CREATIVITY RESEARCH JOURNAL, *20*(3), 309–318, 2008

ISSN: 1040-0419 print/1532-6934 online
DOI: 10.1080/10400410802278776

Routledge
Taylor & Francis Group

Creativity in the Classroom: A Multilevel Analysis Investigating the Impact of Creativity and Reasoning Ability on GPA

Philipp Alexander Freund and Heinz Holling
Westfälische Wilhelms-Universität Münster

Compared to other psychological constructs, such as intelligence, creativity usually is a less valid predictor for scholastic achievement. The majority of the studies from which these results are derived apply Ordinary Least Squares (OLS) methods that assume random sampling of respondents and, hence, do not take into account the possibility of between-classroom variance that is due to clustered data structures. In this study, data from the standardization sample of the Berlin Structure of Intelligence Test for Youth: Assessment of Talent and Giftedness (BIS-HB) were analyzed using multilevel modelling techniques. To test the hypothesis that the impact of creativity on GPA may vary between different classrooms, a multilevel model incorporating explanatory variables on two levels (level 1: students and level 2: classrooms) was specified. The role of reasoning ability was also investigated. The results allow for a more detailed interpretation of the role of different variables in the context of predicting scholastic achievement. More specifically, it could be shown that especially the predictive power of creativity changes across classrooms, indicating that some teachers value creativity in their students more than others.

Creativity is generally valued highly in contemporary (Western) societies. In fact, it is one of the most coveted psychological qualities (Sternberg & Lubart, 1999). Creativity is considered so important, there is undisputed belief that it should be fostered in children from early on, also or even especially in an educational context (Williams, 2002). On the other hand, though, Westby and Dawson (1995) were able to show that while teachers at school generally value creativity in their students, they quite often do, in fact, devalue creative behaviors exhibited by the very same students in the classroom.

Of course, the advantages and promises associated with high levels of creativity are almost countless, and will not be discussed in detail here (but see, for instance, the edited volume by Sternberg, 1999). Many different attempts at measuring creativity have been made, resulting, for instance, in criteria for the assessment of the degree of creativity in products or outputs (e.g., MacKinnon, 1978), checklists (e.g., Hocevar, 1981), self-report questionnaires (e.g., Runco, Plucker, & Lim, 2000), or psychometric tests (e.g., Torrance, 1966). Apart from the "pure" creativity domain, the construct can also be used for the prediction of other external criteria. In educational and organizational settings, academic and work performance, respectively, are of major importance, and one may be inclined to ask if creativity can be used to predict academic and/or work performance. In the present study, the focus is on the predictive power of creativity for such criteria.

In the past, some studies were conducted to shed light on that question. Concentrating on a professional domain, Milgram and Hong (1993), for example, showed creativity to be a more valid predictor of outstanding adult accomplishments than measures of intelligence or school grades. On the other hand, there already is a psychological construct that has repeatedly been found to correlate the highest with professional

Correspondence should be sent to Philipp Alexander Freund, Westfälische Wilhelms-Universität Münster, Psychological Institute IV, Fliednerstr. 21, 48149 Münster, Germany. E-mail: pafreund@uni-muenster.de

performance in general, compared to any other predictor. This is the construct of (analytical resp. general) intelligence, or *g*. Using meta analysis, Schmidt and Hunter (1998), or Ones, Viswesvaran, and Dilchert (2005), for example, showed the superiority of *g* in the prediction of professional performance compared to any other variable used in personnel selection.[1]

Hence, if creativity could be used to finer differentiate between high performers, this would emphasize its eminence. It is important to note at this point that in some theories of intelligence, creativity is viewed as representing an integral part of the intelligence construct as a whole. This view can be found in the Jäger model (1967, 1982, 1984), for example, and Guilford's (1967) Structure of Intellect (SOI) model also includes creativity.[2] In this article, *reasoning ability* is referred to as the most central component of analytical intelligence, and hence it is used as a synonym for the term intelligence.

Looking at the relationship of creativity and achievement in a setting that has a more short-term character, i.e., the prediction of school performance, the empirical evidence is somewhat less convincing than one would hope for. Most studies have yielded correlation coefficients of around $r = .20$ to $r = .30$ (e.g., Cline, Richards, & Needham, 1963; Cortis, 1968; Niaz, de Nunez, & de Pineda, 2000), but for intelligence, larger *r*s of around .40 to .50 are usually obtained (e.g., Rindermann & Neubauer, 2004; Süß, 2001). Jensen (1998) stated that this correlation is even higher for early school grades, but subsequently decreases with higher age, a phenomenon that he attributed to the constant drop-out of students from the lower tail of the ability distribution and the resulting restriction of range.

