Intelligence 38 (2010) 433–446

Contents lists available at ScienceDirect

Intelligence

Sex differences in adults' motivation to achieve

Sophie van der Sluis [a,*], Anna A.E. Vinkhuyzen [b], Dorret I. Boomsma [b], Danielle Posthuma [a,c]

[a] Functional Genomics, Center for Neurogenomics and Cognitive Research (CNCR), VU University Amsterdam, de Boelelaan 1085, 1081 HV, Amsterdam, The Netherlands
[b] Department of Biological Psychology, VU University Amsterdam, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands
[c] Medical Genomics, VU Medical Centre, van der Boechorststraat 7, 1081 BT, Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 26 August 2009
Received in revised form 8 March 2010
Accepted 20 April 2010
Available online 14 May 2010

Keywords:
Sex differences
Multi-group covariance and mean structure analysis
Motivation
Achievement

ABSTRACT

Achievement motivation is considered a prerequisite for success in academic as well as non-academic settings. We studied sex differences in academic and general achievement motivation in an adult sample of 338 men and 497 women (ages 18–70 years). Multi-group covariance and means structure analysis (MG-CMSA) for ordered categorical data was used to establish the location of possible sex differences, i.e., on the level of the latent factors or on the level of the observed items (i.e., sex-related item bias). Five of the 28 achievement motivation items showed severe bias with respect to sex, exemplifying the usefulness of MG-CMSA in locating the source of sex differences. The Academic Achievement Motivation scale consisted of two latent factors: Dedication and Persistence. Sex differences were observed for the factor Dedication only, with women showing more dedication towards their academic work than men. The General Achievement Motivation scale consisted of five latent factors: Pressure, Accomplishment, Work Approach, Future Orientation, and Competition. Sex differences were significant for the factor Future Orientation, with women contemplating less about the future than men, and a trend towards significance ($p = .06$) was observed for the factor Competition, with women being less actuated by competitive motives than men. These results suggest that sex-related item bias merits attention in achievement motivation research, but that men and women still differ in aspects of achievement motivation when biased items are eliminated from the analyses.

1. Introduction

Achievement motivation is considered a prerequisite for success, not only in academic, but also in sports- and job-related situations. In academic settings, the interest in motivation is partly inspired by the notion that students' motivation, operationalized, e.g. as their competency beliefs and value beliefs, could be more malleable than their cognitive ability, and as such could prove to be a potential lead for the educational system for improving learning and achievement processes in students (e.g., Spinath, Spinath, Harlaar, & Plomin, 2006).

Sex differences in achievement motivation have been studied widely (cf. Meece, Glienke, & Burg, 2006). In the context of academic achievement, gender role stereotypes are confirmed when motivation is studied domain-specifically, with boys being more confident and interested in mathematics and science compared to girls, while girls prefer, and feel more confident about language-related domains compared to boys. Researchers have studied whether these sex differences in motivation can predict sex differences in academic achievement (e.g., Steinmayr & Spinath, 2008; Freudenthaler, Spinath, & Neubauer, 2008). In all these studies, motivation-related items and subscale scores are compared directly between boys and girls. It has, however, never been verified whether these items or subscales are actually directly comparable, i.e., are measurement invariant across sex (see below). Yet, if the factor structure of a motivational instrument is not equal in boys and girls, differences in item-, or sumscores should be interpreted with caution. That is, when the measurement model is not

* Corresponding author.
E-mail address: sophie.van.der.sluis@cncr.vu.nl (S. der Sluis).

0160-2896/$ – see front matter © 2010 Published by Elsevier Inc.
doi:10.1016/j.intell.2010.04.004

equivalent across sex, differences between boys and girls in test scores do not necessarily reflect differences in achievement motivation. The present study is concerned with sex differences in academic achievement motivation and general achievement motivation in an adult sample, and explicitly deals with the question of whether the motivational instrument is measurement invariant across sex.

In 2006, Meece et al. (2006) published a comprehensive review of studies on sex differences in motivation. Studies on motivation have mainly focused on the school-going population, and report sex differences for motivation-related constructs such as expectations for success, causal attribution of failure/success, competency beliefs, value beliefs (i.e., perceived importance, usefulness, interest, and costs of academic activities), and self-efficacy judgements (i.e., one's confidence in learning, performing and succeeding academically). These sex differences mostly follow gender norms and stereotypes. Boys are more confident than girls with respect to math, science, and sports related abilities. In addition, boys value these abilities more highly, and attribute their success in these domains to ability while girls attribute their math or science related success mostly to effort and hard work. Contrarily, girls are more confident than boys in domains concerning verbal and language abilities, value these abilities more highly than boys do, and attribute their success in these domains to their own ability. Noteworthy, however, is that these findings are not consistent, and seem to depend not only on the achievement domain for which motivation is measured, but also on socioeconomic status, ethnicity, and actual ability level. In addition, all these studies focus on sex differences observed in adolescents, while sex differences in adults' work- or career-related achievement motivation have not received much attention.

In adolescents, studies focussed on academic achievement motivation, examining whether motivation predicts academic success independently of cognitive ability. Sex differences in academic achievement have been observed in many countries (USA: Epstein, Elwood, Hey, & Maw, 1998; Grant & Rong, 1999; Japan: Wong, Lam, & Ho, 2002; Belgium: van Houtte, 2004; Netherlands: de Knecht-van Eekelen, Gille, & van Rijn, 2007; CITO Terugblik en resultaten, 2009). The question logically following from this is whether these sex differences in achievement can be explained by sex differences in motivation.

In 17 year olds, Steinmayr and Spinath (2009) report that motivational aspects like hope for success, fear of failure, and need for achievement contributed to the prediction of academic achievement over and above general IQ and prior achievement. Although the additional effects of the motivational constructs to the prediction of academic success were smaller ($R^2<10\%$) than the effects of general intelligence ($R^2\approx12\%$) and of prior achievement ($R^2\approx24$–52%), the authors emphasized the importance of motivation because of its possible susceptibility to intervention.

In another study in 17 year olds, Steinmayr and Spinath (2008) observed sex differences for almost all motivation-related predictors included in their study. On average, girls expressed less hope for success, less work avoiding behaviour, and less confidence in their math-related ability, while at the same time rating math as less interesting, important and useful than boys. Boys, on the other hand, showed less fear of failure, less interest in learning as a goal in itself, and they were less confident about their German language-related ability, but also valued language as less important and useful than girls. Sex differences were however not apparent for performance avoidance (i.e., avoiding mistakes) and performance-approach (seeking other people's appreciation of one's own intellectual ability), and the relations of these motivational predictors to academic achievement were similar across sex.

Conversely, Freudenthaler et al. (2008) did report sex differences in prediction of academic achievement in 14 year olds. In boys, self-esteem, intrinsic motivation, performance avoidance, and school anxiety predicted academic achievement over and above IQ, while in girls, only work avoidance (i.e., doing no more than strictly required) and self-esteem did. In yet another study in 13-year old female students, Gagné and St Père (2001) observed no relation whatsoever between self-reported motivation on the one hand, and IQ and academic achievement on the other. In that study, only self-reported persistence was slightly related to academic achievement.

In sum, sex differences in motivational constructs, and sex differences in the relation between these motivational constructs on the one hand, and actual academic achievement on the other, have been found, but not consistently. Mediating effects of socioeconomic status, ethnicity, age, and actual ability level have been put forward as explanations for the inconsistencies. Another possible source of inconsistency, however, is that the tests and items used to measure motivation are not identical across studies, leaving open the possibility that the inconsistencies between studies are due to the use of different instruments. In addition, inconsistency may result when test- and item scores are not directly comparable between boys and girls, i.e., when items do not measure exactly the same constructs in boys and girls, e.g., because the connotation of the item is sex dependent. Such item bias could result in different relationships between items in boys and girls (and thus different underlying factor structures), and the sex differences observed on such biased items may not be indicative of sex differences in actual achievement motivation. If a test or items of a test are biased with respect to sex, then sex differences in the scores on this test are difficult to interpret.

One flexible framework for testing, and accommodating, group differences within the context of factor models is multi-group covariance and mean structure analysis (MG-CMSA; Sörbom, 1974; Little, 1997; Widaman & Reise, 1997). This method, which has been used in studies on group differences in intelligence (e.g., Dolan & Hamaker, 2001; Dolan et al., 2006; van der Sluis et al., 2006; van der Sluis et al., 2008; Wicherts et al., 2004), provides a model-based means to investigate the main source(s) of group differences. MG-CMSA allows one to test whether sex differences observed at the level of specific items are indeed a function of sex differences on the level of the latent trait(s) underlying the response to these items. When differences in scores on individual motivation items are not indicative of differences in actual motivation, then this may indicate that the item differences reflect a situation- or ability-specific difference between boys and girls, rather than a difference in motivation per se. In the context of MG-CMSA, items are considered 'biased' when the mean differences observed on the level of the item cannot be explained by mean differences on the level of the latent factor. The term 'bias' does not imply that the observed sex difference on the item is not real, but simply that the difference observed for the item is smaller or greater than the difference expected based on the means of the underlying factor, and can therefore not be taken as indicative of a sex difference in the latent trait. MG-CMSA can be used to locate such bias.

