A Differential Item Functioning Analysis of the German Academic Self-Regulation Questionnaire for Adolescents

Timo Gnambs

Osnabrück University

Barbara Hanfstingl

University of Klagenfurt

Author Note

Timo Gnambs, Institute of Psychology, Osnabrück University, Germany; Barbara Hanfstingl, Institute of Instructional and School Development, University of Klagenfurt, Austria.

Correspondence concerning this article should be addressed to Timo Gnambs, Institute of Psychology, Osnabrück University, Seminarstr. 20, 49069 Osnabrück, Germany, E-mail: timo.gnambs@uni-osnabrueck.de

Accepted for publication in the *European Journal of Psychological Assessment*

The definitive version of the article will be available at psycontent.metapress.com

Abstract

The German Academic Self-Regulation Questionnaire (SRQ-A[G]) for adolescents assesses four regulatory styles within Deci and Ryan's (1985) self-determination theory: intrinsic, identified, introjected, and external regulation. The study on N = 2,123 students (1,057 girls) from secondary schools in Austria analyzes the effects of differential item functioning (DIF) on individual and group-level estimates of the latent regulatory styles. The scale demonstrated small DIF for sex and the ages from 10 to 17. The DIF items favored, if anything, younger students and lead to a slight overestimation of their introjected motivation level. However, the practical impact on group-level means was negligible. The SRQ-A[G] represents a reliable instrument to capture sex- and age-related differences in the four regulatory styles throughout adulthood.

Keywords: differential item functioning, motivation, self-determination theory, response style

A Differential Item Functioning Analysis of the German Academic Self-Regulation Questionnaire for Adolescents

The formation and enduring maintenance of students' motivation is a central task in educational settings, as perceived self-determined motivations have strong effects on a variety of favorable academic outcomes, such as positive affect (Harter, Whitsell, & Kowalski, 1992), academic engagement (Otis, Grouzet, & Pelletier, 2005), deep conceptual learning strategies (Rijavec, Saric, & Miljokvich, 2003), and even academic achievement (Lepper, Corpus, Iyengar, 2005). Research on students' motivations requires instruments with known psychometric properties; the interpretation of group differences on a given scale, in particular, requires measurement equivalence across groups. We present a differential item functioning (DIF) analysis of the German Academic Self-Regulation Questionnaire for adolescents (SRQ-A[G]; Müller, Hanfstingl, & Andreitz, 2007) and assess the impact of DIF on individual and group-level estimates of the latent constructs.

Self-Determination Theory

In general, motivation is divided into two components: intrinsic (i.e., inherent to the task itself) and extrinsic (i.e., originating outside the task). However, a two-dimensional view of motivation disguises an important micro-structure of extrinsic motivation that has differential outcomes in the academic context. For example, a student's learning activities might be a consequence of external pressure from parents, whereas another student might learn to obtain a high-school diploma for future university studies. Both students are extrinsically motivated because it is not the task itself that initiates their learning, but extrinsic demands outside the task. However, both demonstrate different *qualities* of extrinsic motivation that stem from varying degrees of self-determination. Deci and Ryan (1985) describe different regulatory styles of extrinsic motivation that can be distinguished on the basis of their degree of self-determination. *External regulation* conforms to the classical definition of extrinsic motivation, including the lowest level of perceived self-determination and the highest degree of control. It depends on consequences administered by others in the

form of rewards or punishments. For example, a student might engage in learning to receive praise for good grades or, vice versa, avoid criticism. *Introjected regulation* describes behaviors related to self-esteem. Individuals do not engage in tasks because they truly consider them important but, rather, because it is expected from them and conforms to established social norms (e.g., students might learn to avoid feeling ashamed). *Identified regulation* focuses on the individual importance of learning and is often goal-driven; thus, the governing motive for the behavior lies in its specific outcome. For example, students might engage in a learning activity with the prospect of obtaining a high-school degree. These three regulatory styles represent different qualities of extrinsic motivation that vary in their degree of self-determination. The strongest form of self-determination represents *intrinsic regulation*, which describes students' behavior that is primarily motivated by the task itself. Students learn because they are truly interested in a topic, have fun dealing with it, and are eager to broaden their horizons. Self-determined motivation styles are particularly relevant in the academic realm due to their positive effects on, for example, academic engagement and achievement (e.g., Lepper et al., 2005; Otis, et al., 2005). Regulatory styles frequently display marked sex differences and are, on average, higher for girls than for boys (Marsh, Martin, & Cheng, 2008). Moreover, they do not seem to be stable with increasing age: intrinsic motivations gradually decline throughout adulthood (Corpus, McClintic-Gilbert, & Hayenga, 2009; Lepper, et al., 2005). To allow for meaningful interpretations of these group differences, it has to be established that the administered measurement instrument functions comparably in all groups, that is, measurement equivalence must hold; otherwise the observed score differences might be attributed to differences in the measurement model rather than to true differences in regulatory styles.

