Measurement Invariance of Divergent Thinking Across Gender, Age, and School Forms

Jörg-Tobias Kuhn and Heinz Holling

Westfälische Wilhelms-Universität Münster, Germany

Abstract. The present study explores the factorial structure and the degree of measurement invariance of 12 divergent thinking tests. In a large sample of German students (N = 1328), a three-factor model representing verbal, figural, and numerical divergent thinking was supported. Multigroup confirmatory factor analyses revealed that partial strong measurement invariance was tenable across gender and age groups as well as school forms. Latent mean comparisons resulted in significantly higher divergent thinking skills for females and students in schools with higher mean IQ. Older students exhibited higher latent means on the verbal and figural factor, but not on the numerical factor. These results suggest that a domain-specific model of divergent thinking may be assumed, although further research is needed to elucidate the sources that negatively affect measurement invariance.

Keywords: divergent thinking, creativity, measurement invariance, latent means analysis

Introduction

At the most general level, creativity has been defined as the capability to produce work that is novel, original, and useful and that fits within task constraints (Lubart, 1994). Many structural theories of intelligence comprise a factor corresponding to person-related facets of creativity (Jäger, 1984), underlining its importance as part of the human cognitive taxonomy. In addition, individual traits that enable creativity appear to be related to the development of reasoning ability and psychological health, as well as late-life adaptation and growth (Guignard & Lubart, 2006; Runco & Charles, 1997).

At the level of the individual person, a cognitive ability central to creativity is divergent thinking (DT), which can be defined as the capability to generate diverse and numerous ideas (Runco, 1991). DT tests are often used to estimate the individual potential for creative problem solving, while creativity itself can be regarded as a rather complex social construct that cannot be measured at the individual level (Westmeyer, 2001). Guilford (1950) identified three basic components as factors of DT: fluency (the total number of ideas generated), flexibility (the number of categories in the ideas), and originality (the number of unique or unusual ideas). Numerous DT tests allow for scoring each of these components, although the results are usually highly correlated, with the use of anything but fluency scores adding little useful information (Hargreaves & Bolton, 1972).

Almost all DT tests reported in the literature focus on verbal or figural content, neglecting the numerical domain (Cropley, 2000). This is surprising, as numerical content plays a major role in research on convergent thinking (i.e., reasoning and intelligence) and real-life problems, which currently is not adequately mirrored in DT research. Hence, numerical content appears to be a promising venue for research on creative problem-solving and DT. Several papers have investigated the factorial structure pertaining to different DT *operations* (Kim, Cramond, & Bandalos, 2006), but it is unclear whether a model including domain-specific and content-related DT factors adequately describes the data. An investigation of this question, therefore, seems warranted.

An important assumption that has rarely been tested when comparing fluency scores across groups is the measurement invariance (MI) of the DT tests utilized. MI refers to the extent to which items or tests have the same meaning across groups of examinees (Gregorich, 2006). That is, an investigation of MI can reveal whether a test is systematically biased against a specific subpopulation of participants, or whether an array of tests refers to the same latent variables, to the same degree, across groups. Research questions in this context might be whether the same factor model of DT is valid for males and females, or whether a possible advantage of older students compared to younger ones on specific DT test scores is caused by a higher latent DT ability or measurement artifacts that are unrelated to DT (e.g., test sophistication). Assessing MI, therefore, helps to decide whether observed scores can be attributed to latent (factor) scores or unrelated sources (Wicherts, Dolan, & Hessen, 2005). Further, as discussed below, MI is a central prerequisite for latent mean comparisons across groups.