In addition, MG-CMSA allows one to evaluate and compare the fit of different models that reflect different hypotheses. In most research on motivation, researchers have used sum scores. The implicit assumption with respect to the sum score model is that the factor model underlying the test is 1-dimensional, and that all items are equally informative of the trait of interest. Whether such a highly-restricted model fits the data, i.e., describes the variance–covariance and means structure of the data adequately, is usually not tested. However, if that model does not describe the data adequately, then the sex differences are tested within the context of a poorly fitting model, which could result in incorrect conclusions with respect to the presence, and source, of sex differences.

The aim of the present paper is to investigate sex differences in academic achievement motivation and general achievement motivation in adults using MG-CMSA. Specifically, we investigate whether sex differences in achievement motivation test scores are really indicative of sex differences in the achievement motivation trait, or more likely of sex-related item bias. Below, we will first outline the MG-CMSA procedure for categorical data that we used to investigate the sources of sex differences in our motivational instruments. For convenience, results are presented separately for academic achievement motivation and general achievement motivation.

2. Method

2.1. Participants

All participants in this study were volunteer members of the Netherlands Twin Register (Boomsma et al., 2006) who participated in a larger ongoing study on the interplay between genes and environment on cognition. As part of this extended family study, participants were asked to fill out a questionnaire, which included the 28 questions on achievement motivation, which are used in the present study. At the time of publication, data were available from 284 families, including data from twins, and their siblings, and the parents, children, and partners of these twins and sibling (note that not all relations were represented in all families). The sample comprised 835 subjects in total: 338 men and 497 women. The overrepresentation of women in our study sample may affect the generalizability of this study's results to other populations (see Discussion). It does not, however, detract from the illustrative value of using MG-CMSA in the context of motivation research.

Because of the nature of the data collection, the age range was considerable (from 18 to 70, $M=45.37$, $SD=14.08$), but age did not differ significantly between men and women ($t(833)<1$, ns). Age was included as a covariate in all confirmatory factor analyses.

2.2. Instrument

The items used in this study were part of a larger questionnaire on life experiences, which was administered as part of the study on the interplay between genes and environment on cognition. The entire questionnaire took about 50 min to complete. The 28 multiple-choice achievement motivation items were adopted from the Dutch 'Prestatie Motivatie Test' (Dutch Achievement Motivation Test, DAMT, Hermans, 2004).[1] Ten of the 28 achievement motivation items focused on the academic achievement motivation subscale (AAM, e.g., "When I was in school, the demands that I made on myself concerning studying were very high / high / pretty high / low"; "Studying hard in school was something I did not like at all / did not like much / liked a lot"), while the other 18 focused on the general achievement motivation subscale (GAM, e.g., "The demands that I make on myself at work are very high / high / pretty high / not that high"; "The urge to surpass myself is very strong / pretty strong / not very strong").[2] All items were categorical in nature with 2 to 4 ordered answer options (see Appendix A for more example items). Negative items were recoded such that for all 28 items, higher scores reflect higher achievement motivation. The reliability of the AAM and the GAM subscales was .83, and .75, respectively.

Like many motivation instruments, AAM and GAM are self-report measures. In addition, our adult participants were asked to retrospectively evaluate their academic and general achievement motivation. Both the retrospective character and the self-reporting nature of the scales formed a potential source of bias in the evaluation of a person's motivation to achieve (see Discussion). It does not, however, detract from the illustrative value of using MG-CMSA in the context of motivation research.

3. Statistical analyses

3.1. Exploratory factor analysis

The factor structure of the two subscales of the DAMT has not been studied before. We therefore first conducted exploratory factor analyses for ordered categorical items to investigate the number of factors required to describe the structure of the AAM and GAM subscales, and, if multiple factors were required, to establish the pattern of factor loadings. These exploratory analyses were conducted in *Mplus* version 5 (Muthén & Muthén, 1998–2007), for men and women separately, and were followed by an oblique rotation (geomin).

[1] The original DAMT consists of three more subscales, tapping into positive and negative fear of failure and social desirability, but these were not included in the larger questionnaire for reasons of efficiency.

[2] The original general achievement motivation subtest consists of 20 rather than 18 items. Two items were, however, eliminated because they did not correlate with the other 18, which hindered the factor model fitting. As the content of these two items was also very different from the other 18 (one item asked whether one likes to organize things, the other asked the participant's opinion on the expression 'time is money'), we decided to discard these two items from all subsequent analyses.

3.2. Confirmatory factor analysis and testing for the presence of measurement invariance

To examine sex differences with respect to the latent factors of academic and general achievement motivation, one first needs to establish whether the AAM and GAM subscales are measurement invariant with respect to sex. Measurement invariance with respect to sex implies that the distribution of the observed scores of subjects i on an item j (y_{ij}), given a fixed level of the latent factor (η), depends on the score on the latent factor η only, and not on sex, i.e., $f(y_{ij}|\eta,\text{sex})=f(y_{ij}|\eta)$ (Mellenbergh, 1989). That is, given equal latent factor scores η, men and women should score similarly on item j. In the case of continuous items, and given normally distributed data, measurement invariance can be defined in terms of the means and variances of y_{ij} given η. With ordered categorical data the definition is however somewhat different.

In factor models for ordered categorical data, the observed scores for item y_{ijk}, i.e., the jth ordered categorical measure for the ith person in the kth group (where sex defines the two groups in the present paper), are assumed to be determined by the unobserved scores on the latent response variate y^*_{ijk}. These latent response variates are continuous in scale, and the observed measures y_{ijk} can be considered a categorized versions of the latent variates y^*_{ijk}, where the scores on the categorized items y_{ijk} depend on the threshold parameters $\nu_{jk(0 \ldots c-1)}$, where c is the number of categories, of the jth item in the kth group (Millsap & Yun-Tein, 2004).

Given p items, the scores on the vector of latent response variates for the ith person in the kth group, $\boldsymbol{y}^*_{ik,}$ are within each subgroup assumed to be multivariate normally distributed ($\boldsymbol{y}^*_{ik}$~MVN($\boldsymbol{\mu}^*_k$, $\boldsymbol{\Sigma}^*_k$)), where $\boldsymbol{\mu}^*_k$ is a $p\times 1$ vector of means of the latent response variates, and $\boldsymbol{\Sigma}^*_k$ is a $p\times p$ covariance matrix for the latent response variates, each estimated separately in each subgroup k.

Given the latent response variate y^*_{ijk}, the factor model is specified as:

$$y^*_{ijk} = \tau_{jk} + \lambda_{jk}\eta_{ik} + \varepsilon_{ijk}, \tag{1}$$

where τ_{jk} is a latent intercept parameter, $\boldsymbol{\lambda}_{jk}$ is a $r\times 1$ vector of factor loadings of the jth variate on the r factors, $\boldsymbol{\eta}_{ik}$ is the $r\times 1$ vector of factor scores of the ith person in the kth group, and ε_{ijk} denotes the jth unique factor score for that person. If $\boldsymbol{\varepsilon}_{ik}$ is the $1\times$p vector of unique factor scores, it is assumed that $\boldsymbol{\eta}_{ik}$~MVN($\boldsymbol{\kappa}_{k,}$ $\boldsymbol{\Psi}_k$), where $\boldsymbol{\kappa}_k$ is the $r\times 1$ vector of factor means and $\boldsymbol{\Psi}_k$ denotes the $r\times r$ factor covariance, and that $\boldsymbol{\varepsilon}_{ik}$~MVN(0, $\boldsymbol{\Theta}_k$), where $\boldsymbol{\Theta}_k$ denotes the $p\times p$ (usually diagonal) matrix of residual (or unique) variances, i.e., the variance not explained by the latent factors $\boldsymbol{\eta}$.

The model implied expected values for the vector of latent response variates $\mathbf{y}^*_{ik}$ are given as:

$$E(y^*_{ik}) = \mu^*_k = \tau_k + \Lambda_k\kappa_k, \tag{2}$$

and the model implied covariance matrix is given as:

$$\text{Cov}(y^*_{ik}) = \Sigma^*_k = \Lambda_k\Psi_k\Lambda'_k + \Theta_k, \tag{3}$$

where $\boldsymbol{\Lambda}_k$ is the $p\times r$ matrix of factor loadings, with $\boldsymbol{\Lambda}'_k$ denoting the transpose of this matrix.

Note that to begin with, all factor model parameters ($\boldsymbol{\tau}_k$, $\boldsymbol{\Lambda}_k$, $\boldsymbol{\kappa}_k$, $\boldsymbol{\Psi}_k$, $\boldsymbol{\Theta}_k$) are estimated separately in the different groups (as denoted by subscript k). However, not all parameters may be identified, especially when the observed items are ordered categorical.

To establish measurement invariance with respect to sex in a factor model for ordered categorical data, one needs to establish whether the relation between the observed item scores y_{ijk} (via the latent variates y^*_{ijk}) and the underlying latent factor(s) $\boldsymbol{\eta}$ is the same in men and women. Measurement invariance with respect to sex can be examined through a series of constraints on the model parameters (Meredith, 1993; Millsap & Yun-Tein, 2004), which are, to begin with, estimated separately in men and women.

To test whether the mean structure and the covariance structure of the AAM and GAM subscales were measurement invariant across sex, multi-group confirmatory factor analysis for ordered categorical data had to be carried out. Below we will give a short overview of the constraints required to identify the factor model, and to test for measurement invariance when data are categorical. We refer to Millsap and Yun-Tein (2004) for more details on and the rationale behind these constraints. All steps required to test for measurement invariance were previously described and discussed in detail by Horn and McArdle (1992) and Widaman and Reise (1997).

