6), or a math type that is not preparatory for university (n = 7). Grades of the Dutch educational system range from 1.0 to 10.0 with higher grades representing better performance (5.5 is minimum to pass).

Math Self-Concept and Math Self-Efficacy

The Dutch versions of the PISA 2012 Math self-concept scale (5 items) and Math self-efficacy scale (8 items) were used to assess math self-concept and math self-efficacy (Cito, 2012; OECD, 2013). Participants indicated how much they agreed with statements on their self-confidence in

learning math on a 4-point rating scale (1: *strongly disagree*; 4: *strongly agree*). Higher mean scores indicated higher math self-concept and math self-efficacy. Cronbach's alphas in the current sample were .87 (95%-confidence interval: .85–.89) and .79 (95%-confidence interval: .75–.82), respectively.

Learning a Foreign Language: Anxiety and Performance

The adapted Dutch version of the Foreign Language Classroom Anxiety Scale (FLCAS; Horwitz et al., 1986; Dutch translation: Schmitz et al., 2019) was used to assess anxiety for learning English, a foreign language for Dutch students (33 items). Participants indicated how much statements on anxiety when learning English applied to them on a 5-point rating scale (1: *strongly disagree*; 5: *strongly agree*). Higher mean scores indicated higher anxiety. Validity evidence for FLCAS scores was reported by Horwitz et al. Cronbach's alpha was .94 (95%-confidence interval: .92–.96) in the current sample. Also, participants reported their English grades as an indicator of foreign language performance. These measures were included to test whether AMAS scores were only related to math and not to anxiety or performance in a different school subject (discriminant validity).

Procedure

Secondary school sample

In all studies, the online assessments took place in computer rooms at schools. After classroom instruction, participants completed the assessment individually. Table 1 shows the measures that were included in the current analyses, by study. Studies 1 and 2 started and Study 3 ended with questions on demographics and grades. The order of the AMAS and remaining tasks was random. Assessment ended with a short debriefing and lasted approximately 40 minutes.

University Sample

All assessments were online and took place in university computer rooms. Assessment started with demographic questions, followed by administration of the AMAS. The questionnaires were introduced as focusing on secondary school math. For most students, secondary school experiences were only a few months ago as assessment took place in the first semester of the university year. Participants were asked how they would describe themselves in general. Finally, participants reported their math grade.

Table 1. Variables per Study Next to AMAS.

Sample	Final *n*	Variables
Secondary school		
Study 1	996	Math grade, English grade
Study 2	154	Math grade, English grade, State-Trait-Anxiety Inventory for Children, Foreign Language Classroom Anxiety Scale
Study 3	354	Math grade, math self-concept, self-efficacy
University		
	629	Math grade

Data Analyses

To test for measurement invariance across the four subgroups (i.e., male and female students of secondary school and university), a multi-group Confirmatory Factor Analysis (CFA; following Brown, 2015) was performed using Mplus Version 7.31 (Muthén & Muthén, 2015). A sequence of models with increasing equality constraints across groups was tested, and more stringent invariance would be supported by non-significant difference tests of model fit between the more and less restricted models. First, as a prerequisite, it was tested whether the two-factor model (Model_0) showed acceptable fit and significant factor loadings in all groups, separately. Second, an identical two-factor structure across groups (i.e., *configural invariance*) was tested using an equal form model (i.e., same number of factors and same pattern of indicator-factor loadings) with all groups simultaneously (Model_1). Third, *metric invariance* was tested by constraining the factor loadings to be equal across groups (Model_2). Fourth, *scalar invariance* was tested by additionally constraining the indicator intercepts to be equal across groups (Model_3). Lastly, *strict factorial invariance* was tested by constraining the indicator residual variances to be equal across groups (Model_4). Any subsequent restriction was only applied if the previous restriction was allowed.

The robust Mean-adjusted Maximum Likelihood (MLM) procedure was used, which provides the Satorra–Bentler scaled χ^2 statistic (SB χ^2; Satorra & Bentler, 2010) because the data did not follow a normal distribution (Brown, 2015; see OSM Table 1.2 for details on skewness and kurtosis). A significant SB χ^2 test ($p < .05$) indicates poor model fit, but is very sensitive to deviations from the model in large samples. Therefore, evaluation of overall model fit was based on CFI (values $\geq .95$ indicate good model fit) and RMSEA (values $\leq .06$ and $\leq .08$ indicate good and acceptable model fit, respectively; Brown, 2015).