DOI 10.1027/1015-5759.25.1.1

Establishing MI requires fitting a sequence of nested, increasingly restricted confirmatory factor analysis (CFA) models (Gregorich, 2006; Vandenberg & Lance, 2000). MI is best assessed using multiple-groups CFA, because this approach allows all central aspects of MI to be statistically tested. The MIMIC model (Muthén, 1989) has been used to assess aspects of MI as well, but this model assumes all factor loadings and residual variances to be equal across groups, thereby defeating a central aspect of MI analysis. We therefore utilized multiple-groups CFA in this study. We investigated five increasingly restrictive forms of MI. *Configural MI* requires equal patterns of factor loadings across groups, constraining construct dimensionality to be equivalent. *Metric MI* constrains all factor loadings to be invariant across groups, and is nested in the configural MI model. A further, more restrictive model nested in the metric MI model additionally assumes equal residuals across groups, thereby assuming equal reliabilities of measures across groups. The next two models impose restrictions on the measurement intercept structure: *Strong MI* assumes both factor loadings and intercepts to be invariant across groups, whereas *strict MI* imposes constraints on intercepts, factor loadings, and residuals, respectively. Whereas strong MI is nested in the metric MI model, strict MI is nested in the model assuming equal residuals. A weaker form of strong measurement invariance is *partial strong MI*, where some, but not all, intercepts are fixed across groups. Although partial invariance can be met within any type of MI (Gregorich, 2006), we investigated only partial strong MI here because it is a minimum requirement for latent mean comparisons (Thompson & Green, 2006), which are of interest in this paper. We refer the reader to Gregorich (2006) for a description of valid group comparisons at different levels of MI.

According to Meredith (1993), strict MI is necessary in order to unequivocally ascribe observed mean differences to latent mean differences. Some authors, however, have argued that strong MI is sufficient to compare latent means, because group-specific residuals merely reflect differing reliabilities across groups (e.g., Little, 1997). In addition, Byrne, Shavelson, and Muthén (1989) have pointed out that releasing selected measurement intercepts still allows for latent mean comparisons (see Thompson & Green, 2006). With respect to unequal residuals across groups, Lubke and Dolan (2003) showed that sample sizes of 100 participants are sufficient to detect mean differences of .5 standard deviations even if group reliabilities are different. In line with these authors, we regard strong (or strict) MI as a desideratum for latent mean comparisons, although a cautious interpretation of latent mean differences under partial strong MI is possible. In order to enhance comparability, we computed standardized effect measures of latent mean differences (Hancock, 2001).

Currently, the only study we are aware of that systematically tested for MI in creativity research was recently conducted by Kim et al. (2006). These authors investigated MI across gender and age in a sample of 3000 kindergartners, third graders, and sixth graders, using the Torrance Tests of Creative Thinking (TTCT; Torrance, 1990). In this study, MI across gender was largely achieved, whereas a lack of MI was reported across grades, implying that a comparison of observed TTCT scores across grades needs caution for interpretation.

Given that MI holds, the comparison of latent means offers a theoretically interesting alternative to analyzing group differences based on observed means, because construct differences, not indicator differences, are at the core of individual differences research. In addition, under normality assumptions, latent mean comparisons have been shown to be more powerful than statistical tests comparing observed means, even in the case of noninvariant latent variable systems (Yuan & Bentler, 2006). Therefore, in this study we used latent mean comparisons to investigate differences in DT across groups.

The Present Study

The aims of the present study were threefold. First, a domain-specific model of DT tests across verbal, figural, and numerical domains was estimated in order to assess its model fit. Second, we tested whether the MI of this model could be assumed across gender, age, and school forms (lower and middle tracks vs. higher track). Third, based on these results, a comparison of latent means was conducted in order to test the hypothesized latent mean differences across groups.

Research on DT has provided a mixed picture with respect to its stability across different subpopulations. Concerning gender differences, Kim and Michael (1995) found significantly higher fluency scores for female students in one visual and two verbal creativity subtests from the TTCT (Torrance, 1990). Dudek, Strobel, and Runco (1993) also reported higher mean TTCT scores for girls, although a small interaction effect of gender and socioeconomic status (SES) resulted in marginally better scores for boys with a higher SES status. In contrast, Cheung, Lau, Chan, and Wu (2004) were unable to find gender differences in fluency in a Hong Kong-based student sample, a finding that the authors describe as culture-specific (Lubart, 1999). Based on these results, higher DT ability for girls can be expected in Western cultures (Hypothesis 1).