The first step (Model 1) in testing for measurement invariance concerned the test for 'configural invariance', i.e., the test of whether the pattern of factor loadings (and correlated residuals, if present) was the same in men and women, while the actual values of these parameters were allowed to differ across sex. Several constraints were required to identify this model. In all subsequent analyses, we chose the male group as a reference group. In this group, the latent intercepts $\boldsymbol{\tau}$ and the factorial means $\boldsymbol{\kappa}$ needed to be fixed to 0, and all thresholds $\boldsymbol{\nu}$ were estimated freely. In the women, however, we needed to constrain one threshold per item to be sex-invariant, i.e., to be equal to the threshold of the men. In addition, we needed to pick r reference items (i.e., one for each latent factor) for which the second threshold was constrained to be sex-invariant as well. All remaining thresholds were estimated freely in the women, just as the factorial means $\boldsymbol{\kappa}$, which were identified due to the constraints on the thresholds. The latent intercepts $\boldsymbol{\tau}$ were however fixed to 0 in women as well. As with continuous data, one needs to fix the arbitrary scale of the latent factor; we chose to fix the factorial variances to 1 in both groups. The categorical nature of the observed data requires one to also adopt a scale for the continuous latent variates underlying the categorical response data. To this end, the residual variances were fixed to 1 in the male reference group (i.e., the so-called theta parameterization in *Mplus*, see Muthén & Muthén, 1998-2007), but these parameters could then be estimated freely in the women (unless an item is dichotomous in nature, in which case its' residual variance needs to be fixed to 1 in the women as well). We refer to Millsap and Yun-Tein (2004) for an elaborate discussion of these constraints.

In the second step (Model 2), we tested for 'metric invariance'. Metric invariance implies that the relations between the observed items on the one hand and the latent factor on the other are the same across sex. The test for metric invariance thus involves constraining all factor loadings to be equal across sex. Note that metric invariance is a prerequisite for meaningful comparison of the latent factors across sex: only when the factor loadings are equal across sex, can we be sure that the latent factors themselves are identical, and thus comparable, between men and women. Metric invariance is said to be tenable when the equality constraints on the factor loadings do not result in a significant deterioration of the overall model fit. Note that as a result of these constraints on the factor loadings, fixation of the factorial variances in both groups became superfluous: the factorial variances remained fixed to 1 in the male reference group, but could now be estimated freely in the women.

In the third step (Model 3), we tested for 'strong factorial invariance'. Strong factorial invariance implies that the mean differences that are observed between men and women on the level of the observed items can all be accounted for by the latent factor, i.e., are indicative of mean differences on the latent trait of interest. The test for strong factorial invariance thus involves constraining all thresholds to be equal across sex. These constraints allowed free estimation of the factorial means of the female group, while the factorial means in the male reference group remained fixed to 0 for identification purposes. Modelled as such, the factorial mean of the women should be interpreted as deviation from the factorial means of the men (i.e., deviations from zero). Note that in this model, sex differences in observed scores y_{ijk} can only result from sex differences in factorial means, because, at this point in the model fitting sequence, these factorial means are the only parameters that differ across sex in the regression of the items on the latent factors. In other words, if the constraints implied by strong factorial invariance hold, i.e., do not lead to a significant deterioration of the model fit, then the assumption that the expected observed scores depend only on a subject's factor score and not on the subject's sex holds, i.e., $E(y_{ijk}|\eta,\text{sex}) = E(y_{ijk}|\eta)$. If these constraints do however result in a significant deterioration of the model fit, then the latent factors cannot account for the sex differences in observed scores, i.e., one or more of the differences in thresholds between men and women cannot be accounted for by the latent factors. Comparing men and women with respect to their latent factor means is only meaningful if strong measurement invariance holds. Those items, for which the sex differences observed on the level of the thresholds cannot be explained by sex differences on the level of the latent factor, are considered biased with respect to sex.

The fourth step (Model 4) tested for strict factorial invariance. Strict factorial invariance implies that the residual variances, i.e., the parts of the observed items that are not explained by, or related to, the latent factor, are also equal across sex. Strict factorial invariance thus involves constraining the residual variances to be also equal across sex. Note that because of the categorical nature of the items, the residual variances were fixed to 1 in the male reference group, and were estimated freely in the women. In the context of categorical data, the test for strict factorial invariance thus implies fixing the residual variances in women to 1 as well. If these constraints were tenable, we concluded that all sex differences with respect to the observed scores on the items, and the relations between the items, could be accounted for by sex differences on the level of the latent factor. Note however that for the comparison of threshold or factor means between men and women, strict factorial invariance is not required (i.e., strong factorial invariance suffices).

Finally (Model 5), when at least strong factorial invariance holds (i.e., the constraints in Model 4 are tenable), we were ready to test whether the factorial means were the same in men and women. Note that for reasons of identification, the factorial means were fixed to 0 in the male reference group, and were freely estimated in the women. The test for equal factor means thus involves fixing the factorial means of the women to zero as well. If this constraint resulted in a significant deterioration of the model fit, then we concluded that men and women differed with respect to the latent trait of interest (i.e., achievement motivation in the present study).

3.3. General model fitting strategies

For reasons of convenience in reporting results and estimation of parameters, all analyses were conducted separately for the academic achievement motivation (AAM) subscale and the general achievement motivation (GAM) subscale. Note that in theory, the factor structure and the model fitting results could be different for subsets of items, compared to the results for the complete item set, e.g. because items of the AAM subscale cannot load on the factors of the GAM subscales if they are analyzed separately. However, when the separate factor models of the AAM and the GAM were eventually combined in one overall model (the Total Model in the Results section), this model showed good fit, and no large modification indices (indices of local misfit in the model) or large residuals (i.e., parts not explained by the model). The choice to start with exploratory and confirmatory factor analyses in the two subscales separately turned therefore out to be justified.

All items were regressed on a standardized measure of age to correct for possible age effects.

Because the data were collected within families, the observations could not be considered independent. As treating within-family data as if they are independently distributed observations results in incorrect standard errors and incorrect χ^2 goodness of fit statistics, all analyses were performed in *Mplus* version 5 (Muthén & Muthén, 1998-2007), which computes corrected standard errors and Satora–Bentler scaled χ^2-tests with adjusted number of degrees of freedom, taking into account the dependence of the observations. The fit of nested models can then be compared through a weighted χ^2-difference test (Satorra, 2000). More restricted (i.e., nested) models are accepted if their fit is not significantly worse than the fit of the less restricted model, i.e., if the weighted χ^2-difference test (henceforth χ^2_{diff}) is not significant. Below, we will not report the scaled χ^2-values for each model, as these are not informative, but rather report the weighted χ^2_{diff} tests for the comparisons of competing models.

The fit of ensuing models to the data were also evaluated using the Root Mean Square Error of Approximation (RMSEA) and the Comparative Fit Index (CFI, Bentler, 1990; Bollen & Long, 1993; Jöreskog, 1993; Schermelleh-Engel, Moosbrugger, & Müller, 2003). The RMSEA is a measure of the discrepancy (i.e., error of approximation) between the covariance and mean structure implied by the fitted model, and the covariance and mean structure in the population. Calculating the discrepancy per degree-of-freedom, this fit index favours more parsimonious models. Generally, as a rule of thumb, RMSEA values <.05 are taken as indicative of good fit (i.e., good approximation), RMSEA values between .05 and .08 indicate acceptable fit, and values larger than .08 indicate poor fit (Browne & Cudeck, 1993; Schermelleh-Engel et al, 2003). The CFI is based on the comparison between the independence model, i.e., the model in which all variables are modelled as unrelated, and the user-specified model. The CFI, for which theoretically values range between 0 and 1.00, favours more parsimonious models, and takes on larger values when the difference between the independence model and the hypothesized model increases. Usually, values >.95 are taken to indicate good model fit, and values between .90 and .95 indicate acceptable fit (Hu & Bentler, 1999; Schermelleh-Engel et al., 2003).

The RMSEA and the CFI were used only as indication of the general fit of models, while the scaled χ^2-tests and weighted χ^2_{diff} tests were used specifically when testing the effects of the constraints required for measurement invariance. Modification indices, which express the expected drop in scaled χ^2 if constrained parameters are estimated freely, were used to detect local misfit in models.

Raw data maximum likelihood estimation was used to accommodate missingness (mean percentage of missingness across the entire 28-item DAMT was 1.47 % (SD = 1.05) with a maximum of 4.4% for one of the academic motivation items).[3]

For all analyses, α was set at .05.

4. Results

4.1. Preliminary analyses

Table 1 shows the endorsement rated in valid percentages for the 10 items of the AAM subscale for men and women separately. Effect size r is calculated as the Z-score obtained from a Mann–Whitney test (i.e., the non-parametric test comparing two independent groups with respect to their ranks scores on a categorical measure: the Z-score is a measure of whether the smallest sum of ranks deviates from the expected sum of ranks), divided by the square root of the total number of observations, i.e., $Z/\sqrt{N}$ (Rosenthal, 1991). Most effect sizes for the AAM items were small and positive, implying that women scored overall somewhat higher than men, i.e., were somewhat more motivated or more zealous. The largest effect size was observed for item AAM4 ("In school, people thought I was quite lazy/ not very diligent / diligent"), where women remembered themselves more often as being considered more zealous than men.