Measurement Equivalence

The comparison of groups or individuals on the basis of an observed score requires that the test score is a comparable indicator of the latent construct of motivation. If systematic differences exist in the measurement properties of a scale, scores obtained from such

instruments lead to distorted inferences about an individual's standing on the latent trait. Measurement equivalence of an instrument holds when individuals with the same position on the latent trait have the same response probability at the item and subscale level (Drasgow & Kanfer, 1985). Within the framework of the linear measurement model, the prevalent approach for the assessment of measurement invariance is multi-group confirmatory factor analysis. However, simulation studies indicate that this approach might be inadequate for categorical indicators resulting from response scales in a rating format (Kankaras, Vermunt, Moors, 2011). In these cases, item response theory (IRT), which explicitly acknowledges the ordinal nature of the data, seems more appropriate. In the context of IRT, many different methods have been proposed to test for measurement equivalence (i.e., differential item functioning) across subgroups of respondents (e.g., boys and girls) or measurement occasions, such as the likelihood ratio test (Thissen, Steinberg, & Wainer, 1993), the DFIT approach (Flowers, Oshima, & Raju, 1999), linear logistic test models (Gnambs & Batinic, 2011) or the logistic regression test (Swaminathan, & Rogers, 1990; Zumbo, 1999). All these techniques have in common that they test for the equality of expected true scores across two or more groups when the latent trait is held constant.

This paper focuses on the regression framework because it is rather flexible and can easily be applied to the test for measurement equivalence across multiple groups and even continuous variables (e.g., age). Furthermore, simulation studies have demonstrated comparable or even superior power of regression analyses in the detection of DIF as compared to other procedures (cf. Clauser, Nungster, Mazor, & Ripkey, 1996; Swaminathan & Rogers, 1990). The regression framework for DIF analysis involves the comparison of three logistic (in case of polytomous items: ordinal) regressions (Choi, Gibbons, & Crane, 2011): First, the cumulative response probabilities for each item are regressed on the observed trait score without considering any covariates. Instead of using the observed trait score, it has also been suggested (Crane, Gibbons, Jolley, & van Belle, 2006) to use the IRT-derived latent trait score in this regression because—unless the Rasch model holds for an item set (i.e. equal

discrimination parameters)—the observed score is a rather biased matching criterion (Millsap & Everson, 1993). In this case, the respective regression is nearly equivalent to a conventional IRT formulation and represents the hypothesis that the item responses are solely dependent on the trait. In the second step, a covariate representing group membership (e.g., male vs. female) is added to the model. If the latter explains significantly more variance than the first regression, the item exhibits uniform DIF; the item score does not only depend on the latent trait, but also on the group membership. To test if the degree of DIF varies depending on the latent trait, a third regression is estimated that also includes an interaction term between group membership and the latent trait. If the latter explains significantly more variance than the previous regression, the item exhibits non-uniform DIF. Model comparisons are typically based on the loglikelihood difference test (Swaminathan, & Rogers, 1990). As this test is highly sensitive to sample size and, given that a large enough sample size identifies even negligible differences, it has been recommended to base the identification of DIF on an effect size measure (Meade, Johnson, & Braddy, 2008). DIF is considered negligible when the difference in R^2 is < .035 (Jodoin, & Gierl, 2001) or the difference in regression weights is less than 5% or even 1% (Crane, Gibbons, Ocepek-Welikson, Cook, & Cella., 2007).

Overview

The study focuses on the German Academic Self-Regulation Questionnaire (Müller et al., 2007) designed to assess four motivational regulatory styles in adolescents: intrinsic, identified, introjected, and external regulation. Several authors have noted a marked decline in self-determined motivations across adolescents (Corpus et al., 2009; Lepper et al., 2005) and also higher levels of motivation for girls (Marsh, et al., 2008). However, these analyses assumed measurement invariance of their instruments without actually evaluating it. Therefore, the present study examines DIF and its consequences on individual and group-level estimates of the four regulatory styles for two focal characteristics in educational research: sex and age.

Method

Participants and Procedure

Responses to the SRQ-A[G] were obtained from 2,138 adolescents (1,066 girls) from 112 secondary schools across rural and urban localities in Austria. To reach a diverse sample of students, all major school types were included: about 28% attended higher general secondary schools, 66% went to secondary schools providing vocational education, and the remaining 5% encompassed students from several specialized school branches. Students from grades 5 to 12 were eligible to participate in the study. Their ages ranged between 10 and 17 (M = 13.54, SD = 2.02). A total of 15 participants had to be excluded from the analysis due to a high number of missing data (more than 3 items). The proportion of missing values in the remaining sample (N = 2,123) ranged between zero and three percent for each item, which falls well below the tolerable threshold of five percent (Little & Rubin, 1987). Moreover, simulation studies indicated unbiased results of DIF analyses within the logistic regression framework for small rates of responses (i.e. 10 percent) missing completely at random (Robitzsch & Rupp, 2009). Data collection was conducted during class in groups of 20 to 30 students by teachers having received prior training in the assessment procedure.