The impact of age on DT results varies widely between studies as well. Using a longitudinal design, Torrance (1968) was the first to note that DT test scores of students significantly decreased around the fourth grade, but later showed a subsequent increase, a phenomenon known as the fourth-grade slump. These results pertain to a specific age cohort, however. Concerning older students, a longitudinal study (Claxton, Pannells, & Rhoads, 2005) found no significant change in fluency scores between the fourth, sixth, and ninth grade at all, but the sample size was very small (N = 25). In contrast, Smith and Carlsson (1985) reported

a steady rise in DT test scores after age 12 until age 16 (N = 146), which supports the assumption of Cropley (2003) that creativity peaks at the end of adolescence and early adulthood. Therefore, a steady rise in DT ability between the ages of 12 and 16 can be expected (Hypothesis 2).

A third factor exerting an influence on DT scores in students is school type and, related to it, the average intelligence of students. Because the German education system separates students after grade 4 into different tracks based on their prior level of scholastic achievement, school type can be expected to have an implicit impact on DT scores, because more gifted students will attend the higher track of the German school system as compared to the middle and lower track. The reason for this is that scholastic achievement and intelligence are highly correlated (r = .81 in Deary, Strand, Smith, & Fernandes, 2007). The only study we are aware of comparing DT scores between school forms was conducted by Grampp and Grampp (1977). These authors analyzed TTCT scores from 72 students from all tracks of the German education system, resulting in significantly better fluency scores for the lower track students compared to the middle track students on the two figural tasks of the TTCT. However, the sample used in this single study was small and unrepresentative, and did not provide information on intelligence test scores. Evidence for a positive relation between intelligence and DT ability was provided in a meta-analysis by Kim (2005), who reported a mean correlation between intelligence test scores and TTCT scores of r = .22. Based on this finding, we hypothesized a higher average DT ability in the higher tracks of the German education system compared to the lower tracks (Hypothesis 3).

Method

Sample

Data came from the standardization sample of the Berliner Intelligenzstruktur-Test für Jugendliche: Begabungs- und Hochbegabungsdiagnostik (BIS-HB [Berlin Structure-of-Intelligence test for Youth: Diagnosis of Talents and Giftedness]; Jäger et al., 2005). A total of N = 1328 students were tested (728 males and 598 females, two participants gave no information concerning gender). In a subsample (n = 713), the German adaptation of the Culture Fair Intelligence Test (Cattell & Cattell, 1960) was administered in order to assess IQ. A one-way ANOVA revealed significant mean differences in IQ between school forms, $F(3, 709) = 9.54$, $p < .01$. Tukey's post hoc multiple comparisons revealed that the lower and the middle track significantly differed from the higher and gifted track, respectively, with no significant difference between lower and middle track, or higher and gifted track. We therefore merged the lower and middle track subsamples into one group, and the top and gifted track subsamples into another. The lower and middle tracks were attended by n = 424 participants, whereas n = 894 participants attended the top and gifted track (10 participants gave no information on school). Mean IQ of students in the lower and middle tracks was 108 compared to 122 in the top and gifted track group. Mean age of the sample was 14.5 years (SD = 1.1 years). The range was from 12 years and 6 months to 16 years and 6 months), with 305 participants falling into the age interval from 12.6 to 13.5 years (Age Group 1), 354 from 13.6 to 14.5 years (Age Group 2), 358 from 14.6 to 15.5 years (Age Group 3), and 311 from 15.6 to 16.5 years (Age Group 4). These age intervals correspond to the standardized age groups of the BIS-HB.

Measures

The BIS-HB contains 12 DT tests overall, which are based on the Berlin model of intelligence structure (BIS; Jäger, 1984), with four falling into each of the three content domains: (a) Verbal: Specific traits, Masselon, Insight test, Possible object uses; (b) Numerical: Divergent computing, Inventing telephone numbers, Equations with numbers, Puzzle with numbers; (c) Figural: Layout, Symbol competition, Object designing, Symbol combining (for a detailed description of the DT tests used in this study see Bucik & Neubauer, 1996). Each DT test was administered with a prespecified time limit and scored for fluency. Flexibility scores were available for five subtests, but discarded because they showed very high correlations with fluency scores (r = .78–.87) and did not change the results. The unadjusted intraclass correlation coefficient as a measure of objectivity of scoring between the ratings of two independent raters showed satisfactory values for all DT tests (M = .94, SD = .04). The overall internal consistency (Cronbach's αs) over the 12 DT tests scored for fluency was found to be satisfactory (α = .84; verbal DT: α = .76; figural DT: α = .65; numerical DT: α = .60).