Significant SB χ^2 difference tests of the less and more restricted models indicated that the more stringent invariance should be rejected. In addition, because this test is overly sensitive to small deviations in the model in large samples, alternative difference tests were used as well (Chen, 2007; Cheung & Rensvold, 2002). For large samples, values $\geq -.010$ for the robust statistic ΔCFI, supplemented by values $\geq .015$ for ΔRMSEA, indicated that the more stringent invariance should be rejected (Chen; Cheung & Rensvold). A model was accepted if the majority of fit measures were acceptable and superior to those of alternative models.

Results

Four-Group Test of Measurement Invariance

The original two-factor model (Model_x.0) resulted in acceptable fit values for all groups (Table 2), but modification indexes suggested allowing for correlated residuals for items 2 and 4 as well as for items 6 and 7 (Model_a.0; Table 2). Items 2 and 4 both concerned anticipation of a math test and were part of the original MEA factor, whereas items 6 and 7 both concerned listening to math explanations and were part of the original LMA factor. Model_a.0 had acceptable to good fit for all groups and all factor loadings were significant. Since the correlated residuals did not conflict with the underlying factors and post hoc modeling is often necessary to obtain well-fitting models (Byrne et al., 1989), the correlated residuals were included. It was concluded that a factor model with an equal form fits the data in all groups and Model_a.0 was subsequently used to test for configural invariance.

Fitting the equal form model (Model_a.1) in all four groups simultaneously resulted in acceptable fit values, and all factor loadings were significant in all groups. However, for female university students, it was tested whether loading item 5 (i.e., homework) on the LMA factor

Table 2. Model Fit Statistics of Multi-Group CFA.

Model		SB χ^2	*df*	Δ*df*	$\Delta\chi^2$	*p*	CFI	ΔCFI	RMSEA (90% CI)	CFit	ΔRMSEA
Per subgroup											
Model_x.0[a]	Secondary males	116.49***	26				.934		.071 (.058–.084)	.005	
Model_x.0[a]	Secondary females	154.94***	26				.923		.078 (.067–.091)	.000	
Model_x.0[a]	University males	50.14**	26				.952		.073 (.041–.103)	.106	
Model_x.0[a]	University females	101.49***	26				.952		.080 (.064–.097)	.001	
Model_a.0	Secondary males	45.03**	24				.985		.035 (.019–.051)	.935	
Model_a.0	Secondary females	63.05***	24				.977		.045 (.032–.059)	.711	
Model_a.0	University males	35.27	24				.978		.052 (.000–.086)	.436	
Model_a.0	University females	79.18***	24				.965		.071 (.054–.089)	.022	
Multi-group analyses											
Equal form (configural)											
Model_a.1		214.71***	96				.977		.048 (.040–.057)	.625	
Model_b.1[b]	Cross loading 5	187.48***	95				.982		.043 (.034–.052)	.907	
Equal factor loadings (metric)											
Model_b.2		218.22***	115	20	30.85	.057[c]	.980	−.002[c]	.041 (.033–.049)	.964	−.002[c]
Equal indicator intercepts (scalar)											
Model_b.3.a		461.03***	136	21	330.13	<.001[d]	.936	−.044[d]	.067 (.060–.074)	.000	.026[d]
Model_b.3.b	Intercept 3 free	347.50***	133	18	168.57	<.001[d]	.958	−.022[d]	.055 (.048–.062)	.119	.014[d]
Model_b.3.c	Intercepts 3,8 free	274.01***	130	15	67.11	<.001[d]	.972	−.008[d]	.046 (.038–.053)	.828	.005[d]
Equal residuals											
Model_b.4.a		467.19***	157	27	169.66	<.001[e]	.939	−.033[e]	.061 (.055–.067)	.003	.015[e]
Model_b.4.b	Residual 1 free	429.87***	154	24	142.29	<.001[e]	.946	−.026[e]	.058 (.052–.064)	.022	.012[e]
Model_b.4.c	Residuals 1,7 free	403.30***	151	21	119.15	<.001[e]	.950	−.022[e]	.056 (.049–.063)	.067	.010[e]

Note. CFI ≥.95 indicates good model fit. For RMSEA, values ≤.06 and ≤.08 indicate good and acceptable model fit, respectively. CFit = test of close fit (RMSEA ≤.05). $\Delta CFI = CFI_{constrained} - CFI_{unconstrained}$. $\Delta RMSEA = RMSEA_{constrained} - RMSEA_{unconstrained}$.