Instrument

The SRQ-A[G] (Müller, et al., 2007) is a modified short version of the Academic Self-Regulation Questionnaire (SRQ-A; Ryan & Connell, 1989) adapted for the German-speaking countries and assesses the four regulatory styles as described above. The 16 items of the SRQ-A[G] (see appendix) were selected using factor loadings from an item pool including the SRQ-A and several newly constructed items. All items were designed for adolescents with sufficient reading comprehension and as such are appropriate for students in secondary schools from age 10 and upwards. Responses to all items were indicated on five-point response scales from “strongly agree” to “strongly disagree”. Previous studies (cf. Müller et al., 2007) identified hypothesized validity correlations for the four scales with the satisfaction of three basic psychological needs that—according to self-determination theory—are essential for the development of intrinsic, self-determined motivations (Deci & Ryan, 1985). Intrinsic

regulation was moderately correlated with the needs for autonomy, competence and social relatedness, whereas the other three scales showed gradually decreasing validity correlations; external regulation was not correlated with need satisfaction. Furthermore, in line with an educational-psychological theory of interest (Krapp, 2002) intrinsic regulation was also moderately correlated with perceived topic relevance and teacher involvement; whereas external regulation was not. Overall, the four scales exhibited hypothesized validity correlations with different constructs in educational research.

In the current sample, the scores of the four subscales resulted in means of $M = 3.36$ ($SD = 1.06$) for intrinsic regulation, $M = 3.77$ ($SD = 1.04$) for identified regulation, $M = 2.96$ ($SD = 1.02$) for introjected regulation, and $M = 2.95$ ($SD = 0.95$) for external regulation. Students from higher general secondary schools exhibited significantly, $p < .05$, lower identified and introjected regulation scores than students attending secondary schools providing vocational education. However, with Cohen's *d*s of -0.17 and -0.11 the respective effects were rather small. Intrinsic and external regulation did not result in significant differences, $d = -0.08$ and -0.03. Variances were comparable across school types, all *p*s > .05. Latent factor reliabilities (Hancock & Mueller, 2001) for the four subscales were satisfactory, with .94, .93, .83, and .79, respectively.

Results

Latent Trait Modeling

Parametric item response models rely on rather strong assumptions, including unidimensionality, local independence, and monotonicity. Prior to DIF analyses, it is important to test if these assumptions are met; otherwise potential DIF effects cannot be distinguished from effects resulting from poor model fit.

Model assumptions. Unidimensional IRT models assume that a respondent's observed responses reflect a single latent proficiency dimension. The implied factorial structure of the 16 items was investigated with exploratory factor analyses (EFA). To account for the ordered response format, the EFA was conducted on a polychoric correlation matrix. A

principal axis factor analysis with promax rotation ($\kappa = 4$) clearly rediscovered the four subscales. All items had high loadings, *Mdn*(λ) = .76 [*Min* = .47, *Max* = .96] on their respective factor and minor cross-loadings, *Mdn*($|\lambda|$) = .05 [*Min* = .00, *Max* = .31], on the other factors. Generally, unidimensionality is sufficiently supported for IRT parameter calibration when the first factor extracted from an item set accounts for at least 20% of the variances of the items (Reckase, 1979). For each of the four subscales, a single factor explained between 42% (external regulation) and 75% (intrinsic regulation) of the variances of the items, indicating an adequate latent factor. A second assumption pertains to local independence, that is, for respondents with the same latent trait, two items are expected to be statistically independent of each other (i.e., their covariance approaches zero). After extraction of the first factor, residual correlations greater than .10 are indicative of minor dependencies between items, whereas values greater than .20 indicate serious dependencies (Amtmann et al., 2010). The residual correlations for all four subscales were very low, $|\bar{r}_{res}| < .04$ (*Min* = -.08, *Max* = .11), suggesting that the assumption of local independency is tenable for the item set. A third assumption refers to monotonicity of the item response curves, meaning that the probability of endorsing an item should monotonically increase for individuals with higher traits. Monotonicity was assessed by creating discrete proficiency groups based on the respondents' scale scores (cf. Van Schuur, 2011). Then, the proportion of respondents in each group endorsing an item was tabulated. Items with monotonically increasing response functions should exhibit gradually increasing values with each proficiency group. Descriptive analyses indicated small violations of the monotonicity assumption for four items. However, McNemar tests conducted with the *mokken* software (Van der Ark, 2012) did not corroborate these results at an alpha level of 5%, but suggested that the observed lack of monotonicity was due to sampling error. In conclusion, these results indicate that the 16 items exhibit adequate properties for calibration with parametric IRT models.

Parameter calibration. To determine the optimal response model for each scale four different polytomous IRT models were fitted to the data with the *ltm* software (Rizopoulos,

2006): a generalized partial credit model (Muraki, 1992), a graded response model (GRM; Samejima, 1969), and respective models with equal discrimination parameters for all items. On the basis of Schwarz's bayesian information criterion (BIC; Schwarz, 1978) the GRM was deemed the optimal response model for all four scales. As the SRQ-A[G] supposedly includes four correlated subscales, we fitted a multidimensional variant of the GRM to the data that specified a simple structure for all latent traits (i.e. each item loaded on a single factor and had no cross-loadings). By including several latent factors in a single model, multidimensional IRT models also estimate the covariance structure between different latent traits. The respective model was estimated in Mplus 7 (Muthén & Muthén, 1998-2012) using a full-information maximum likelihood algorithm. Generally, the items had satisfactory loadings on their respective factor, with all slopes falling between $\alpha = 0.70$ and 2.25; in the factor-analytic metric, this corresponds to loadings between $\lambda = .57$ and .91. The external regulation scale had slightly lower loadings, $Mdn(\lambda) = .64$, than the intrinsic, $Mdn(\lambda) = .87$, identified, $Mdn(\lambda) = .82$, and introjected regulation subscales, $Mdn(\lambda) = .70$. The thresholds for the items fell within a range of $\delta = [-2.15, 1.89]$; thus, the items are able to differentiate between individuals about two standard deviations below and above the mean. In line with Deci and Ryan's (1985) self-determination theory, the four subscales were modestly correlated. The highest correlations resulted between proximal regulatory styles, on the one hand, between intrinsic and identified regulation, $r = .49$, $p < .001$, and, on the other hand, between introjected and external regulation, $r = .50$, $p < .001$. Distal intrinsic and external regulations were nearly orthogonal, $r = -.11$, $p = .001$.