Statistical Modeling Procedures

All CFA analyses were conducted using MPlus, version 4.2 (Muthén & Muthén, 2006). The following indices were used for assessing model fit : (1) a rescaled χ^2 test, the Satorra-Bentler χ^2 statistic (SB-χ^2; Satorra & Bentler, 2001), which is robust with respect to data significantly departing from multivariate normality; (2) the comparative fit index (CFI); (3) the Bayesian information criterion (BIC); (4) the root mean squared error of approximation (RMSEA), adjusted for multiple groups; (5) a χ^2-difference test; and (6) Steiger's (1989) $\hat{\gamma}$, as well as McDonald's (1989) noncentrality index (Mc), which are unaffected by model complexity and sample size in MI evaluation (Cheung & Rensvold, 2002). CFI values above .90 indicate an adequate model fit, whereas RMSEA values less than .08 can be regarded as acceptable (Browne & Cudeck, 1993). The RMSEA is substantially less affected by sample size than other fit indices, and confidence intervals can be computed for this fit index even under nonnormality conditions.

Table 1. Fit indices for MI model comparisons across gender

MI [a]	*df*	SB-χ^2	Compare with	ΔSB-χ^2	CFI	BIC	$\hat{\gamma}$	Mc	RMSEA (95%CI)
COI	102	325.91			.944	70405	.972	.92	.058 (.049–.066)
MEI	111	331.57	COI	5.34	.944	70346	.973	.92	.055 (.047–.063)
ER	123	360.37	MEI	23.17*	.940	70297	.971	.91	.054 (.046–.062)
SOI	120	420.79	MEI	89.01**	.924	70374	.963	.89	.061 (.054–.069)
SOI-p [b]	117	337.50	MEI	5.99	.944	70308	.973	.92	.053 (.045–.061)
SOI-pm [b]	120	382.34	SOI-p	43.74**	.934	70333	.968	.91	.057 (.050–.065)
SII	132	449.85	ER	89.01**	.920	70327	.961	.89	.060 (.053–.068)

Note. [a]MI-model: COI = configural MI; MEI = metric MI; ER = equal residuals; SOI = strong MI; SOI-p = partial strong MI; SOI-pm = partial strong MI with equal means across groups; SII = Strict MI. [b]Latent intercepts of three DT tests (Insight test, Possible object uses, Symbol combining) varied between groups. *$p < .05$, **$p < .01$.

Results

Univariate skewness of all DT tests was found to significantly depart from normality, except for Masselon, $Z(\sqrt{b_r}) = .90$, $p > .05$. Further, analysis of univariate kurtosis revealed that five DT tests substantially deviated from normality as well. Therefore, univariate normality was not given in the data. This finding was complemented by the absence of multivariate normality (Mardia's multivariate skewness: $b_{1,12} = 5.53$, $A = 1224.49$, $p < .01$; Mardia's multivariate kurtosis: $b_{2,12} = 183.75$, $Z = 15.65$, $p < .01$). Hence, we estimated all model parameters, standard errors, fit indices, and ΔSB-χ^2 utilizing the aforementioned rescaling procedure.

Gender

To investigate Hypothesis 1 regarding gender differences, we first tested a three-factor solution model. We found that the postulated three-factor solution including a verbal, a numerical, and a figural factor, as incorporated in the configural invariance model, showed satisfying model fit (Model COI in Table 1). A model assuming a single DT factor exhibited a significant drop in model fit (BIC = 70647), therefore, we retained the model with three factors. A higher-order model with a second-order general DT factor was not fit to the data, because it was underidentified for multiple group comparisons with three first-order factors (Chen, Sousa, & West, 2005). Next, we investigated metric invariance (Model MEI) and found this model to be tenable. This indicates that equal factor loading across groups could be assumed without a significant drop in model fit. However, according to ΔSB-χ^2, the restriction imposed on the residual variances (Model ER) resulted in a significant deterioration of model fit when compared with the metric invariance model. This could be a result of the large sample size, because all other fit indices indicated that the misfit from this restriction was not very large. More importantly, subsequent SB-χ^2 difference tests revealed that strong and strict variance (Models SOI and SII, respectively) both showed insufficient fit. This implies that measurement intercepts across gender were not equal, and that mean differences in test scores cannot unequivocally be explained by latent mean differences between male and female groups.