** $p < .01$. *** $p < .001$.

[a]Model_x.0 does not include correlated residuals; all other models did.

[b]Model_b included a cross loading for item 5 on LMA for female university students only.

[c]Compared to Model_b.1.

[d]Compared to Model_b.2.

[e]Compared to Model_b.3.c.

would improve the model, similar to previous studies (Cipora et al., 2015; Schillinger et al., 2018) and as supported by modification indices. An equal form model (Model_b.1), that allowed for the estimation of loadings for item 5 on both LMA and MEA for female university students only, had improved fit compared to Model_a.1. Also, all factor loadings in Model_b.1 were significant, and there was no obvious source of misfit according to the modification indices. Alternatively, the cross loading was allowed in all groups, but this model was unacceptable because the latent variable covariance matrix was positive definite for the male secondary school students. Therefore, it was concluded that the conditions of configural variance were mostly met and that a similar factor structure could be applied for all groups (except for the cross-loading of item 5 in the female university group). Model_b.1 was accepted and used to test metric invariance, SB χ^2 = 187.48, CFI = .982, RMSEA = .043.

Next, equal factor loadings were constrained for all groups except for the factor loadings of item 5 for female university students (Model_b.2). Compared to Model_b.1, Model_b.2 did not result in significantly worse fit (Table 2). Therefore, Model_b.2 was accepted, and it was concluded that metric invariance across groups was acceptable, although factor loadings of item 5 on both factors were included for female university students only and were not subjected to equality constraints.

Next, Model_b.2 was used to evaluate scalar invariance, by additionally constraining the indicator intercepts for groups (Model_b.3.a). Compared to Model_b.2, Model_b.3.a resulted in significantly worse fit, indicating that constraining all intercepts to be equal across subgroups was not tenable. As recommended, we tested partial invariance of intercepts by freeing estimation of intercepts for one item at the time (Byrne et al., 1989; Cheung & Rensvold, 2002). Freeing the intercept of item 3 in Model_b.3.b (slightly) worsened the model fit, but freeing the intercepts of both items 3 and 8 in Model_b.3.c did not significantly worsen model fit compared to Model_b.2. Hence, Model_b.3.c was accepted, and it was concluded that scalar invariance across the four groups was partially tenable, SB χ^2 = 274.01, CFI = .972, RMSEA = .046.

Finally, Model_b.3.c was used to test for equivalence of residual variance by additionally constraining residuals for each item to be equal across groups. Compared to Model_b.3.c, all models with equal residuals resulted in significantly worse fit based on chi-square and ΔCFI and just acceptable worsened fit based on ΔRMSEA (Model_b.4.a: all residuals constrained to be equal; Model_b.4.b, Model_b.4.c: part of the residuals constrained). Therefore, models with equal

Table 3. Factor Loadings of Model b.3.c: Equal Factor Loadings Across Groups.

	Factor	
	Learning Math Anxiety	Math Evaluation Anxiety
AMAS1	.43	.00
AMAS2	.00	.77
AMAS3	.77	.00
AMAS4	.00	.83
AMAS5	.00 (.45^{FU})	.72 (.35^{FU})
AMAS6	.56	.00
AMAS7	.64	.00
AMAS8	.00	.70
AMAS9	.51	.00

Note. ^{FU} indicates deviant estimate for female university students.

residuals were rejected, and Model_b.3.c was the final model. Table 3 shows that factor-indicator loadings in Model_b.3.c were positive and substantial.

Summarized, the multi-group CFA resulted in a two-factor model with equal factor loadings except for item 5 for female university students. Equality restrictions held on 7 out of 9 equal intercepts (i.e., intercepts of items 3 and 8 differed between groups). Hence, for the first research aim, we conclude that metric and scalar invariance across sexes and age-groups were partially supported. Equivalence of residual variances was rejected, but this is not strictly necessary for comparing group means. Therefore, it was concluded that minimal conditions to meaningfully comparing group means on AMAS were met (Brown, 2015; Byrne et al., 1989; Cheung & Rensvold, 2002). For female university students, mean scores on the LMA subscale should also include item 5.