Differential Item Functioning

Differential item functioning was analyzed within the ordinal regression framework (Swaminathan & Rogers, 1990; Zumbo, 1999) using the iterative approach suggested by Crane et al. (2007) [1]. We fitted three regression models to each item using the IRT-derived

[1] DIF analyses might yield spurious results when the latent trait score used as matching criterion is biased because it was estimated from items having DIF. The iterative DIF procedure (Crane et al., 2007) estimates

latent trait score as matching criterion: a baseline model without a group-specific covariate (model 1), a model with a group-specific main effect to test for uniform DIF (model 2), and a model to test for non-uniform DIF that also included an interaction term between the respondents' latent trait and the group-specific covariate (model 3). Due to the well-known problems of significance tests in large samples, DIF is identified when, in addition to a significant loglikelihood test for model comparison, an effect size exceeds the threshold for at least small DIF, that is, either a difference in McFadden's $R^2 > .035$ (Jodoin & Gierl, 2001) or a percentage change in regression weights greater than 1% (Crane et al., 2007). To account for the multilevel data structure corresponding random effects were also included that acknowledged the grouping of students within classes. DIF was analyzed with regard to sex and age using a modified version of the *lordif* software (Choi et al., 2011). The results of the DIF analyses are summarized in Table 1. Of the 16 items, we identified 2 items with sex-related DIF and 2 items with age-related DIF. In all instances, the effect sizes for DIF were very small. It has to be noted that DIF was classified according to the strictest criterion in the literature, that is, the 1% rule for β change (Crane et al., 2007). It is still a matter of debate whether this magnitude indeed represents non-negligible DIF for practical applications. Previous research (Crane, Gibbons, Jolley, & van Belle, 2006; Crane et al., 2007) also used higher thresholds (e.g., 5% or 10%) to classify DIF. Neither the more liberal thresholds for β change nor the R^2 rule (Jodoin & Gierl, 2001) would have identified DIF for any item.

Effect of response styles. Individuals vary in their tendency to use extreme response options in Likert scales. These response styles reflect individual differences that are not related to the particular item content. Because the adoption of extreme response styles has

---

corrected latent trait scores that account for the DIF of the items: First, latent trait scores are estimated from the original item set. If DIF is identified using these trait scores, the latent trait scores are estimated anew using the DIF-free items that have parameters estimated from the whole sample and DIF items that have different parameters estimates in the comparison groups (e.g., for boys and girls). In the next step, the DIF analyses are repeated with these new trait scores. If different items are identified as having DIF the previous steps are repeated. Otherwise, the procedure is aborted because the DIF results are stable and not biased by a distorted matching criterion.

been associated with sex and age (Austin, Deary, & Eagn, 2006; De Jong, Steenkamp, Fox, & Baumgartner, 2008; Wetzel, Böhnke, Carstensen, Ziegler, & Ostendorf, 2013), differences in response styles might account for the previously identified DIF. Adopting the modeling approach in Wetzel et al. (2013), respondent homogeneity was examined separately for the identified, introjected, and external regulation subscales.[2] For each subscale a mixture graded response model was specified in Mplus 7 (Muthén & Muthén, 1998-2012) to identify individuals adopting either extreme response styles (ERS) or non-extreme response styles (NRS). These analyses clearly identified two qualitatively distinct groups of students; that is, mixture models with two groups (BIC = 20,470 / 24,264 / 23,575) provided superior fits than respective single-group models (BIC = 20,808 / 24,655 / 23,861). Moreover, constraining the loadings across the two groups did not result in a loss of fit (BIC = 20,445 / 24,244 / 23,561). This indicates that the items measured the same latent traits in both groups but differed with regard to the adopted response style. In the NRS group the first and fourth thresholds were rather widely spaced whereas in the ERS group the threshold parameters for each item were clustered together (cf. Wetzel et al., 2013).

Students could be assigned to the NRS and ERS groups with high certainty: the median probability of the most probable group for each student was *Mdn* = .90. About 39% to 57% of students were classified as adopting ERS. Based on the constrained mixture models, each respondent could be described by an ERS score that was calculated as the logarithmized odds ratio of being in the ERS group as opposed to the NRS group. These ERS scores were derived for each subscale and correlated at $\bar{r} = .36$; thus, the adopted response styles showed remarkable consistency across the three scales. An exploratory factor analysis identified a single factor that explained about 37 percent of variance in the three ERS indicators. The respective factor scores correlated with age at $r = -.13$, $p < .001$; thus, younger students

[2] For the intrinsic regulation subscale measurement invariance of loadings across groups did not hold. Therefore, ERS was not examined further for this scale.

exhibited slightly stronger ERS than older students. Sex was not related to ERS, $r = .02$, $p = .29$.