Hence, in the next step, we tried to fit a partial strong MI model. An inspection of the univariate Lagrange multiplier tests (LM), indicating which parameter constraints strongly affected model fit, showed large values for the intercepts of three DT tests (Insight test: LM = 54; Possible object uses: LM = 18; Symbol combining: LM = 10). We freed these intercepts across groups, resulting in a significantly improved model fit (Model SOI-p). We used this partial strong MI model as the baseline model to conduct latent mean comparisons, keeping in mind that factor mean differences were now only based on the nine DT tests whose intercepts were fixed. An additional model was specified (Model SOI-pm) that was identical with Model SOI-p, except that latent means were fixed to be equal across groups. As can be seen from Table 1, Model SOI-pm fit the data significantly worse than Model SOI-p, indicating that overall, latent means differed significantly between males and females. In order to assess the magnitude of these latent mean differences, standardized effect sizes (ES) were computed (similar to Cohen's *d*; Hancock, 2001) for all factors, yielding ES = .20 for the verbal factor, ES = .29 for the numerical factor, and ES = .50 for the figural factor, respectively, with consistently higher latent means in the female group. According to Cohen's effect-size conventions (Cohen, 1988), these differences can, therefore, be classified as small to medium. It should be noted that the partial strong MI model used here is exploratory in character, and that mean differences on the tests with free intercepts across groups are now unexplained by their respective factor. Hence, most, but not all, differences between males and females in DT test scores could be explained by differences in DT ability.

Age Groups

To test Hypothesis 2, we first investigated MI across age groups (Table 2). We found that a good model fit could be observed up to the assumption of equal residual variances,

Table 2. Fit indices for MI model comparisons across age

MI [a]	*df*	SB-χ^2	Compare with	ΔSB-χ^2	CFI	BIC	$\hat{\gamma}$	Mc	RMSEA (95%CI)
COI	204	422.53			.945	71086	.973	.92	.057 (.048–.066)
MEI	231	458.29	COI	34.57	.942	70929	.972	.92	.054 (.046–.063)
ER	267	478.41	MEI	16.10	.946	70704	.974	.92	.049 (.040–.057)
SOI	258	518.77	MEI	60.14**	.934	70795	.968	.91	.055 (.047–.063)
SOI-p [b]	252	486.96	MEI	28.76	.941	70806	.971	.91	.053 (.045–.061)
SOI-pm [b]	261	557.82	SOI-p	71.29**	.925	70814	.964	.89	.059 (.051–.067)
SII	294	538.30	ER	60.50**	.938	70570	.970	.91	.050 (.042–.058)

Note. [a]MI-model: COI = configural MI; MEI = metric MI; ER = equal residuals; SOI = strong MI; SOI-p = partial strong MI; SOI-pm = partial strong MI with equal means across groups; SII = strict MI. [b]Latent intercepts of two DT tests (Possible object uses, Puzzles with numbers) varied between groups. **$p < .01$.

Table 3. Fit indices for MI model comparisons across school forms

MI[a]	*df*	SB-χ^2	Compare with	ΔSB-χ^2	CFI	BIC	$\hat{\gamma}$	Mc	RMSEA (95%CI)
COI	102	280.26			.947	69561	.978	.93	.052 (.043–.060)
MEI	111	294.26	COI	13.22	.945	69510	.977	.93	.050 (.042–.058)
ER	123	387.17	MEI	71.19**	.921	69532	.968	.90	.057 (.050–.065)
SOI	120	446.15	MEI	155.15**	.903	69604	.960	.88	.046 (.038–.054)
SOI-p[b]	115	301.63	MEI	7.58	.944	69610	.977	.93	.050 (.042–.058)
SOI-pm[b]	118	520.22	SOI-p	218.59**	.880	69698	.951	.86	.072 (.065–.080)
SII	132	533.89	ER	151.94**	.880	69622	.952	.86	.054 (.047–.062)