Reliability, Norms, and Group Comparisons for AMAS scores

Cronbach's alphas for total AMAS scores (α = .88, 95%-confidence interval: .87–.89), LMA scores (α = .80, 95%-confidence interval: .78–.81 excluding female university students; α = .84, 95%-confidence interval: .81–.86 for female university students), and MEA scores (α = .82, 95%-confidence interval: .81–.84) indicated good internal consistency (Evers et al., 2009). Appendix A contains means, standard deviations, percentile scores, and internal consistencies by subgroup.

To test for group differences on AMAS, a 2 Sex (females vs. males) x 2 Age-group (secondary school vs. university students) ANOVA was performed for total AMAS score, and a 2 Sex x 2 Age-group MANOVA for LMA and MEA subscales simultaneously. Results revealed significant main effects of sex for Total AMAS, F (1, 2129) = 90.59, p < .001, η^2 = .04, and subscales, F (2, 2128) = 56.99, p < .001, η^2 = .04 for LMA and η^2 = .05 for MEA, significant main effects of age-group for Total AMAS, F (1, 2129) = 60.27, p < .001, η^2 = .03, and subscales, F (2, 2128) = 51.79, p < .001, η^2 = .02 for LMA and η^2 = .05 for MEA. The main effects were qualified by significant interaction effects between sex and age-group for Total AMAS, F (1, 2129) = 14.53, p < .001, η^2 = .01, and subscales, F (2, 2128) = 16.53, p < .001, η^2 = .02 for LMA and η^2 = .01 for MEA. Post hoc independent t-tests with Bonferroni correction (p-value = .003) revealed that in university students the difference between females and males was larger than in secondary school students. Female university students' AMAS scores were the highest level of all groups (Figure 1; Table 4 shows p-values and effect sizes). The main effects indicated that females reported higher math anxiety than males and that university students reported significantly higher math anxiety than secondary school students. For the second research aim, we conclude that internal consistency of AMAS scores was good, and that AMAS scores differed between sexes and age-groups, which necessitated supplying norms for each subgroup separately.

Validity Evidence of AMAS scores

The CFAs above support the interpretation of AMAS scores as anxiety for learning math and math evaluation. To examine further validity evidence, Pearson correlations were calculated between AMAS total and subscale scores and outcome measures (Table 5). Absolute correlations <.3 were interpreted as weak, between .3 and .7 as moderate, and ≥.7 as strong (Dancey & Reidy, 2007). Convergent validity was supported, as AMAS scores correlated weakly negatively with math grade, r = −.29, moderately negatively with math self-concept, r = −.61, and math self-efficacy, r = −.48, and weakly to moderately positively with trait anxiety, r = .34. Discriminant validity was also supported, because correlations between AMAS scores and anxiety and performance in English were very weak, r = .09; r = .11, respectively. Lastly, the pattern of correlations was