To examine whether the previously identified DIF for the four items resulted from differences in response styles the DIF analyses presented above were repeated using the ERS factor scores as covariates in the regression models. These analyses did not result in markedly different results. Three of the previously identified items were again identified as having DIF. Only the item with the smallest effect size (IO4 in Table 1) failed to exhibit age-related DIF after controlling for ERS. Thus, ERS had no consistent effect on DIF identification.

Individual level impact. In light of the rather small DIF effects, we subsequently analyzed the impact of the identified DIF for the four items on individual trait estimates. For reasons of brevity, we focus the following presentation to DIF in terms of age. As the precision of parameter estimates in IRT is affected by sample size, we created four age groups, 10 to 11 years ($N = 342$), 12 to 13 years ($N = 741$), 14 to 15 years ($N = 582$), and 16 to 17 years ($N = 426$), to achieve the recommended minimum sample size of 250 in each group (Embretson & Reise, 2000). Following the approach in Crane et al. (2006), we first estimated the item parameters in each age group separately. In the second step, the item parameters from each group were transformed to a common metric using the Stocking and Lord (1983) equating procedure. For this transformation the DIF-free items were used as anchors and the DIF items were treated as unique items in each group. As a result, DIF-free items have parameters estimated from the whole sample, whereas items exhibiting DIF have item parameters estimated separately in the different age groups. Estimates of the latent traits are derived by using the common DIF-free item parameters in conjunction with the group-specific parameters of the DIF items. Figure 1 (top left panel) illustrates the test characteristic curves (TCC) for the introjection scale with regard to the four age groups. For visual clarity, the figure includes only the TCC for the youngest (10 to 11 years) and oldest (16 to 17 years) respondents. The TCC for the two remaining age groups fell in between the two depicted curves. For older adolescents (dashed line in the top left panel of Figure 1) the TCC is shifted

markedly to the right on the latent dimension. Ignoring the DIF for this item would result in a slight overestimation of younger students' latent traits or, vice versa, an underestimation of older students' traits. A visual inspection of the difference in TCC (grey line in Figure 1) also indicates that the bias due to DIF was not constant across the latent trait continuum but was slightly larger for students with lower trait levels. Hence, if the DIF for these items were ignored, the bias would be more pronounced for individuals with below-average introjected regulation. In general, however, this bias is expected to be small. The average absolute score difference in the latent trait space amounted to 0.44 points on the response scale, with the largest difference of 0.89 points at a latent trait of $\theta = -0.70$; in other words, students with a true introjected regulation level of -0.70 are, on average, expected to achieve a test score of 5.39 if they are about 10 to 11 years of age and a score of 4.49 at the ages of 16 to 17.

The results for the external subscale with regard to sex-related DIF are presented in the top right panel of Figure 1. For adolescents with the same true trait level, girls are expected to achieve slightly higher test scores than boys. Again, this bias is more pronounced for below-average trait levels; at a true external regulation of $\theta = -1.70$ girls are, on average, expected to achieve an observed test score of 3.01 whereas the expected score amounts to 2.31 for boys.

Group level impact. As the SRQ-A[G] is primarily designed as a screening instrument for educational research that predominantly focuses on between-group comparisons, we also investigated the impact of DIF for the four items on group level estimates. Figure 1 (bottom left panel) displays the mean trait level for the four age groups with and without DIF correction. In line with the statistical analyses identifying small DIF effects, the figure reveals only marginal differences with regard to group-level means in introjected regulation. Ignoring the DIF for the items would result in mean differences between 0.00 and 0.08 logits on the latent trait continuum: for the youngest students this would translate into a mean trait overestimation of $\Delta = .08$ ($SD = .06$) and for the oldest students in an underestimation of $\Delta = .05$ ($SD = .09$). With regard to external regulation (see bottom right panel of Figure 1) the observed differences for sex were even smaller. Hence, the

DIF identified in the four items confounded the estimation of the latent regulation styles only very modestly. Neglecting the small DIF would most likely be irrelevant for most practical applications.

Discussion

The relevance of self-determined motivations in the educational context has been recognized for a long time. An important challenge for teachers is the formation and enduring maintenance of self-determined motivations in their students. Research on different instructional methods, the conditions under which they are most effective, or potential academic outcomes of self-determined motivations requires measurement instruments with known psychometric properties. For a meaningful interpretation of mean group differences, in particular, it is necessary that the construct of motivation is assessed comparably in all groups, that is, measurement equivalence must hold. In this paper, we examined the measurement invariance of the German Academic Self-Regulation Questionnaire for adolescents (Müller et al., 2007), a short self-report scale for the assessment of four regulatory styles within Deci and Ryan's (1985) self-determination theory. Although DIF analyses identified small effects of sex and ages from 10 to 17 on the measurement of the perceived regulatory styles, the practical implications on individual and group-level estimates were rather small. Ignoring the identified DIF would somewhat impair the fairness of the comparison of results for individual students, because introjected motivation is slightly underestimated for older students and external motivation is overestimated for girls. However, when turning to group-level comparisons, which is the primary intention of the SRQ-A[G] as a tool for educational research, DIF had negligible consequences. Differences between uncorrected means and means corrected for DIF were very small and are most likely to have no noteworthy effects in practice.