The polychoric correlations between the 10 AAM items are shown in Table 2 for men and women separately.

The endorsement rates (in valid percentages) for the 18 GAM items are shown in Table 3. The effect sizes for the GAM items were mostly small but the more sizable ones were negative, implying that women scored somewhat lower than men. The largest effect size was observed for item GAM4 ("As the manager of a factory you are often very busy and overworked. I would certainly not want such a job / would not readily accept such a job / would really like such a job".)

Table 1
Endorsement rates of the 10 academic achievement motivation (AAM) items for men and women separately.

	Men				Women				Effect size r
	Cat 1	Cat 2	Cat 3	Cat 4	Cat 1	Cat 2	Cat 3	Cat 4	
AAM1	8.9	50.3	40.8	–	6.7	42.5	50.8	–	.10
AAM2	59.9	40.1	–	–	58.6	41.4	–	–	.01
AAM3	21.0	61.0	18.0	–	16.4	62.9	20.7	–	.06
AAM4	28.1	25.4	46.5	–	20.7	8.0	71.3	–	.21
AAM5	9.7	44.8	40.9	4.5	4.5	34.1	52.9	8.5	.17
AAM6	36.7	47.4	15.9	–	27.0	53.9	19.1	–	.10
AAM7	22.5	42.3	29.1	6.0	18.5	37.6	34.1	9.9	.09
AAM8	61.8	31.5	6.7	–	54.7	34.3	11.0	–	.08
AAM9	14.3	31.3	40.4	14.0	18.1	26.7	41.9	13.3	−.01
AAM10	10.5	38.4	51.1	–	7.7	37.9	54.4	–	.04

Note: Number of ordered answer options varies across items (range: 2–4). Higher categories correspond to higher motivation. Effect size r is calculated as $Z/\sqrt{N}$, where Z is obtained in a Mann–Whitney test, and N is the effective sample size (men + women) for each individual item. Positive effect sizes denote higher academic achievement motivation for women.

The polychoric correlations between the 18 GAM items are shown in Table 4 for men and women separately. Important to note is that, although some correlations were higher than .35, many correlations between these categorical items were lower than .20.

Whether the small differences observed between men and women on the categorical items, were indicative of differences on the latent level, was further examined using multi-group covariance and means structure analysis (MG-CMSA). First, however, the factor structure of the AAM and the GAM was established using exploratory factor analysis.

4.2. Exploratory factor analyses (EFA)

Because the factor structure of the subscales AAM and GAM has not been studied before, exploratory factor analyses were conducted in order to get a first impression of the pattern of factor loadings. As explained before, analyses were conducted separately for the AAM and the GAM.

4.2.1. Academic achievement motivation

With respect to the AAM, an exploratory factor solution with two correlated factors showed a good fit in both men (CFI = .98, RMSEA = .036) and women (CFI = .98, RMSEA = .046). Table 5 shows the geomin rotated factor loadings of the 10 AAM items on the two correlated factors for men and women separately.

The items loading on the first factor all represent Dedication (willingness to study and allocate time to homework), while the items loading on the second factor mostly refer to focus or Persistence (the ease with which one could start and continue doing school work in spite of distraction). Item 1 loaded on both factors, and item 10 loaded mainly on factor 1 in men, and on factor 2 in women. Based on the content of these items and the model fit statistics, however, we choose to let these items load on the Dedication factor only in all subsequent confirmatory factor analyses. In these analyses, the Persistence factor was

[3] Note that missingness on some of the academic achievement motivation items was significantly related to the age of the participants, with missingness being more frequent in older subjects. This could suggest that questions about academic achievement motivation are more difficult to answer when the school years are in the remote past, or that academic training was less often granted to the older participants, rendering questions about e.g. homework unsuitable.

Table 2
Polychoric correlations between the 10 academic achievement motivation (AAM) items for men (below diagonal) and women (above diagonal) separately.

	AAM1	AAM2	AAM3	AAM4	AAM5	AAM6	AAM7	AAM8	AAM9	AAM10
AAM1		.37	.44	.41	.57	.34	.55	.39	.54	.55
AAM2	.41		.33	.29	.40	.44	.33	.33	.27	.23
AAM3	.34	.35		.46	.35	.27	.54	.42	.50	.35
AAM4	.49	.27	.46		.47	.28	.55	.47	.57	.32
AAM5	.49	.47	.34	.43		.48	.47	.57	.50	.52
AAM6	.35	.44	.37	.31	.45		.30	.42	.29	.39
AAM7	.38	.27	.57	.58	.44	.36		.44	.57	.44
AAM8	.38	.41	.37	.37	.50	.45	.38		.44	.37
AAM9	.53	.22	.39	.48	.38	.32	.47	.33		.50
AAM10	.47	.27	.39	.39	.37	.39	.38	.35	.48	

thus indicated by 4 items (items 2, 6, 8, and 9), and the Dedication factor by 6 items (items 1, 3, 4, 5, 7, and 10). This configuration of factor loadings was used for the multi-group CFA analyses, with the bold factor loading of Table 5 estimated freely, and all other factor loadings fixed to zero.

4.2.2. General achievement motivation

Exploratory factor analyses on the 18 items of the GAM subscale showed that a factor solution with 5 factors described the data structure adequately in both men (CFI = .96, RMSEA = .029) and women (CFI = .96, RMSEA = .032). Table 6 shows the geomin rotated factor loadings of the 18 items on the 5 correlated factors for men and women separately.

Based on the content of the items, factor 1 represents the extent to which subjects experience time pressure as a result of their work (Pressure; items 2, 4, 17 and 20), factor 2 represents the intrinsic motivation to accomplish goals and to surpass oneself (Accomplishment; items 1, 5, 6, 10, 11, and 20), factor 3 gives an indication of work approach or avoidance, i.e., how much subjects are inclined to work in general (Work Approach; items 3, 9, 14 and 15), factor 4 gives an indication of how future-oriented subjects are (Future Orientation; items 4, 7 and 16), and factor 5 represents the extrinsic motivation of subjects to compete with others and to earn respect (Competition; items 10, 12, 13, 15 and 19). The pattern of factor loadings of items 4, 5 and 9 was somewhat different for men and women, but based on the content of these items, it was decided to start with a confirmatory factor model in which item 5 loaded on the Accomplishment factor, item 9 on the Work Approach factor, and item 4 on both factors Pressure and Future. This configuration of factor loadings was used for the multi-group CFA analyses, with the bold factor loading of Table 6 estimated freely, and all other factor loadings fixed to zero.

Table 3
Endorsement rates of the 18 general achievement motivation (GAM) items for men and women separately.

	Men				Women				Effect size r
	Cat 1	Cat 2	Cat 3	Cat 4	Cat 1	Cat 2	Cat 3	Cat 4	
GAM1	2.4	15.7	56.4	25.5	1.4	22.2	48.0	28.4	−.01
GAM2	1.2	14.5	54.7	29.6	.6	13.3	57.6	28.5	.00
GAM3	19.6	55.2	25.2	–	13.9	64.2	21.8	–	.01
GAM4	22.0	49.7	28.3	–	33.8	55.7	10.5	–	−.21
GAM5	6.5	74.7	18.8	–	6.4	82.8	10.9	–	−.09
GAM6	13.4	39.2	38.3	9.2	9.9	51.1	32.3	6.7	−.05
GAM7	35.7	34.5	25.9	3.9	42.6	35.3	18.7	3.4	−.09
GAM9	5.3	24.9	69.7	–	9.0	34.2	56.8	–	−.13
GAM10	47.6	36.5	15.9	–	46.2	39.5	14.4	–	.00
GAM11	15.5	37.8	29.2	17.6	15.1	34.0	33.0	17.9	.03
GAM12	27.5	46.2	26.3	–	20.6	50.7	28.7	–	.06
GAM13	14.9	38.2	36.4	10.4	10.6	39.7	40.7	9.0	.03
GAM14	14.9	15.8	69.3	–	12.7	18.2	69.1	–	.00
GAM15	13.9	40.6	45.5	–	11.6	47.0	41.4	–	−.02
GAM16	13.6	50.1	36.2	–	22.8	55.8	31.4	–	−.03
GAM17	6.6	42.9	46.8	3.6	3.3	39.8	50.6	6.3	.08
GAM19	49.1	22.5	23.1	5.4	59.7	22.0	15.7	2.6	−.12
GAM20	4.8	37.0	49.4	8.7	5.1	33.7	50.3	10.8	.03

Note: Number of ordered answer options varies across items (range: 3–4). Higher categories correspond to higher motivation. Effect size r is calculated as $Z/\sqrt{N}$, where Z is obtained in a Mann–Whitney test, and N is the effective sample size (men + women) for each individual item. Positive effect sizes denote higher academic achievement motivation for women.

4.3. Multi-group covariance and means structure analysis (MG-CMSA)

4.3.1. Academic achievement motivation

The results and fit statistics of the multi-group CFA of the AAM items are presented in Table 7.

In Model 1, we tested for configural invariance, with 4 items loading on the Persistance factor, and 6 on the Dedication factor. No cross-loadings were modelled (i.e., congeneric structure), and all residual terms were modelled as uncorrelated. To correct for possible age effects, all 10 items were regressed on age in men and women separately. The CFI (.98) and the RMSEA (.06) indicated that Model 1 fitted the data well.