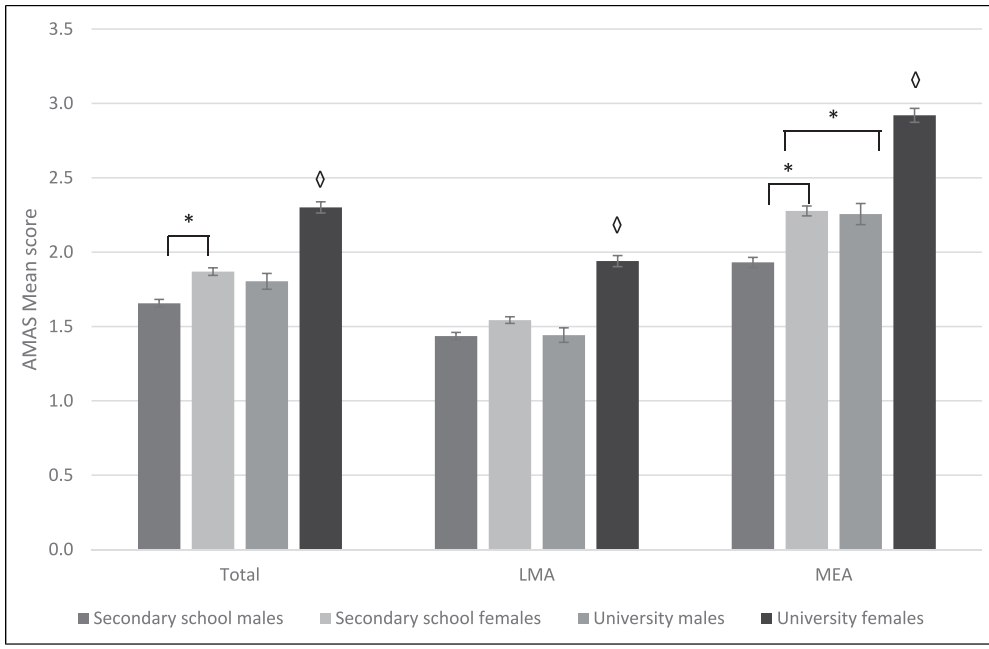

Figure 1. Mean scores for Total AMAS, and LMA and MEA subscales per subgroup.
Note. Error bars show +/− 1 standard error of the group means. (*) indicates a significant group difference for $p < .001$. (◊) indicates that female university students differed significantly from all other groups ($p < .001$). Effect sizes are in Table 4.

Table 4. Post Hoc Comparisons of Mean AMAS Scores Between Subgroups (Total and Subscales) by Means of Independent T-tests.

Scale	Group		Secondary Males		Secondary Females		University Males	
			p	*d*	*p*	*d*	*p*	*d*
Total	Secondary	Males						
		Females	<.001	0.30				
	University	Males	.097	0.21	>.100	0.09		
		Females	<.001	0.86	<.001	0.57	<.001	0.66
LMA	Secondary	Males						
		Females	.014	0.17				
	University	Males	>.100	0.01	.440	0.16		
		Females[a]	<.001	0.70	<.001	0.56	<.001	0.70
MEA	Secondary	Males						
		Females	<.001	0.38				
	University	Males	<.001	0.36	>.100	0.02		
		Females	<.001	1.05	<.001	0.66	<.001	0.69

Note. p = p-value of the post hoc test for group differences. d = Cohen's d.
[a]LMA included item 5 as well but for female university students only.

Table 5. Pearson Correlations Between AMAS (Total and Subscales) and Outcome Measures.

		AMAS			
	Study	Total	LMA	MEA	N
Math grade	All	−.29***	−.24***	−.28***	2097
Trait anxiety	2	.34***[a]	.24**	.38***	153
Math self-concept	3	−.61***	−.48***	−.61***	354
Math self-efficacy	3	−.48***	−.43***	−.44***	354
English grade	2	.11***	.09**	.12***	1147
English anxiety	2	.09[a]	.04	.12	148

[a]Also reported in Schmitz et al. (2019).
p <.01;*p <.001.

similar for the subscales. Note that the correlation between LMA and MEA subscales was strong, $r = .72$, $p < .001$. Regarding the third research aim, we conclude that significant relations with associated constructs and insignificant relations with different constructs provide further validity evidence for the Dutch AMAS scores.

Discussion

The current study's aims were to investigate measurement invariance, reliability, sex and age differences, and evidence for validity of Dutch AMAS scores in a sample of female and male secondary school and university students. The original two-factor structure was replicated (Hopko et al., 2003), supporting the interpretation of Dutch AMAS scores as stemming from "Learning Math Anxiety" and "Math Evaluation Anxiety". The associations with scores on related constructs (trait anxiety, self-concept, and self-efficacy) and the independence of unrelated constructs (performance and anxiety for learning a foreign language) provide further validity evidence of AMAS scores.