Note. [a]MI-model: COI = configural MI; MEI = metric MI; ER = equal residuals; SOI = strong MI; SOI-p = partial strong MI; SOI-pm = partial strong MI with equal means across groups; SII = Strict MI. [b]Latent intercepts of five DT tests (Divergent computing, Masselon, Possible object uses, Equations with numbers, Symbol combining) varied between groups. **$p < .01$.

implying equal reliabilities (Model ER). Again, a three-factor solution fit significantly better than a model with a single DT factor (BIC = 71231), and was retained for further analyses. However, a deterioration in fit was observed with respect to the models requiring intercepts to be fixed across groups, that is, according to the ΔSB-χ^2 statistic, neither strong nor strict MI were tenable, although the rest of the fit indices showed satisfactory fit. We therefore specified a partial strong MI model (Model SOI-p), releasing the intercepts of the two DT tests with the largest LMs across groups (Possible object uses: LM = 9; Puzzles with numbers: LM = 11). This model showed an improved fit and did not differ significantly from the metric invariance model. Similar to before, we then compared Model SOI-p with a model that was identical except that all latent means were fixed to be equal across groups (Model SOI-pm). A significant drop in model fit was observed, indicating that latent mean differences existed between age groups. In order to quantify the magnitude of these effects, we computed standardized ESs. Age-related differences were found to be small (Cohen, 1988), with ES = .15 for the verbal factor, ES = .11 for the numerical factor, and ES = .20 for the figural factor, respectively. As was expected, older participants had consistently higher verbal and figural factor means. However, interestingly, the second and third age groups had higher means on the numerical factor than the fourth age group. In sum, age had a significant, albeit small, impact on latent means, where 10 out of 12 test score differences could be ascribed to latent mean differences.

School Form

To test Hypothesis 3, a comparison of MI models across school forms (lower and middle track vs. higher and gifted track) was conducted (see Table 3). A configural MI model (Model COI) fit the data well, as did a metric MI model, and again fit significantly better than a one-factor model (BIC = 69859), hence, the three-factor model was retained. MI models of equal residuals, strong and strict invariance, however, showed a large drop in model fit. Therefore, we released the intercepts of five subtests, based on their LM values, in order to investigate model fit of a partial strong MI model, which showed a significantly better model fit than the same models with latent means constrained to be equal across groups. Concerning standardized ESs, latent mean differences between the higher track group compared to the middle and lower track group yielded ES = .65 for the verbal factor, ES = .40 for the numerical factor, and ES = .92 for the figural factor, respectively. According to Cohen's (1988) effect-size conventions, these are medium to high effect sizes. Students with a higher intelligence, therefore, had better DT skills, especially in the figural domain.

Discussion

The goal of the present study was to analyze the factor structure of 12 DT tests and to investigate whether these

tests were measurement invariant across different subsamples. Results in all subsamples investigated were supportive of a model comprising three factors, verbal DT, figural DT, and numerical DT, respectively. The measurement of ideational fluency, therefore, does not appear to be limited to the verbal or figural domain, and the numerical domain should be taken into consideration when DT is being assessed.

Further, we evaluated the degree of MI, using a series of nested and increasingly restrictive CFA models. Because strong MI was not achieved in any model, we selectively released measurement intercept parameters in gender and age group models. We are well aware that selectively releasing parameter constraints to improve the model fit always runs the risk of capitalization on chance. However, we chose this exploratory approach because our interest was in conducting latent mean comparisons, which have both the advantage of comparing constructs that are free of measurement error and of having a higher statistical power (Yuan & Bentler, 2006).

For different gender groups, it was found that both configural and metric MI could be assumed, and a model with equal residuals exhibited excellent model fit as well. Strong factorial invariance was untenable because of some group-specific measurement intercepts. A partial strong MI model was used to compare latent means, providing evidence for a significantly higher figural DT capability in female students. This result is in line with Kim and Michael (1995), who found that female students scored higher on one of the TTCT subtests. In contrast to other studies (Dudek et al., 1993), we found only small gender differences in verbal or numerical DT. However, we are not aware of any other study in this area that both checked for the MI of its model and compared latent means. Hence, it is not clear to which extent the results of these studies can be generalized to the latent level.