In Model 1a we tested whether the age regressions could be constrained to be equal in men and women, but this was not the case (Model 1a vs Model 1: $\chi^2_{\text{diff}}(6) = 30.19$, $p < .001$). In all subsequent models, age effects were therefore modelled separately in men and women. Note that this part of the model was saturated (i.e., all regressions on age were estimated) so that the age correction could not contribute to model misfit.

To test for metric invariance, all factor loadings were constrained to be equal in men and women in Model 2, and the factor variances were estimated freely in the women (and fixed to 1 for reasons of identification in the male reference group). The model fit did not deteriorate significantly as a result of these constraints (Model 2 vs Model 1: $\chi^2_{\text{diff}}(7) = 11.34$, ns), implying that metric invariance across sex was tenable for the AAM subscale.

In Model 3, strong factorial invariance was tested by constraining all thresholds to be equal across sex, and estimating the factor means freely in women, while these remained fixed to 0 in the male reference group for reasons of

Table 4
Polychoric correlations between the 18 general achievement motivation (GAM) items for men (below diagonal) and women (above diagonal) separately.

	GAM1	GAM2	GAM3	GAM4	GAM5	GAM6	GAM7	GAM9	GAM10	GAM11	GAM12	GAM13	GAM14	GAM15	GAM16	GAM17	GAM19	GAM20
GAM1		.30	.06	.31	.18	.41	.22	.06	.52	.24	.28	.20	-.02	.08	.20	.23	.38	.39
GAM2	.31		.10	.21	.18	.16	.14	.10	.27	.10	.18	.16	.06	.02	.20	.44	.10	.43
GAM3	.16	.08		.13	.19	.15	−.04	.16	.10	.10	.05	.16	.50	.22	.04	−.02	.12	.15
GAM4	.29	.30	.12		.25	.15	.25	.14	.28	.16	.11	.11	.09	.26	.36	.08	.34	.16
GAM5	.25	.15	.16	.27		.11	.06	.14	.10	.04	.04	.06	.10	.28	.16	−.02	.21	.16
GAM6	.36	.12	.13	.04	.02		.12	.16	.38	.16	.15	.06	.10	.10	.07	.11	.21	.16
GAM7	.25	.17	.16	.28	.26	.09		−.11	.23	.15	.08	.21	−.07	.20	.64	.06	.25	.06
GAM9	.13	.15	.04	.15	.09	.13	.11		.06	.06	−.03	.03	.43	.07	.15	−.09	−.02	.20
GAM10	.49	.17	.12	.25	.18	.35	.36	.02		.26	.33	.20	.07	.23	.32	.21	.50	.30
GAM11	.23	.20	.09	.21	.17	.08	.12	.14	.21		.11	.16	.13	.12	.23	.02	.18	.33
GAM12	.07	.05	.08	−.02	.02	−.04	.17	−.15	.22	−.00		.17	.01	.13	.05	.10	.45	.20
GAM13	.16	.02	.24	.14	−.05	.14	.25	−.01	.17	.06	.18		.17	.26	.18	.13	.26	.16
GAM14	.18	.17	.52	.23	.12	.07	.27	.22	.02	.19	.03	.15		.34	.15	−.03	−.01	.19
GAM15	.19	.03	.32	.33	.16	−.04	.22	−.01	.11	.09	.28	.28	.31		.36	.01	.34	.12
GAM16	.17	.11	.23	.32	.24	.09	.66	-.01	.34	.16	.14	.19	.27	.45		.12	.27	.21
GAM17	.09	.41	.03	.19	−.09	.02	.18	.03	.12	−.03	.05	−.05	.17	.04	.16		.09	.21
GAM19	.43	.22	.05	.33	.27	.13	.27	.01	.58	.13	.28	.37	.10	.33	.39	.17		.15
GAM20	.41	.44	.13	.27	.31	.18	.20	.19	.26	.33	.06	.15	.22	.12	.17	.27	.26	

identification. These constraints did however result in a significant deterioration of the model fit (Model 3 vs Model 2: $\chi^2_{\text{diff}}(9) = 23.61$, $p < .01$), implying that not all threshold differences observed between men and women could be accounted for by differences on the level of the factors. The modification indices indicated that the misfit was mainly due to item 4. Note that this is the diligence item for which the largest effect size was observed in the item-specific analyses (Table 1). In Model 3a, we constrained all thresholds equal across sex except the thresholds of item 4. This set of constraints did not result in a significant drop in model fit (Model 3a vs Model 2: $\chi^2_{\text{diff}}(8) = 9.23$, ns). For the AAM, strong factorial invariance was thus established for 9 out of 10 items, while the sex difference on item AAM4 was too large to be accounted for by the model, i.e., this item is biased in the context of this model. In the subsequent models, the thresholds for item 4 were therefore estimated freely in both groups. Note that free estimation of the thresholds for this item implies that this item no longer contributes to the estimation of the differences between men and women in the mean of the latent factor Dedication (Byrne, Shavelson, & Muthén, 1989). The mean of the Dedication factor was thus not biased, but directly comparable between men and women.

Strict factorial invariance was tested in Model 4 by restricting all residuals in the women to be equal to the residuals in the male reference group, i.e., equal to 1. The fit did not deteriorate significantly (Model 4 vs Model 3a: $\chi^2_{\text{diff}}(8) = 9.14$, ns), implying that strict factorial invariance was tenable. Table 8 shows the factorial correlations and factor means taken from Model 4 for men and women separately.

Given that the factor model was invariant across sex, we could subsequently meaningfully test whether men and women differed with respect to the means of the two factors, Persistence and Dedication. In Model 4, the factor means were fixed to 0 in the male reference group for reasons of identification, while they were freely estimated in women, such that these estimates can be considered deviations from the factor means of the men. In Model 5, the mean of the Persistence factor was fixed to 0 in women, which did not result in a significant drop in model fit (Model 5 vs Model 4: $\chi^2_{\text{diff}}(1) = 2.36$, ns), meaning that men and women did not differ significantly with respect to persistence. In Model 6, the

Table 5
Geomin rotated (oblique) exploratory factor solution for the 10 academic achievement motivation (AAM) items for men and women separately.

	Men ($N = 338$)		Women ($N = 497$)	
	Factor 1	Factor 2	Factor 1	Factor 2
	Dedication	Persistence	Dedication	Persistence
AAM1	**.313**	.412	**.374**	.405
AAM2	−.231	**.829**	.038	**.514**
AAM3	**.634**	.013	**.695**	−.017
AAM4	**.613**	.088	**.723**	.085
AAM5	**.637**	.096	**.629**	.086
AAM6	.056	**.692**	.000	**.823**
AAM7	**.878**	−.128	**.812**	−.004
AAM8	.013	**.629**	−.146	**.734**
AAM9	.005	**.662**	.185	**.516**
AAM10	**.377**	.260	**.192**	.497

Note: Factor loadings in bold print are estimated freely in the subsequent confirmatory multi-group covariance and means structure analyses.

Table 6
Geomin rotated (oblique) exploratory factor solution for the 18 general achievement motivation (GAM) items for men and women separately.

	Men (N = 338)					Women (N = 497)				
	Pressure	Accomplishment	Work Approach	Future Orientation	Competition	Pressure	Accomplishment	Work Approach	Future Orientation	Competition
GAM1	.081	**.563**	.074	−.017	.258	.037	**.732**	−.112	.005	.096
GAM2	**.787**	.015	−.025	−.078	.020	**.916**	−.010	.014	.016	−.004
GAM3	−.044	.070	**.664**	−.026	.029	.020	.023	**.559**	−.155	.141
GAM4	**.299**	.065	.087	**.221**	.113	**.053**	.151	.117	**.231**	.200
GAM5	.131	**.198**	−.046	.283	.015	.077	**.018**	.251	.044	.196
GAM6	−.128	**.456**	.058	−.026	.041	−.115	**.574**	.070	−.065	.001
GAM7	.042	.068	.005	**.691**	.017	−.008	.035	−.184	**.642**	.120
GAM9	.154	.254	**.031**	.077	−.205	−.006	.243	**.417**	.040	−.215
GAM10	.000	**.361**	−.189	.227	**.491**	.036	**.490**	−.021	.074	**.321**
GAM11	.104	**.284**	.091	.084	−.066	−.038	**.315**	.109	.117	.002
GAM12	.022	−.241	.112	−.026	**.493**	.142	.115	−.038	−.136	**.489**
GAM13	−.158	.022	.345	−.057	**.407**	.118	−.017	.175	.062	**.269**
GAM14	.192	.029	**.707**	.050	−.163	.026	.023	**.870**	−.045	−.026
GAM15	.011	−.243	**.474**	.140	**.361**	−.078	−.129	**.409**	.179	**.392**
GAM16	−.040	−.067	.065	**.914**	.011	.048	.012	.019	**.990**	−.024
GAM17	**.592**	−.208	−.011	.018	.006	**.483**	.027	−.135	−.008	.049
GAM19	.118	.059	−.024	.068	**.771**	−.022	.093	.003	.008	**.825**
GAM20	**.459**	**.300**	.071	−.003	.041	**.341**	**.353**	.149	.022	−.025

Note: Factor loadings in bold print are estimated freely in the subsequent confirmatory multi-group covariance and means structure analyses.

mean of the Dedication factor was fixed to 0 in the women, resulting in a significant deterioration of the model fit (Model 6 vs Model 5: $\chi^2_{\text{diff}}(1) = 9.52$, $p < .01$). The factor mean of the women was estimated at .33 (SD = 1.05), implying that, on average, women remembered themselves to be more dedicated to their academic work than men.