Thus, although there is a rather clear-cut picture showing reasoning ability to be a valid predictor of school grades, the same cannot be said unambiguously about creativity. First, the number of studies analyzing the predictive value of creativity is way smaller, and second, if creativity is the focus of such endeavours, the question of operationalization is a crucial one. Although the Torrance (1966) tests are one instrument that has been used rather extensively, there is, just like with reasoning tests, a large variety of different psychometric tests in use. In general, it can be said that different creativity tests correlate much lower than common reasoning tests because different aspects of the creativity construct may be tapped by the different tests (e.g., originality, fluency, etc.; see below). This renders the state of knowledge derived from the respective studies less cohesive. Finally, and most important, virtually all these studies did not take into account the clustered structure of respondents that researchers usually meet in educational contexts (students are grouped in classrooms, which are, in turn, clustered at the school level, etc.). The next section deals more comprehensively with this aspect.

Concluding from the empirical evidence obtained so far, then, it may tentatively be stated that creativity can be used to predict extracurricular achievement in some domains, possibly including educational and professional contexts, although intelligence is a proven predictor of academic success (and extracurricular success in non-creative domains).

MULTILEVEL MODELLING

In educational settings, the application of traditional analysis techniques, like multiple regression or ANOVA, results in the investigation of responses at the student level, and methods like Ordinary Least Squares (OLS) are usually applied. These methods assume random and independent sampling of respondents, a requirement that is generally not met in educational settings. Figure 1 depicts two sampling schemes. Panel A shows the case of random sampling of respondents. Every subject (represented by the little dots; filled dots indicate selection) has the same probability of being included in the study, the respondents are sampled randomly and independently from each other. In contrast, Panel B shows the realistic case most often met in educational research endeavours. Here, the respondents are clustered within units at a higher level (e.g., classrooms). Then, these classrooms are grouped within schools, schools within districts, and so on.

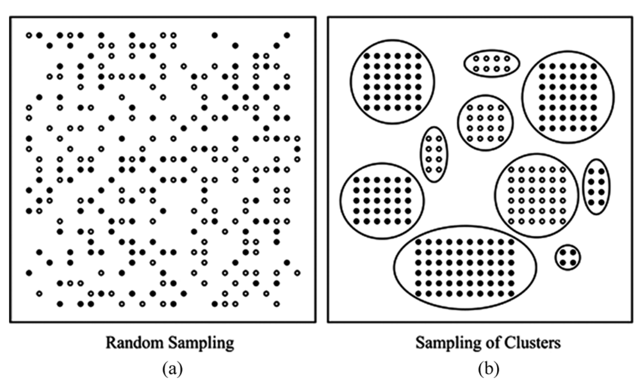

FIGURE 1 Random sampling of respondents vs. sampling of respondent clusters.

[1]Naturally, some studies report contradictory evidence, as results largely depend on the criteria used, and the sample studied. Overall, though, the meta-analytic results are very clear.

[2]Guilford, however, thought that divergent production was orthogonal to convergent production, and he strived to measure the former with tests of originality and the latter with intelligence tests.

When using OLS, the hierarchical nature of this clustering and the additional information it might provide is ignored. Students attending *a priori* identifiable higher level units (classroom, schools, etc.) share a common (educational, social, environmental, etc.) background and may be more alike than students from other higher level units. This may have an influence on all sorts of dependent variables that could be of interest for researchers. In other words, the assumption of random sampling is violated, and therefore, it is a good idea to take into account the cluster structure of the respondents and model it statistically (Goldstein, 2003). In an educational research context, this approach ensures two advantages over student-level-only analyses. The first is that bias is removed from the statistical inferences for the independent variables in the model. Generally, the standard errors in multilevel analyses will be larger than in OLS analyses, resulting in more conservative interpretations of effects. The second advantage is that the influence of the clustering levels can be assessed (Goldstein, 1997). The statistical modelling technique appropriate in these cases is called multilevel modelling. Also known as hierarchical modelling, the classic exposition of this method can be found in Aitkin and Longford (1986), and multilevel modelling has since then experienced growing popularity, which is expressed through a large body of applications and further refinements.