In contrast to Kim et al. (2006), we found that for different age groups, a partial strong MI model adequately fit the data. Latent mean comparisons clearly showed that older students had consistently higher factor scores on verbal DT and figural DT, respectively, but not on numerical DT. This result is complementary to findings from Wu, Cheng, Ip, and McBride-Chang (2005), who reported that sixth-grade students scored consistently higher on figural DT tests than university students. More cross-sectional research using MI modeling is necessary to elucidate the reasons for these empirical differences.

We could also fit a partial strong MI model with respect to school type, resulting in higher latent means for students in the higher track, as was expected. Interestingly, the difference was largest on the figural DT factor, which falls into the same domain as the Culture Fair Intelligence Test utilized in this study. The difference was much lower for the numerical DT factor, albeit still of medium size.

To conclude, the present study provided support for partial strong MI across gender, age, and school forms. Substantial latent mean differences emerged for all DT factors between school forms and a figural DT factor in females, while smaller gender differences were found with respect to a verbal and numerical DT factor. Age groups differed moderately in latent means, with an unexpectedly low latent mean on the numerical factor for the oldest age group. However, strong or strict MI were untenable in all group comparisons, thus, conclusions with respect to latent mean differences must be made cautiously. A possible venue for future research would be a closer investigation of the intercept variability in several subtests across groups, which is unrelated to the actual latent ability investigated. Why, for example, do females exhibit a higher intercept in the Insight test, but males in the Possible object uses test (both DT verbal)? These and other questions need to be addressed in future research.

References

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.

Bucik, V., & Neubauer, A.C. (1996). Bimodality in the Berlin model of intelligence structure (BIS): A replication study. *Personality and Individual Differences, 21,* 987–1005.

Byrne, B.M., Shavelson, R.J., & Muthén, B.O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466.

Cattell, R.B., & Cattell, A.K.S. (1960). *Culture Fair Intelligence test: Scale 2.* Champaign, IL: Institute for Personality and Ability Testing.

Chen, F.F., Sousa, K.H., & West, S.G. (2005). Testing measurement invariance of second-order models. *Structural Equation Modeling, 12,* 471–492.

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255.

Cheung, P.C., Lau, S., Chan, D.W., & Wu, W.Y.H. (2004). Creative potential of school children in Hong Kong: Norms of the Wallach-Kogan Creativity Tests and their implications. *Creativity Research Journal, 16,* 69–78.

Claxton, A.F., Pannells, T.C., & Rhoads, P.A. (2005). Developmental trends in the creativity of school-age children. *Creativity Research Journal, 17,* 327–335.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Cropley, A.J. (2000). Defining and measuring creativity: Are creativity tests worth using? *Roeper Review, 23,* 72–79.

Cropley, A.J. (2003). *Creativity in education and learning: A guide for teachers and educators.* London: Kogan.

Deary, I.J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35,* 13–21.

Dudek, S.Z., Strobel, M.G., & Runco, M.A. (1993). Cumulative and proximal influences on the social environment and children's creative potential. *The Journal of Genetic Psychology, 154,* 487–499.

Grampp, G., & Grampp, H. (1977). Divergentes Denken bei Schülern verschiedener Schularten [Divergent thinking in students of different school forms]. *Psychologie in Erziehung und Unterricht, 24,* 319–325.

Gregorich, S.E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care, 44,* 78–94.

Guignard, J.-H., & Lubart, T.I. (2006). Is it reasonable to be creative? In J.C. Kaufman & J. Baer (Eds.), *Creativity and reason in cognitive development* (pp. 244–268). New York: Cambridge University Press.

Guilford, J.P. (1950). Creativity. *American Psychologist, 5,* 444–454.

Hancock, G.R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika, 66,* 373–388.

Hargreaves, D.J., & Bolton, H. (1972). Selecting creativity tests for use in research. *British Journal of Psychology, 63,* 451–462.

Jäger, A.O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven [Research on the structure of intelligence: Competing models, new developments, perspectives]. *Psychologische Rundschau, 35,* 21–35.