The biased item (item 4) only loaded on the Dedication factor. If we would have calculated simple sum scores across the items of the Dedication factor (rather than subjecting the items to a factor model), and compared men and women with respect to these sum scores, as is common practice, then the presence of the biased item would have lead to an overestimation of the effect size of the sex difference in sum scores of .06 (effect size is .33 with, and .27 without the biased item).

4.3.2. General achievement motivation

The results and fit statistics of the multi-group CFA of the GAM are presented in Table 9.

In Model 1, we tested for configural invariance, with 4 items loading on Pressure, 6 on Accomplishment, 4 on Work Approach, 3 on Future Orientation, and 5 on Competition, and four cross-loadings (i.e., items 4, 10, 15 and 20 all loaded on two factors : non-congeneric structure). All residual terms were modelled as uncorrelated. In addition, all 18 items were regressed on age in men and women separately, in order to correct for possible age effects. The RMSEA (.059) indicated that Model 1 described the data adequately, while the CFI was rather low (.90).

It should be noted that because of the way the CFI is calculated (i.e., as the difference in fit between the

Table 7
Results of the multi-group covariance and means structure analyses (MG-CMSA) for the academic achievement motivation (AAM) subscale.

		CFI	RMSEA	Vs model	DF	χ^2_{diff}	p
Model 1	Configural invariance	.98	.058				
Model 1a	Age correction equal across sex	.97	.068	Model 1	6	30.19	<.001
Model 2	Metric invariance	.98	.055	Model 1	7	11.34	ns
Model 3	Strong factorial invariance	.97	.056	Model 2	9	23.61	.005
Model 3a	Strong factorial invariance, bar item 4	.98	.051	Model 2	8	9.23	ns
Model 4	Strict factorial invariance	.98	.047	Model 3a	8	9.14	ns
Model 5a	Mean Persistence factor equal across sex	.98	.048	Model 4	1	2.36	ns
Model 6	Mean Dedication factor equal across sex	.97	.061	Model 5	1	9.53	<.01

Table 8
Correlations between the two latent academic achievement motivation factors persistence and dedication for men (below diagonal) and women (above diagonal), and the means and SD for men and women on these factors.

Correlations

	Persistence	Dedication
Persistence	1	.90
Dedication	.79	1

(Men below diagonal, women above diagonal)

Means (SD)

	Men (N = 338)	Women (N = 497)	Effect size
Persistence	0 (1)	.14 (1.01)	.14
Dedication	0 (1)	.33 (1.05)	.32

Note: The means of the women should be interpreted as deviations from the mean of the men. The mean for the factor Dedication was significantly higher in women (see Model 6, Table 7).

Table 9
Results of the multi-group covariance and means structure analyses (MG-CMSA) for the general achievement motivation (GAM) subscale.

		CFI	RMSEA	Vs model	*DF*	χ^2_{diff}	*p*
Model 1	Configural invariance	.90	.059				
Model 1a	Age correction equal across sex	.90	.058	Model 1	13	28.44	<.01
Model 2	Metric invariance	.91	.057	Model 1	14	21.63	ns
Model 3	Strong factorial invariance	.90	.059	Model 2	18	67.61	<.001
Model 3a	Strong factorial invariance, bar items 4, 5, 9, and 12	.91	.056	Model 2	15	25.61	.04
Model 4	Strict factorial invariance	.90	.057	Model 3a	15	50.08	<.001
Model 4a	Strict factorial invariance bar residuals items 6 and 12	.91	.055	Model 3a	15	23.83	ns
Model 5	Factorial means equal across sex	.91	.054	Model 4a	4	7.86	ns
Model 5a	Factorial means equal across sex bar for factors Future Orientation and Competition	.92	.053	Model 4a	3	.92	ns

independence model and the hypothesized model), this fit index can never take on high values if the intercorrelations between the modelled items are small to begin with. In that case, the fit of the independence model will not be very bad, and the difference with the hypothesized model can therefore not become large. Experience thus teaches that the CFI is never high when the intercorrelations between the modelled items are low overall, and in the present data, many intercorrelations were smaller than .20. As the RMSEA indicated adequate fit, and the residual terms (i.e., the part of the data not predicted by the model) were all small, Model 1 was accepted as baseline model for further testing for measurement invariance across sex.

In Model 1a, we tested whether the age effects could be constrained equal across sex, but as this was not the case (Model 1a vs Model 1: $\chi^2_{\text{diff}}(13) = 28.44$, $p < .01$), we chose to leave this part of the model saturated. That is, all age regressions are estimated separately in men and women in all following models, and this part of the model did therefore not contribute to any model misfit.

Metric invariance was tested in Model 2 by fixing all factor loadings to be equal across sex. Factorial variances were estimated freely in the women, but remained fixed to 1 in the male reference group for identification. This set of constraints proved tenable (Model 2 vs Model 1: $\chi^2_{\text{diff}}(14) = 21.63$, ns), implying that metric invariance across sex was tenable for the GAM subscale.

To test for strong factorial invariance, all thresholds were constrained to be equal across sex in Model 3. This set of constraints, however, resulted in a significant deterioration of the model fit (Model 3 vs Model 2: $\chi^2_{\text{diff}}(18) = 67.61$, $p < .001$). By systematically testing for strong factorial invariance for each of the 5 factors separately, it appeared that 4 of the 18 items (items 4, 5, 9, and 12) were biased with respect to sex. Note that in the item-specific analyses (Table 3), large effect sizes were observed for items GAM4 and GAM9. The effect sizes for items GAM5 and GAM12 were smaller, but the factor loadings for these items were not that large (although significant), meaning that these items were not strongly related to the latent factors. Constraining all thresholds, bar the thresholds of these 4 items, to be equal across sex, did just result in a significant drop of the model fit (Model 3a vs Model 2: $\chi^2_{\text{diff}}(15) = 25.61$, $p = .04$), but as the overall fit of Model 3a was satisfactory, we accepted this model. For the GAM, strong factorial invariance was thus established for 14 out of 18 items, while sex differences on 4 items were too large to be accounted for by the model. In the subsequent models, the thresholds of these 4 items were therefore estimated freely in both groups, and as such no longer contributed to the means of the underlying latent factors. The means of these factors were thus directly comparable between men and women.

In Model 4 we tested for strict factorial invariance, by constraining all residual variances in the women equal to those of the male reference group, i.e., equal to 1. These constraints were however not tenable (Model 4 vs Model 3a: $\chi^2_{\text{diff}}(15) = 50.08$, $p < .001$). In Model 4a, all residuals except the residual for item 12, were fixed to be equal in men and women, and this set of constraints was tenable (Model 4a vs Model 3a: $\chi^2_{\text{diff}}(15) = 23.83$, ns). This implies that the reliability of item 12 was not equal across sex: the residual variance was smaller in women, implying that the reliability of this item was higher in women. Table 10 shows the factor correlations and factor means taken from Model 4a for men and women separately.

Given that the greater part of the factor model was invariant across sex, while the parts that were not were freely estimated and thus no longer contributed to any sex differences, we could now meaningfully compare the five factor means across sex. In Model 4a, all factor means in the women's group were freely estimated while the factor means in the male reference groups were fixed to 0 for reasons of identification. In Model 5, we fixed all factor means to 0 in the women. This did not result in a significant drop in model fit (Model 5 vs Model 4a: $\chi^2_{\text{diff}}(4) = 7.86$, $p = .10$). However, this omnibus test disguised what was already apparent from the effect sizes shown in Table 10, namely the fact that the mean of the factor Future Orientation was actually significantly different between men and women (Future Orientation: $\chi^2_{\text{diff}}(1) = 5.67$, $p < .05$) while a trend was observed for the factor Competition ($\chi^2_{\text{diff}}(1) = 3.22$, $p = .07$). In the final model, Model 5a, we freely estimated these means in the women, and fixed the means of the other three factors to zero (Model 5a vs Model 4a: $\chi^2_{\text{diff}}(3) = .92$, ns). The means of the factors Future Orientation and Competition were negative in women, implying that women were somewhat less future-oriented, and less driven by motives related to competition with others.

The four biased items (4, 5, 9, and 12) affected all factors of the GAM. If we would have calculated simple sum scores across the items of each of the five factors (rather than subjecting the items to a factor model), and compared men and women with respect to these sum scores, as is common practice, then the presence of the biased items would have lead to considerable

Table 10
Correlations between the five latent general achievement motivation factors pressure, accomplishment, work approach, future orientation and competition for men (below diagonal) and women (above diagonal), and the means and SD for men and women on these factors.