The general rationale behind multilevel modelling is the relation of a dependent variable to membership of different institutions at a higher aggregation level. Thus, analyzing the effects of the hierarchical clustering of the respondents becomes possible, and models with increasing complexity allow the researcher to test specific hypotheses about the nature of the relationships between certain variables at the different levels specified (see also Goldstein, 2003, or Raudenbush & Bryk, 2002).

THE BERLIN INTELLIGENCE STRUCTURE (BIS) MODEL

With the BIS model, Jäger (1967, 1982, 1984) suggested a hierarchical model, which was derived empirically through a research synthesis of all existing intelligence and creativity tasks that were, or had ever been, in use until the 1970's and 1980's. Jäger categorized approximately 2000 different kinds of tasks and selected a set of 98 sufficiently dissimilar ones for further analysis, using methods of explorative and confirmatory factor analysis. The emerging model contains seven different kinds of factors or abilities, which can be arranged in a rhomboid-like figure.

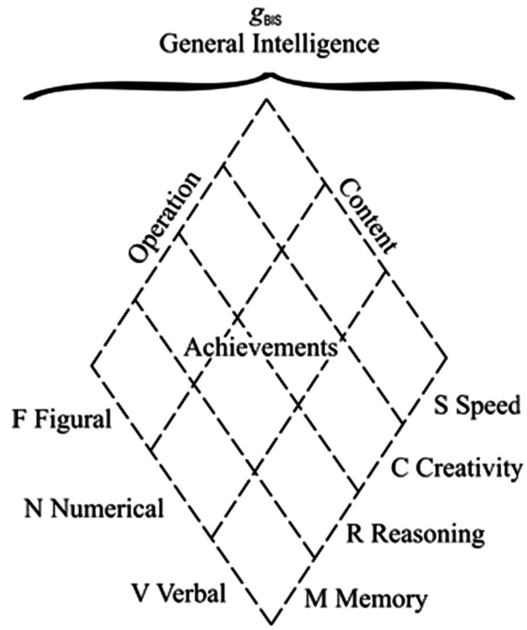

FIGURE 2 The BIS model (cf. Jäger et al., 2006).

Figure 2 shows that these seven abilities can be divided into two modalities, namely operations and contents, both of which contain additional special abilities. The BIS model states that, at the lowest level, achievements always rely on two factors, one for each modality. This is similar to Guilford's ideas about operations and contents, but the selection of sub-abilities is more parsimonious (see below), and there is no third modality (productions in the SOI model). The operational factors in the BIS are:

- Reasoning ability (R) – the processing of complex information in tasks that are not immediately solvable, but rather require the subject to establish diverse relations and use exact formal-logical reasoning about relevant problem solving information;
- Processing speed (S) – working speed, ease of perception, and concentration capacity, which are decisive in solving low-difficulty tasks;
- Memory (M) – active storage in short-term memory and recognition or reproduction of verbal, numerical, and figural material; and
- Creativity (C) – fluid, flexible, and original production of ideas, requiring diverse information, wealth of imagination and ability to see many different sides, variations, reasons, and possibilities in problem-oriented—not purely imaginative—solutions.

On the content side of the BIS, there are three factors:

- Figural ability (F) – visually based thinking; the ability to clearly understand relations and plasticity

in figural material, forms, shapes, patterns in tasks, appertaining to different operations;
- Numerical ability (N) – number-based thinking; it plays a role in all sets of tasks for different operations linked to numerical material; and
- Verbal ability (V) – language-based thinking; it plays a role in all sets of tasks for different operations linked to language.

Indeed, all factors are always involved in every achievement, but most of them to a far lesser degree than the two factors meeting in the particular achievement facet in question. The bimodality aspect therefore states, for instance, that the overall measure of the factor creativity can be split into three achievements: creative performance with figural material, with numerical material, and with verbal material. General intelligence (BIS *g*) is finally seen as the integral of all these measured abilities, rendering more obvious the hierarchical nature of the BIS model. For a more detailed overview on the BIS-model, see Jäger (1982) or Bucik and Neubauer (1996), who also provide a historical account of model development.