The two-factor structure was validated in all subgroups and both metric and scalar invariance were largely supported although the factor structure for female university students did include a cross-loading for item 5. Because of the partial support for metric and scalar invariance, group comparisons of AMAS scores seemed valid. The higher AMAS scores of females compared to males are consistent with most findings with other translations (Cipora et al., 2015; Hill et al., 2016; Hopko et al., 2003; Primi et al., 2014; Schillinger et al., 2018 but see Vahedi & Farrokhi, 2011). The result that university students had higher AMAS scores than secondary school students is opposite to earlier results (Primi et al., 2014). Possibly, the current results are affected by the remarkably high scores of the female university students. Female Psychology students may represent a selected group with relatively strong feelings of responsibility and low interest and confidence in math (Dempster & McCorry, 2009). Future research into causes of gender and age differences is necessary, taking into account the relationship with state math anxiety (Sachisthal et al., 2021).

The validity evidence, together with the observed high internal consistency and the norms supplied, constitutes a sound base for investigating usability of the AMAS in practice. The AMAS may provide a quick screening of math anxiety, which can be a starting point for treatment of math anxiety. It should however be noted that the relation between AMAS scores and math performance is robust but correlational and moderate in size at most (Carey et al., 2016), necessitating the inclusion of moderators, such as math motivation (Wang et al., 2018).

The current study suffered from several limitations. First, data on math self-concept, math self-efficacy, trait anxiety, and performance and anxiety for foreign language learning were available only for secondary school students, limiting the investigation of convergent validity. Second, sample sizes of the groups differed, specifically male university students were underrepresented, which could have resulted in smaller differences in fit indices and, consequently, the test could have failed to reject invariance (Chen, 2007). However, given good overall model fit and low values of modification indices, there is no reason to assume such a failure. Third, demographics should be assessed after completing the AMAS, which was the case in Study 3 but not in Studies 1 and 2. Indicating one's sex before answering math anxiety questions may trigger gender-stereotype images of math and biased responses (Steele & Ambady, 2006). However, note that the AMAS was never administered immediately after indicating demographics.

Recommended steps for future research are testing invariance of between-group differences on latent means (Brown, 2015; Cheung & Rensvold, 2002) and investigating test–retest reliability. Further, research may focus on tests of measurement invariance across countries. Although cross-national comparisons to date show similarities in factor structure and associations with related constructs (see also Cipora et al., 2015), such comparisons require a test of measurement invariance.

In conclusion, the partial metric and scalar invariance and the validity and reliability evidence support interpreting the Dutch AMAS scores of male and female secondary school and university students equally, as originating from anxiety for learning math and math evaluation anxiety. The partial metric and scalar invariance also support group comparisons, and the observed differences in Dutch AMAS scores between sexes and age groups may guide the design of school practices and interventions aimed at reducing math anxiety.

Appendix A

AMAS Norms and Group Comparisons

Percentiles, Means, Standard Deviations, and Internal Consistencies per Subgroup for AMAS

Scale	Group		Percentile				Mean	*SD*	α
			50	75	90	95			
Total	Secondary	Males	1.44	2.00	2.67	3.11	1.66	0.70	.88
		Females	1.67	2.22	2.89	3.33	1.87	0.71	.86
	University	Males	1.67	2.22	2.78	3.11	1.80	0.70	.89
		Females	2.22	2.89	3.44	3.70	2.30	0.80	.88
LMA	Secondary	Males	1.20	1.60	2.40	3.00	1.44	0.65	.82
		Females	1.40	1.80	2.40	3.00	1.54	0.64	.77
	University	Males	1.20	1.60	2.40	3.00	1.44	0.64	.86
		Females[a]	1.67	2.50	3.17	3.50	1.94	0.79	.86
MEA	Secondary	Males	1.75	2.50	3.25	3.75	1.93	0.88	.80
		Females	2.25	2.75	3.75	4.00	2.28	0.95	.79
	University	Males	2.25	3.00	3.58	4.00	2.26	0.94	.84
		Females	3.00	3.75	4.25	4.50	2.92	0.99	.82

Note. LMA includes items 1, 3, 6, 7, and 9; MEA includes items 2, 4, 5, and 8.
[a]LMA included item 5 as well but for female university students only.

Acknowledgments

We are grateful to Maien Sachisthal and co-contributors for providing data on AMAS (Study 3). We thank all students, adolescents, parents, and schools for their participation. We thank all (under)graduate students for their assistance in school recruitment and data collection.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Brenda R. J. Jansen https://orcid.org/0000-0001-9262-933X

Notes

1. Data were included in the current study.
2. See Online Supplementary Material 1 for detailed information on data cleaning and descriptive statistics per study.
3. Exploratory, AMAS scores were related to other variables in these studies: see Online Supplementary Material 2.