Correlations

	Pressure	Accomplishment	Work Approach	Future Orientation	Competition
Pressure	1	.50	.13	.19	.29
Accomplishment	.44	1	.26	.41	.60
Work Approach	.25	.33	1	.14	.13
Future Orientation	.23	.44	.43	1	.50
Competition	.20	.50	.16	.57	1
(Men below diagonal, women above diagonal)					

Means (SD)

	Men ($N=338$)	Women ($N=497$)	Effect size
Pressure	0 (1)	.08 (.84)	.09
Accomplishment	0 (1)	.02 (1.06)	.02
Work approach	0 (1)	.05 (1.00)	.05
Future orientation	0 (1)	-.22 (.85)	-.24
Competition	0 (1)	-.18 (.70)	-.22

Note: The means of the women should be interpreted as deviations from the mean of the men. The mean for the factor Future Orientation is significantly lower in women, and a trend towards significance was observed for the factor Competition.

over- or underestimation of the effect sizes of the sex difference for the factors Pressure (effect size with the biased item: .05, effect size without the biased item: −.13, difference: −.18), Accomplishment (with: .03, without: .00, difference: −.03), Work Approach (with: .08, without: −.01, difference: −.09), Future Orientation (with: .29, without:.14, difference: −.15), and Competition (with: .02, without: .08, difference: .06).

Finally, in order to estimate the correlations between the academic and the general achievement motivation factors, we combined the final models from the AAM (Model 5) and the GAM (Model 5a) into one overall model. The fit of this model, which we denoted the Total Model, was adequate (CFI = .92, RMSEA = .053), and the correlations between the 7 subscales are shown in Table 11. All correlations between the AAM factors and the GAM factors were positive and most of them were significant. Note that the modification indices of the Total Model were all small, as were the residuals (i.e., the part not explained by the model), which implies that the choice to analyze the AAM and the GAM subscales separately before combining them, was justified (i.e., there were no cross-loading between the AAM and the GAM factors, and no correlated errors, etc).

Table 11
Correlations between the two latent academic achievement motivation factors and the five general achievement motivation factors for men and women separately.

	Men		Women	
	Persistence	Dedication	Persistence	Dedication
Pressure	.06	.11	.07	.25**
Accomplishment	.39**	.33**	.46**	.58**
Work approach	.37**	.20**	.42**	.32**
Future orientation	.18*	.13†	.13†	.14*
Competition	.18*	.32**	.15*	.36**

Note: Signs denote the significance of the observed correlations: ** p<.01, * p<.05, † p<.08

5. Discussion

In this study, sex differences in academic achievement motivation and general achievement motivation were examined in adult subjects using categorical multi-group covariance and mean structure analysis (MG-CMSA).

Academic achievement motivation was measured with 10 items. A two-factor solution, with factors Dedication and Persistence, described the relations between these items adequately. On the level of the latent factors, men and women differed significantly with respect to the mean of the factor Dedication, with women considering themselves to have been more dedicated to their academic work than men. No mean difference was observed for the factor Persistence, i.e., men and women did not differ in their retrospective evaluation of how well they had been able in their school years to allocate time to, and focus on, homework. The questionnaire, of which the DAMT was part, also included two retrospective questions on whether the parents of the participants had considered school important, and whether the participants' school results were discussed at home. The men and women in this study did not respond differently to these questions ($Z = -1.74$, *ns*, and $Z = -.71$, *ns*, respectively). This suggests that the differences in Dedication observed between men and women in this study were most likely not due to a difference in how they experienced their academic upbringing. One academic achievement motivation item proved biased with respect to sex, i.e., the sex difference on this item was too large to be explained by the underlying latent factor Dedication, and this item-specific sex difference was not indicative of a sex difference in Dedication. On this item, which concerned the question of whether others had thought the participant to be diligent in school, women scored much higher than men, i.e., women thought they were perceived as more diligent by others.

General achievement motivation was measured with 18 items. A five-factor solution, with factors Pressure, Accomplishment, Work approach, Future Orientation, and Competition, described the relations between those items adequately.

On the level of the latent factors, sex differences were observed for the factors Future Orientation and a trend was observed for the factor Competition. On both factors, women scored lower than men. This means that women cogitated less about the future and made less future-related plans, compared to men, and achieving more than others was considered less important by women compared to men. Men and women did not differ with respect to the means of the factors Pressure, Accomplishment, and Work approach, i.e., men and women did not differ in their perception of how occupied they are by their work, in their assessment of the demands they put on themselves, and in their evaluation of how important work/employment is in their lives. Four of the 18 general achievement motivation items were biased with respect to sex, i.e., the sex difference observed on the items were not indicative of the sex difference on the underlying factors. On three of the four items, women scored lower than men: women aspired less after a busy management job at a factory (item 4), were less often of the opinion that other people could work harder (item 5), and perceived working on something for a long time as more tiring (item 9). At the same time, women were more concerned about other peoples' opinion about their achievements than men (item 12). It should be noted that in the exploratory factor analysis, the pattern of factor loadings for some of the biased items differed between men and women. In MG-CMSA, item bias is defined within the context of a specific factor model, i.e., an item is considered biased if the mean differences observed for this item cannot be explained by the specified model. This means that in theory, bias can originate from model misspecification in one of the groups. To verify whether this was the case, we ran alternative models in which the biased items were allowed to load on other factors as well. The bias however remained significant, implying that it was not the result of misspecifications in the factor structure.

In our analyses, we chose to leave items for which the bias was uniform (i.e., limited to the intercepts) in the model. This strategy is justified as uniformly biased items no longer contribute to the misfit of the model when their intercepts are freely estimated and thus allowed to vary across the groups. This strategy is, however, not recommended when the bias is *non*-uniform (implying significant differences in factor loadings between groups). In that case, one should remove the item from the model before testing for strong and strict measurement invariance.

Individual items which clearly showed differences in endorsement rates between the sexes were indeed flagged as biased in the MG-CMSA analyses. However, although the sample sizes in our study were considerable ($N=338$ and $N=497$, respectively), the statistical power to detect mean differences between groups on the level of the latent factors was not optimal. For example, even though the sex difference in the factor Competition was associated with an effect size of −.22, the effect was only marginally significant ($p=.07$).

These results show that items that measure motivation-related concepts can be biased with respect to sex. It is possible that the sex difference on these items was just too large to be accounted for by underlying latent factors (i.e., the sex difference is item specific), but it is also possible that the connotation of these items was different for men and women to such an extent that the responses of men and women on the biased items were actually incomparable. MG-CMSA can only point out the location of the bias, but further research into the content and interpretation of these items would be required to uncover the exact nature of the bias. The present results are of course limited to these specific academic and general achievement motivation subscales of the DAMT. Yet, the study shows that researchers should be cautious in directly comparing motivation-related scores of men and women without first studying their comparability. Our calculations based on sum scores showed that the effect sizes of the sex differences in achievement motivation can be very much affected by the presence of a few biased items. As a result, sex differences in achievement motivation may be exaggerated or underestimated in one study, and may fail to replicate in subsequent studies, in which different instruments are used to measure achievement motivation.

One advantage of studying academic achievement motivation and general achievement motivation simultaneously is that one can calculate the correlation between these two types of motivation. In this study, the two academic achievement motivation factors and the five general achievement motivation factors correlated positively. Correlations were particularly strong between the 2 academic achievement motivation factors Persistence and Dedication on the one hand, and the 3 general achievement motivation factors Accomplishment, Work Approach and Competition on the other.

One disadvantage of studying academic achievement motivation in an adult population is that such a study is by definition retrospective. For some of the participants in our sample, which was particularly heterogeneous with respect to age, the schooldays were a distant past. Retrospective assessments of one's own academic achievement motivational levels may not always be reliable. The fact that we find clear factor structures, and significant correlations between academic and general measures of achievement motivation, suggests that the answers must at least have been consistent within subjects, but how reliably such retrospective assessments reflect the past reality, cannot be answered with the present data.

In this study, age effects were accounted for by partialling out the effects of age on the level of the items before fitting the factor models. Of the 28 DAMT items, 12 showed significant age effects in men, and 15 showed significant age effects in women. Moreover, especially for the academic achievement motivation items, age effects were markedly different for men and women, not only in size (e.g., AAM items 1, 7 and 9) but even in sign (e.g., AAM for items 2 and 5). The aim of this paper was not to study the effect of age on motivation. We therefore decided to keep the age-corrective part of the model saturated, which implies that the effects of age were fully controlled for in men and women separately, and not studied in more detail. Yet, the finding that age did affect the responses, and that it did so in a different manner for men and women, does suggest that inconsistencies between former studies in whether or not sex differences in motivation were observed, and whether motivation was related to actual achievement, could be due to differences between these studies in the age range of their study-samples.

The heavy reliance on self-report measures in research on achievement motivation forms another potential source of bias since subjective and objective evaluations of a person's motivation, effort and dedication, may not always be in agreement, especially retrospectively. Especially with respect to the Academic Achievement Motivation scale, we emphasize that we measured our participants' personal recollection of how dedicated and persistent they were when they were in school. It is possible that sex bias, such as observed for the diligence item of the AAM scale, represented a difference between women and men in how they perceived and recalled reactions from their environment, rather than objective differences in diligence with respect to academic work. Besides the problems related to self-report measures, the lack of standardized and validated measures of academic and general achievement motivation hampers the generalizability of results across studies. Still, the present study shows that MG-CMSA is useful in locating the exact source of sex differences in motivation, and that the study of item bias may be advantageous in the field of sex differences in achievement motivation.