Jäger, A.O., Holling, H., Preckel, F., Schulze, R., Vock, M., Süß, H.-M. et al. (2005). *Berliner Intelligenzstruktur-Test für Jugendliche: Begabungs- und Hochbegabungsdiagnostik (BIS-HB)* [Berlin Structure-of-Intelligence test for Youth: Diagnosis of talents and giftedness]. Göttingen, Germany: Hogrefe.

Kim, J., & Michael, W.B. (1995). The relationship of creativity measures to school achievement and to preferred learning and thinking style in a sample of Korean high school students. *Educational and Psychological Measurement, 55,* 60–74.

Kim, K.H. (2005). Can only intelligent people be creative? A meta-analysis. *The Journal of Secondary Gifted Education, 16,* 57–66.

Kim, K.H., Cramond, B., & Bandalos, D.L. (2006). The latent structure and measurement invariance of scores on the Torrance Tests of Creative Thinking – Figural. *Educational and Psychological Measurement, 66,* 459–477.

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

Lubart, T.I. (1994). Creativity. In R.J. Sternberg (Ed.), *Thinking and problem solving* (pp. 289–332). San Diego, CA: Academic Press.

Lubart, T.I. (1999). Creativity across cultures. In R.J. Sternberg (Ed.), *Handbook of creativity* (pp. 339–350). New York: Cambridge University Press.

Lubke, G.H., & Dolan, C.V. (2003). Can unequal residual variances across groups mask differences in residual means in the common factor model? *Structural Equation Modeling, 10,* 175–192.

McDonald, R.P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification, 6,* 97–103.

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

Muthén, B.O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557–585.

Muthén, L.K., & Muthén, B.O. (2006). *Mplus user's guide* (4th ed.). Los Angeles, CA: Muthén & Muthén.

Runco, M.A. (1991). *Divergent thinking.* Norwood, NJ: Ablex.

Runco, M.A., & Charles, R.E. (1997). Developmental trends in creative potential and creative performance. In M. Runco (Ed.), *The creativity research handbook* (pp. 115–152). Cresskill, NJ: Hampton.

Satorra, A., & Bentler, P.M. (2001). A scaled difference χ^2 test statistic for moment structure analysis. *Psychometrika, 66,* 507–514.

Smith, G., & Carlsson, I. (1985). Creativity in middle and late school years. *International Journal of Behavioral Development, 8,* 329–343.

Steiger, J.H. (1989). *EzPATH: Causal modeling.* Evanston, IL: SYSTAT.

Thompson, M.S., & Green, S.B. (2006). Evaluating between-group differences in latent variable means. In G.R. Hancock & G.R. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 119–169). Greenwich, CT: Information Age Publishing.

Torrance, E.P. (1968). A longitudinal examination of the fourth-grade slump in creativity. *Gifted Child Quarterly, 12,* 195–199.

Torrance, E.P. (1990). *The Torrance Tests of Creative Thinking Norms – Technical manual figural (streamlined) Forms A & B.* Bensenville, IL: Scholastic Testing Service.

Vandenberg, R.J., & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4–70.

Westmeyer, H. (2001). Kreativität: Eine relationale Sichtweise [Creativity: A relational perspective]. In E. Stern & J. Guthke (Eds.), *Perspektiven der Intelligenzforschung* [Perspectives in intelligence research] (pp. 233–249). Lengerich, Germany: Pabst.

Wicherts, J.M., Dolan, C.V., & Hessen, D.J.(2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology, 89,* 696–716.

Wu, C.H., Cheng, Y., Ip, H.M., & McBride-Chang, C. (2005). Age differences in creativity: Task structure and knowledge base. *Creativity Research Journal, 17,* 321–326.

Yuan, K.-H., & Bentler, P.M. (2006). Mean comparison: Manifest variable versus latent variable. *Psychometrika, 71,* 139–159.

Jörg-Tobias Kuhn

University of Münster
Psychological Department IV
Fliednerstr. 21
D-48149 Münster
Germany
Tel. +49 251 8334127
Fax +49 251 8339469
E-mail t.kuhn@uni-muenster.de