References

American Educational Research Association (AERA) (2014). *American psychological association, & national council on measurement in EducationStandards for educational and psychological testing*. American Educational Research Association.

Bakker, F. C., van Wieringen, P. C. W., van der Ploeg, H. M., & Spielberger, C. D ((1989). *Zelf-beoordelings-vragenlijst voor kinderen*. ZBV-K. Swets & Zeitlinger.

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

Brown, J. L., & Sifuentes, L. M. (2016). Validation study of the abbreviated math anxiety scale: Spanish adaptation. *Journal of Curriculum and Teaching*, *5*(2), 76–82. http://dx.doi.org/10.5430/jct.v5n2p76

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. http://doi.org/10.1037/0033-2909.105.3.456

Carey, E., Hill, F., Devine, A., & Szücs, D. (2016). The chicken or the egg? The direction of the relationship between mathematics anxiety and mathematics performance. *Frontiers in Psychology*, *6*(1987), 1–7 http://doi.org/10.3389/fpsyg.2015.01987

Caviola, S., Primi, C., Chiesi, F., & Mammarella, I. C. (2017). Psychometric properties of the Abbreviated Math Anxiety Scale (AMAS) in Italian primary school children. *Learning and Individual Differences*, *55*, 174–182. http://doi.org/10.1016/j.lindif.2017.03.006

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. http://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255. http://doi.org/10.1207/S15328007SEM0902_5

Cho, K. W. (2022). Measuring math anxiety among predominantly underrepresented minority undergraduates using the Abbreviated Math Anxiety Scale. *Journal of Psychoeducational Assessment*. Advance online publication. http://doi.org/10.1177/07342829211063286

Cipora, K., Szczygieł, M., Willmes, K., & Nuerk, H. C. (2015). Math anxiety assessment with the abbreviated math anxiety scale: Applicability and usefulness: Insights from the Polish adaptation. *Frontiers in Psychology*, *6*(1833), 1–16. http://doi.org/10.3389/fpsyg.2015.01833

Cito (2012). *OECD programme for international student assessment 2012Hoofdonderzoek PISA 2012*. PISA-2012-llvragenlijst-B-v1.0).

Dancey, C. P., & Reidy, J. (2007). *Statistics without maths for psychology*. Pearson Education.

Dempster, M., & McCorry, N. K. (2009). The role of previous experience and attitudes toward statistics in statistics assessment outcomes among undergraduate psychology students. *Journal of Statistics Education*, *17*(2), 2. https://doi.org/10.1080/10691898.2009.11889515

Dowker, A., Sarkar, A., & Looi, C. Y. (2016). Mathematics anxiety: What have we learned in 60 years? *Frontiers in Psychology*, *7*(508), 1–16. http://doi.org/10.3389/fpsyg.2016.00508

Evers, A. V. A. M., Lucassen, W., Meijer, R. R., & Sijtsma, K. (2009). *COTAN Beoordelingssysteem voor de kwaliteit van tests [COTAN Assessment system for the quality of tests]*. Netherlands Institute of Psychologists.

Hill, F., Mammarella, I. C., Devine, A., Caviola, S., Passolunghi, M. C., & Szűcs, D. (2016). Maths anxiety in primary and secondary school students: Gender differences, developmental changes and anxiety specificity. *Learning and Individual Differences*, *48*, 45–53. http://doi.org/10.1016/j.lindif.2016.02.006

Hopko, D. R., Mahadevan, R., Bare, R. L., & Hunt, M. K. (2003). The abbreviated math anxiety scale (AMAS) construction, validity, and reliability. *Assessment*, *10*(2), 178–182. http://doi.org/10.1177/1073191103010002008

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, *70*(2), 125–132. http://doi.org/10.1111/j.1540-4781.1986.tb05256.x

Kuncel, N. R., Credé, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research*, *75*(1), 63–82. https://doi.org/10.3102/00346543075001063

Lee, J. (2009). Universals and specifics of math self-concept, math self-efficacy, and math anxiety across 41 PISA 2003 participating countries. *Learning and Individual Differences*, *19*(3), 355–365. http://doi.org/10.1016/j.lindif.2008.10.009

Ma, X. (1999). A meta-analysis of the relationship between anxiety toward mathematics and achievement in mathematics. *Journal of Research in Mathematrics Education*, *30*(5), 520–540. http://doi: 10.2307/749772

Muthén, L. K., & Muthén, B. O. (2015). *Mplus user's guide*. Muthén & Muthén.