The present sample consisted of family data, i.e., twins and their (in-law) family members, and one question of interest is whether twin-samples can be considered representative of the general population. In general, twins are born in all strata of society, and they are on average somewhat more willing to participate in research, as are their relatives (Martin, Boomsma, & Machin, 1997). To date, no studies have been performed on whether twins differ from non-twins in motivation-related characteristics. At present, there is however no reason to believe that motivational differences between the sexes should be different for men and women born as twins or coming from twin families. Another question of interest is whether the sex differences observed in the present sample are representative of sex differences in the Dutch population. Generally, men and women differ in their willingness to participate in research (women being somewhat more willing). If this sex difference in willingness to participate in research is in turn related to, or dependent on, social status or success (i.e., men are more willing to participate if they are socially more successful, while women's willingness to participate is independent of their social status) then the sex differences observed in achievement motivation could be a function of the sex difference in the willingness to participate. In principle, this could be tested by comparing the within-pair differences in motivation observed between opposite sex twins or opposite sex siblings,[4] to the sex differences observed in unrelated individuals. That is, if the sex differences in motivation such as observed *across* families are also observed *within* families (where brothers and sisters are matched with respect to social background and social economic status), then the possible distortion (due to sex differences in willingness to participate, or due to studying twins rather than non-twins) is probably minor. Our family data included 47 complete opposite sex twin pairs and opposite sex sibling pairs. Wilcoxon non-parametric signed-rank tests showed that even within this limited number of opposite sex pairs, brothers and sisters scored markedly differently on items AAM4, GAM4 and GAM12, which were all labelled as severely biased in this study. The fact that the sex effects as reported in the total sample were also observed within families, confirms our expectation that selection effects were absent or minor. Yet, the question of whether the development of motivation-related traits is influenced by the presence of a co-twin or sibling, does merit further study. Such studies could possibly even provide insight into the origin of sex differences in motivation. Similarly, the question of whether the willingness to participate in research is itself related to achievement motivation is worth following-up.

This study was the first to examine measurement invariance and sex differences in the context of motivational research using MG-CMSA in a large sample of adult participants. It was shown that five motivation items (1 academic, 4 general achievement motivation) were biased with respect to sex. Once these biased items were effectively removed from the means model, sex difference were still observed for Dedication (academic achievement motivation), and for Future Orientation and Competition (general achievement motivation). Further studies into the nature of the sex bias observed for some items are merited. In addition, it would be interesting to study how the sex differences observed in subjects' self-reported motivation, relate to more objective measures. For example, in the present study, sex differences were absent with respect to the factor Pressure, implying that men and women do not differ in their perception of the extent to which they are engaged by their work. It would be interesting to relate this subjective perception to an objective measure such as the number of hours of work per week. Such studies could be used for validation, but could also enhance our insight into the relation between achievement motivation-related constructs on the one hand, and actual achievement on the other.

Acknowledgement

Sophie van der Sluis (VENI-451-08-025) and Danielle Posthuma (VIDI-016-065-318) are financially supported by the Netherlands Scientific Organization (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, gebied Maatschappij-en Gedragswetenschappen: NWO/MaGW).

We thank all our test assistants for their help with the data collection.

Appendix A. DAMT example items

Academic Achievement Motivation (AAM)

Dedication:

AAM7: “When I was in school, the demands that I made on myself concerning studying were very high, high, pretty high, low” (R)

Persistence:

AAM8: “When I'm studying, my thoughts often wander / I'm not easily distracted / I work ceaselessly”

General Achievement Motivation (GAM)

Pressure:

GAM2: “Usually, I'm busy / quite busy / not very busy / not busy at all” (R)

[4] Note that in principle, father–daughter and mother–son relations could also be included in the within-family comparisons, in addition to brother–sister relations. The advantage of brother–sister comparisons is, however, that these relatives are of approximately the same age, grew up at approximately the same juncture, and were nursed under approximately the same (social economic) circumstances, i.e., these relatives are matched with respect to background variables, while such matching is not as obvious across generations.

Accomplishment (intrinsic motivation):

GAM10: "The urge to surpass myself is very strong / pretty strong / not very strong" (R)

Work approach:

GAM14: "For me, working is something which I would like to do only occasionally / which I like to do, but which generally takes me a lot of effort / which I always enjoy doing"

Future Orientation:

GAM7: "When thinking about my future, I usually plan very far ahead / plan far ahead / I plan ahead quite a bit / I do not usually plan ahead very far" (R)

Competition:

GAM19: "Achieving more than others, is very important for me / is important for me / is quite important for me / is not that important for me" (R)

References

Bentler, P. M. (1990). Comparative fit indixes in structural models. *Psychological Bulletin*, *107*, 238—246.

Bollen, K., & Long, J. S. (Eds.). (1993). *Testing structural equation models*. Newbury Park, CA: Sage.

Boomsma, D. I., de Geus, E. J. C., Vink, J. M., Stubbe, J. H., Distel, M. A., Hottenga, J. H., et al. (2006). Netherlands twin register: From twins to twin families. *Twin Research and Human Genetics*, *9*(6), 849—857.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136—162). Newbury Park, CA: Sage.

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*, 456—466.

CITO Terugblik en resultaten 2009 - Eindtoets Basisonderwijs, versie februari 27, 2009.

de Knecht-van Eekelen, A., Gille, E., & van Rijn, P. (2007). Resultaten PISA 2006. Praktische kennis en vaardigheden van 15-jarigen. Nederlandse uitkomsten van het OESO Programme for International Student Assessment (PISA) op het gebied van natuurwetenschappen, leesvaardigheid en wiskunde in het jaar 2006.

Dolan, C. V., Colom, R., Abad, F. J., Wicherts, J., Hessen, D. J., & van der Sluis, S. (2006). Multi-group covariance and mean structure modeling of the relationship between WAIS-III common factors and gender and educational attainment in Spain. *Intelligence*, *34*(2), 193—210.

Dolan, C. V., & Hamaker, E. L. (2001). In F. Columbus (Ed.), *Investigating balck-white differences in psychometric IQ: Multi-group confirmatory factor analyes of the WISC-R and K-ABC and a critique of the method of correlated vectorsAdvances of Psychological Research*, *vol. 6*. (pp. 31—59) Huntington: Nova Science Publishers.

Epstein, D., Elwood, J., Hey, V., & Maw, J. (Eds.). (1998). *Failing boys? Issues in gender and achievement*. Buckingham: Open University Press.

Freudenthaler, H. H., Spinath, B., & Neubauer, A. C. (2008). Predicting school achievement in boys and girls. *European Journal of Personality*, *22*, 231—245.

Hermans, H. J. M. (2004). *Prestatie Motivatie Test, Handleiding.* Lisse: Harcourt Assessment B.V.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117—144.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fitindexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modelling*, *6*(1), 1—55.

Gagné, F., & St Père, F. (2001). When IQ is controlled for, does motivation still predict achievement? *Intelligence*, *30*, 71—100.

Grant, L., & Rong, X. L. (1999). Gender, immigrant generation, ethnicity, and schooling progress of youth. *Journal of Research and Development in Education*, *33*, 15—26.

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 294—316). Newbury Park, CA: Sage.

Little, T. D. (1997). Mean and covariance structures (MACS) analysis of cross-cultural data: practical and theoretical issues. *Multivariate Behavioral Research*, *32*, 53—76.

Martin, N., Boomsma, D., & Machin, G. (1997). A twin-pronged attack on complex traits. *Nature Genetics*, *17*(4), 387—392.

Meece, J. L., Glienke, B. B., & Burg, S. (2006). Gender and motivation. *Journal of School Psychology*, *44*, 351—373.

Mellenbergh, G. J. (1989). Item Bias and Item Response Theory. *International Journal of Educational Research*, *13*, 127—143.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*(4), 525—543.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479—515.

Muthén, L. K., & Muthén, B. O. (1998-2007). *Mplus user's guide*, Fourth Edition Los Angeles, CA: Muthén & Muthén.

Rosenthal, R. (1991). *Meta-analytic procedures for social research (revised)*. Newbury Park, CA: Sage.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysisA Festschrift for Heinz Neudecker*. (pp. 233—247) London: Kluwer Academic Publishers.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research*, *8*(2), 23—74.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229—239.

Spinath, B., Spinath, F. M., Harlaar, N., & Plomin, R. (2006). Predicting school achievement from general cognitive ability, self-perceived ability, and intrinsic value. *Intelligence*, *34*, 363—374.

Steinmayr, R., & Spinath, B. (2008). Sex differences in school achievement: What are the roles of personality and achievement motivation? *European Journal of Personality*, *22*, 185—209.

Steinmayr, R., & Spinath, B. (2009). The importance of motivation as a predictor of school achievement. *Learning and Individual Differences*, *19*, 80—90.

Van der Sluis, S., Posthuma, D., Dolan, C. V., Boomsma, D. I., de Geus, E. J. C., Colom, R., et al. (2006). Gender differences on the Dutch WAIS-IIIR. *Intelligence*, *34*(3), 273—289.

Van der Sluis, S., Derom, C., Thiery, E., Bartles, M., Polderman, T. J. C., Verhulst, F. C., et al. (2008). Sex differences on the WISC-R in Belgium and the Netherlands. *Intelligence*, *36*, 48—67.

Houtte, Van (2004). Why boys achieve less at school than girls: The difference between boys' and girls' academic culture. *Educational Studies*, *30*(2), 159—173.

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., van Baal, G. C. M., Boomsma, D. I., et al. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence*, *32*(5), 509—537.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant & M. Windle (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281—324). Washington, DC: American Psychological Association.

Wong, K. -C., Lam, Y. R., & Ho, L. -M. (2002). The effect of schooling on gender differences. *British Educational Research Journal*, *28*, 827—843.