We limited our DIF analyses to two criteria, sex and age, that have been frequently used in the past to study individual differences in self-determined motivations (e.g., Lepper et al., 2005; Marsh et al., 2008). Comparable analyses would be required for conducting group

comparisons based on different criteria, for example, for comparing gifted and average-ability students (cf. Preckel, Goetz, Pekrun & Kleine, 2008). While such analyses are beyond the scope of this paper, the adopted strategy using logistic regression tests presents an intuitive method for the study of DIF (Swaminathan, & Rogers, 1990; Zumbo, 1999) which is conveniently implemented in freely available software packages (e.g., Choi et al., 2011). Another advantage of the regression framework is the possibility to examine hypothesized sources of DIF. For example, the presented analyses provided some evidence for individual differences in response styles as a potential cause for DIF. Mirroring previous findings (Austin et al., 2006; De Jong et al., 2008), younger students tended to use extreme response options more frequently than older adolescents. As a consequence, one item previously identified as having age-related DIF failed to do so after controlling for ERS. Thus, response styles associated with age might account for some age-related DIF. Future research should extend this line of research, not only to identify DIF for additional criteria but also to systematically examine further causes of DIF, for example, cognitive competencies or even cultural values.

References

Amtmann, D., Cook, K. F., Jensen, M. P., Chen, W.-H., Choi, S., Revicki, D., … & Lai, J.-S. (2010). Development of a PROMIS item bank to measure pain interference. *Pain, 150,* 173-182.

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences, 40,* 1235–1245.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R package for detecting differential item functioning. *Journal of Statistical Software, 39,* 1-30.

Clauser, B. E., Nungster, R. J., Mazor, K., & Ripkey, D. (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measurement, 33,* 453-464.

Corpus, J., H., McClintic-Gilbert, M. S., & Hayenga, A.O. (2009). Within-year changes in children's intrinsic and extrinsic motivational orientations: Contextual predictors and academic outcomes. *Contemporary Educational Psychology, 34,* 154-166.

Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques. *Medical Care, 44,* 115-123.

Crane, P. K., Gibbons, L. E., Ocepek-Welikson, K., Cook, K., & Cella, D. (2007). A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Quality of Life Research, 16,* 69-84.

De Jong, M. G., Steenkamp, J. B., Fox, J. P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research, 45,* 104–115.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York, NY: Plenum Press.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70,* 662-680.

Embretson, S. E., & Reise, P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement, 23*, 309-326.

Gnambs, T., & Batinic, B. (2011). Evaluation of measurement precision with Rasch-type models. *Personality and Individual Differences, 50*, 53-58.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliabilty within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural Equation Modeling* (pp. 195-216). Lincolnwood, IL: SSI.

Harter, S., Whitesell, N. R., & Kowalski, P. (1992). Individual differences in the effects of educational transitions on young adolescent's perceptions of competence and motivational orientation. *American Educational Research Journal, 29*, 777-807.

Krapp, A. (2002). An educational-psychological theory of interest and its relation to self-determination theory (SDT). In E. L. Deci, & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 405-427). Rochester, NY: University of Rochester Press.

Lepper, M. R., Corpus, J. H., & Iyengar, S. S. (2005). Intrinsic and extrinsic motivational orientations in the classroom: Age differences and academic correlates. *Journal of Educational Psychology, 97*, 184–196.

Little, R., & Rubin, D. (1987). *Statistical analysis with missing data*. New York, NY: Wiley.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349.

Kankaras, M., Vermunt, J. K., & Moors, G. (2011). Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods & Research, 40*, 279-310.

Marsh, H. W., Martin, A. J., & Cheng, J. H. S. (2008). A multilevel perspective on gender in classroom motivation and climate: Potential benefits of male teachers for boys? *Journal of Educational Psychology, 100*, 78-95.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*, 568-592.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297-334.

Müller, F. H., Hanfstingl, B., & Andreitz, I. (2007). *Skalen zur motivationalen Regulation beim Lernen von Schülerinnen und Schülern* [Scales to assess motivational regulation during students' learning]. Wissenschaftliche Beiträge aus dem Institut für Unterrichts- und Schulentwicklung, No. 1. Klagenfurt, Austria: Alpen-Adria-Universität.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Otis, N., Grouzet, F. M. E., & Pelletier, L. G. (2005). Latent motivational change in an academic setting: A 3-year longitudinal study. *Journal of Educational Psychology, 97*, 170–183.

Preckel, F., Goetz, T., Pekrun, R., & Kleine, M. (2008). Gender differences in gifted and average-ability students. *Gifted Child Quarterly, 52*, 146-159.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistic, 4*, 207-230.

Rijavec, M., Saric, Z. R., & Miljkovic, D. (2003). Intrinsic vs. extrinsic orientation in the classroom and self-regulated learning. *Studia Psychologica, 45*, 51-63.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17*, 1-25.