OECD (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD Publishing. https://doi.org/10.1787/9789264190511-en

Primi, C., Busdraghi, C., Tomasetto, C., Morsanyi, K., & Chiesi, F. (2014). Measuring math anxiety in Italian college and high school students: Validity, reliability and gender invariance of the Abbreviated Math Anxiety Scale (AMAS). *Learning and Individual Differences*, *34*, 51–56. http://doi.org/10.1016/j.lindif.2014.05.012

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*(1), 71–90. http://doi.org/10.1016/j.dr.2016.06.004

Richardson, F. C., & Suinn, R. M. (1972). The mathematics anxiety rating scale: Psychometric data. *Journal of Counseling Psychology*, *19*(6), 551–554. http://doi.org/10.1037/h0033-456

Sachisthal, M. S., Raijmakers, M. E., & Jansen, B. R. (2021). Trait and state math EAP (emotion, appraisals and performance) profiles of Dutch teenagers. *Learning and Individual Differences*, *89*(1), 102029. https://doi.org/10.1016/j.lindif.2021.102029

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*(2), 243–248. http://doi.org/10.1007/s11336-009-9135-Y

Schillinger, F. L., Vogel, S. E., Diedrich, J., & Grabner, R. H. (2018). Math anxiety, intelligence, and performance in mathematics: Insights from the German adaptation of the Abbreviated Math Anxiety Scale (AMAS-G). *Learning and Individual Differences*, *61*(1), 109–119. http://doi.org/10.1016/j.lindif.2017.11.014

Schmitz, E. A. (2020). *Missing factors in math anxiety: The role of emotional components, math behaviour, and cognitive biases in adolescents' math anxiety.* [Doctoral dissertation, University of Amsterdam].

Schmitz, E. A., Jansen, B. R. J., Wiers, R. W., & Salemink, E. (2019). Do implicitly measured math-anxiety associations play a role in math behavior? *Journal of Experimental Child Psychology*, *186*(1), 171–188. http://doi.org/10.1016/j.jecp.2019.05.013

Spielberger, C. D., Edwards, C. D., Lushene, R. E., Montuori, J., & Platzek, D. (1973). State-trait anxiety inventory for children. In *Preliminary manual*. Consulting Psychologist Press.

Steele, J. R., & Ambady, N. (2006). Math is Hard!" The effect of gender priming on women's attitudes. *Journal of Experimental Social Psychology*, *42*(4), 428–436. https://doi.org/10.1016/j.jesp.2005.06.003

Sticca, F., Goetz, T., Bieg, M., Hall, N. C., Eberle, F., & Haag, L. (2017). Examining the accuracy of students' self-reported academic grades from a correlational and a discrepancy perspective: Evidence from a longitudinal study. *Plos One*, *12*(11), Article e0187367. https://doi.org/10.1371/journal.pone.0187367

Suárez-Pellicioni, M., Núñez-Peña, M. I., & Colomé, À. (2016). Math anxiety: A review of its cognitive consequences, psychophysiological correlates, and brain bases. *Cognitive, Affective & Behavioral Neuroscience*, *16*(1), 3–22. http://doi.org/10.3758/s13415-015-0370-7

Vahedi, S., & Farrokhi, F. (2011). A confirmatory factor analysis of the structure of Abbreviated Math Anxiety Scale. *Iranian Journal of Psychiatry*, *6*(2), 47–53.

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. http://doi.org/10.1080/17405629.2012.686740.

Wang, Z., Shakeshaft, N., Schofield, K., & Malanchini, M. (2018). Anxiety is not enough to drive me away: A latent profile analysis on math anxiety and math motivation. *Plos One*, *13*(2), Article e0192072. https://doi.org/10.1371/journal.pone.0192072

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practice*, *29*(3), 39–47. http://doi.org/10.1111/j.1745-3992.2010.00182.x