Robitzsch, A., & Rupp, A. A. (2009). Impact of missing data on the detection of differential item functioning: The case of Mantel-Haenszel and logistic regression analysis. *Educational and Psychological Measurement, 69*, 18-34.

Ryan, R. M., & Connell, J. P. (1989). Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology*, *57*, 749-761.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 17*, 1-25.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning*. (pp. 67-113). Mahwah, NJ: Lawrence Erlbaum.

Van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software, 48*, 1-19.

Van Schuur, W. H. (2011). *Ordinal item response theory: Mokken scale analysis*. Los Angeles, CA: Sage.

Wetzel, E., Böhnke, J. R., Carstensen, C. H., Ziegler, M., & Ostendorf, F. (2013). Do individual response styles matter? Assessing differential item functioning for men and women in the NEO-PI-R. *Journal of Individual Differences, 34*, 69-81.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa, Canada: Directorate of Human Resources Research and Evaluation,

Department of National Defense.

Table 1

DIF statistics for the SRQ-A[G]

	Sex					Age				
	Non-uniform DIF		Uniform-DIF			Non-uniform DIF		Uniform-DIF		
	χ^2_{23}	ΔR^2	χ^2_{12}	ΔR^2	%Δβ	χ^2_{23}	ΔR^2	χ^2_{12}	ΔR^2	%Δβ
Intrinsic regulation										
IN1.	0.28	0.00	5.27^{+}	0.09	0.23%	5.68	0.09	1.51	0.02	0.24%
IN2.	0.41	0.01	7.28^{+}	0.13	0.37%	2.81	0.05	4.55	0.08	0.32%
IN3.	2.13	0.03	2.85	0.05	0.01%	4.27	0.07	11.86^{+}	0.19	0.52%
IN4.	1.04	0.02	14.41*	0.24	0.20%	5.49	0.09	11.59^{+}	0.19	0.89%
Identified regulation										
ID1.	1.07	0.02	10.70*	0.17	0.42%	0.26	0.00	12.27	0.20	0.12%
ID2.	0.72	0.01	30.45*	0.46	0.83%	3.02	0.05	4.08	0.07	0.53%
ID3.	0.19	0.00	20.96*	0.33	0.00%	4.00	0.07	9.49	0.17	0.17%
ID4.	2.94	0.05	11.92*	0.18	0.10%	11.01^{+}	0.19	7.96	0.14	0.29%
Introjected regulation										
IO1.	1.09	0.02	7.92^{+}	0.13	0.64%	1.61	0.03	27.73*	0.44	1.14%
IO2.	3.71	0.07	3.39	0.06	0.42%	2.22	0.03	3.58	0.05	0.38%
IO3.	8.83*	0.15	8.62^{+}	0.15	0.06%	2.98	0.05	6.91	0.11	0.05%
IO4.	0.02	0.00	1.66	0.03	0.07%	1.63	0.03	17.08^{+}	0.26	1.09%
External regulation										
EX1.	3.57	0.06	1.48	0.03	0.23%	2.70	0.05	2.37	0.04	0.11%
EX2.	0.00	0.00	0.10	0.00	0.07%	0.52	0.01	2.75	0.05	0.02%
EX3.	0.26	0.00	24.24*	0.41	2.17%	5.15	0.09	12.41^{+}	0.21	0.61%
EX4.	2.42	0.04	28.71*	0.45	2.30%	1.51	0.02	18.63^{+}	0.30	0.03%

Notes. N = 2,123. Mixed effects ordinal logistic regressions; χ^2_{23} … Chi^2 difference for model 2 and 3 (see text); χ^2_{12} … Chi^2 difference for model 1 and 2; $df_{Sex} = 1$, $df_{Age} = 7$; ΔR^2 … Change in McFadden's pseudoR^2 from model 1 and 2 multiplied by 100 (critical value for small DIF: 3.5); %Δβ … Change in regression coefficient from model 1 and 2 (in percent; critical value for small DIF: 1.0); Gray elements indicate small DIF;

$p < .05$ * with and $^{+}$ without adjustment for multiple comparisons

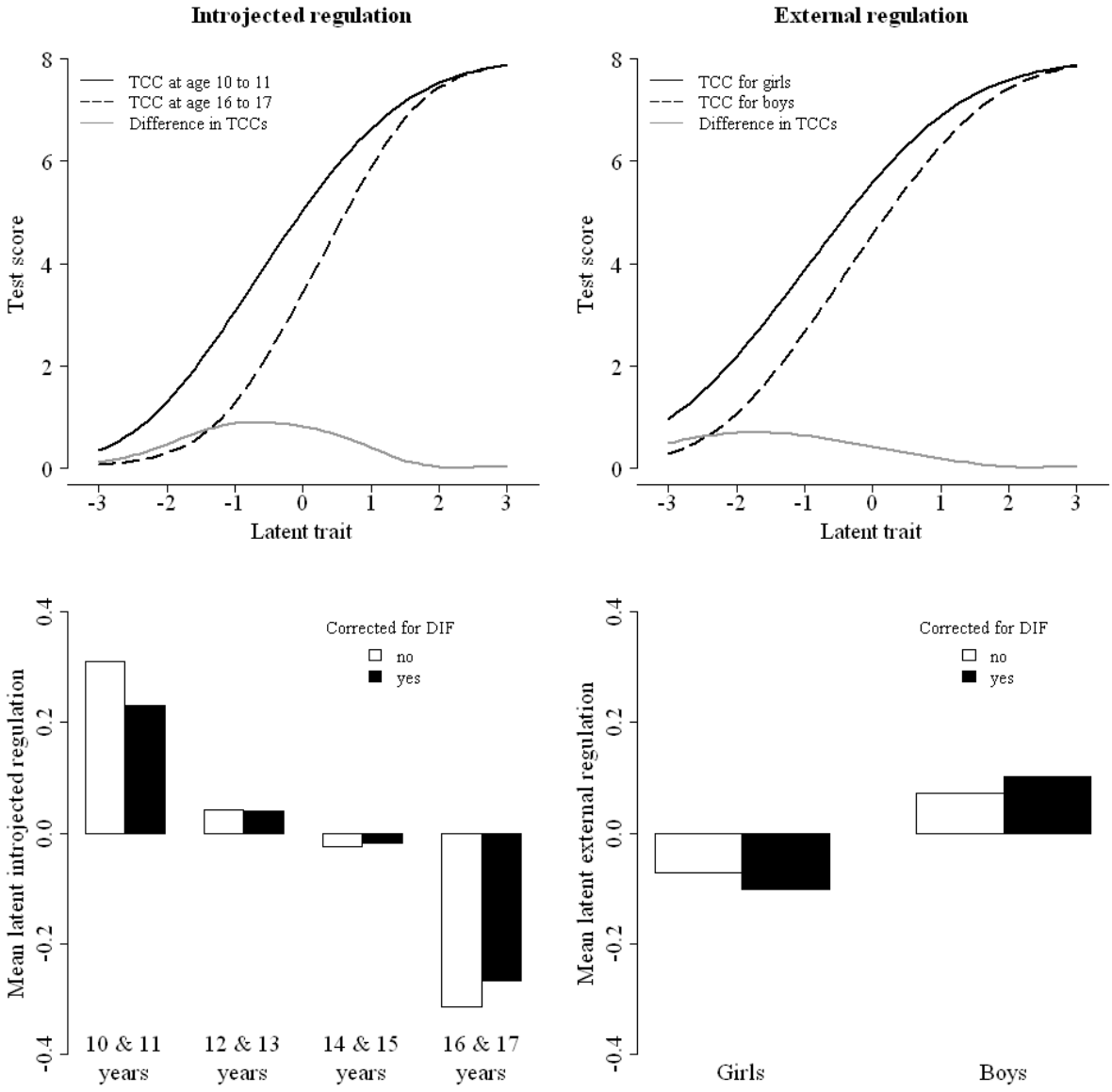

Figure 1. Test characteristic curves (TCC; top panel) and latent means (bottom panel) for regulatory scales with sex- or age-related DIF

Appendix: German Academic Self-Regulation Questionnaire for adolescents (SRQ-A[G])

Item No.	English	German (Müller et al., 2007)
	I work on my classwork…	Ich arbeite und lerne in der Schule…
	Intrinsic regulation	
IN1	because it's fun*	weil es mir Spaß macht
IN2	because I want to learn new things*	weil ich neue Dinge lernen möchte
IN3	because I enjoy thinking and reflecting about things in this subject*	weil ich es genieße, mich mit dem Fach auseinanderzusetzen
IN4	because I enjoy solving tasks in this subject	weil ich gerne Aufgaben aus dem Fach löse
	Identified regulation	
ID1	so in the future, I can continue my education	um später eine bestimmte Ausbildung machen zu können (z.B. Schule, oder Studium)
ID2	because it will give me better career choices	weil ich damit mehr Möglichkeiten bei der späteren Berufswahl habe
ID3	because the knowledge in the subject will allow me to get a better job*	weil ich mit dem Wissen im Fach später einen besseren Job bekommen kann
ID4	because the things that I learn here will be useful in the future*	weil ich die Sachen, die ich hier lerne, später gut gebrauchen kann
	Introjected regulation	
IO1	because I want the teacher to think I'm a good student*	weil ich möchte, dass meine Lehrerin denkt, ich bin ein/e gute/r Schüler/in
IO2	because otherwise I would have a guilty conscience	weil ich ein schlechtes Gewissen hätte, wenn ich wenig tun würde.
IO3	because I want other students to think I am quite good**	weil ich möchte, dass die anderen Schüler von mir denken, dass ich ziemlich gut bin
IO4	because I would feel ashamed of myself if I don't try**	weil ich mich vor mir selbst schämen würde, wenn ich es nicht tun würde
	External regulation	
EX1	because otherwise I would get into trouble at home**	weil ich sonst von zu Hause Druck bekomme
EX2	because otherwise I would get into trouble with my teacher**	weil ich sonst Ärger mit meiner Lehrerin bekomme
EX3	because otherwise I would get bad grades	weil ich sonst schlechte Noten bekomme
EX4	because that's what I'm supposed to do*	weil ich es einfach lernen muss

Notes. Items are presented with a five-point response scale (1 = strongly disagree to 5 = strongly agree). * Original from the SRQ-A (Ryan & Cornell, 1989), ** Adapted from the SRQ-A (Ryan & Cornell, 1989).