Reasoning ability, processing speed, and memory can be measured with tasks generally well-known, such as number series or (figural, as well as verbal) analogies for reasoning ability, digit-symbol tasks for processing speed or simple story memorizing for memory. With creativity, on the other hand, there is less consensus on what constitutes a valid measure of this construct, so a brief description of some exemplary creativity tasks used in the BIS test appears necessary. According to the definition of creativity in the BIS model, these tasks measure the ability to produce problem-oriented solutions, and hence emphasize fluency. In addition, some of the tasks can also be scored for originality. Figure 3 provides three exemplary BIS creativity tasks, one for each content domain. Altogether, there are 12 creativity tasks in the current version of the BIS-HB. It is worth of note that because of economical reasons (test length) they are all administered with a time limit, which stresses the aspect of creating problem-oriented solutions. Thus, the measurement of creativity with the BIS test is limited to fluency and originality. In practice, it is difficult to measure other creative aspects, like flexibility, or richness of imagination, which are central to the idea of divergent thinking, as it is formulated in other cognitive theories of intelligence and creativity (e.g., Carroll, 1993; Guilford, 1967), with the BIS-HB creativity tasks. Although this is a limitation to the assessment of creativity, the BIS model, nevertheless, explicitly lists creativity as an important component of an overarching intelligence construct, and with the BIS test, creativity is measured with a wide array of different tasks.

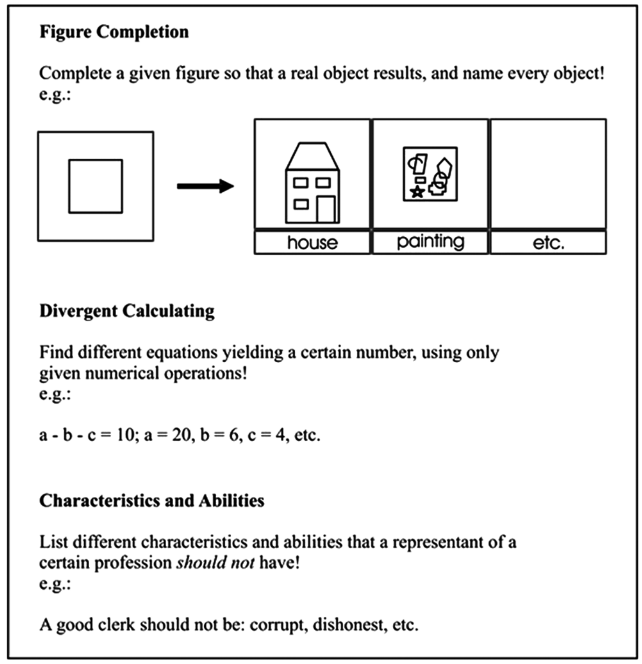

FIGURE 3 Three exemplary BIS creativity tasks (modified from the test).

RESEARCH QUESTIONS

In this article, two main goals were pursued. The first one is to use a multilevel model that will consider between-classroom variation at cluster level 2, as well as between-students variation at level 1 (see above). The second one is to operationalize creativity within the BIS model, leading to the assessment of creativity through the application of tasks belonging to the three different content categories figural, numerical, and verbal. In doing so, it was expected that a more complete picture of the role of creativity in predicting school performance can be attained because there is no over-reliance on one particular creativity aspect, e.g., verbal creativity.

The following two substantive research questions are of interest in this study:

1. How good is the predictive power of creativity for scholastic achievement, assessed with the multilevel approach? In other words, is creativity, as measured with tests of figural, numerical, and verbal content, strongly connected with school performance, and if so, is the relationship similar to that of reasoning ability and school performance?
2. Is creativity a better predictor in some units of analysis (classrooms) than in others? This question is concerned with possible differences in the relationship between creativity and school performance among the higher units of analysis and,

therefore, the multilevel aspect of this study is further tested. Some teachers may value creative answers and problem solving behavior in their students more than others, and thus reward students with better grades for showing creativity.

METHOD

Participants

The data analyzed in this study are part of the standardization sample of the Berlin Structure of Intelligence Test for Youth: Assessment of Talent and Giftedness (BIS-HB; Jäger et al., 2006). They consist of a total of 1,133 students grouped within 60 classrooms. The educational system in the Federal Republic of Germany selects students according to their cognitive abilities after grade 4 and assigns them to schools with different degrees of academic difficulty, with the *Gymnasium* (top track) being the most demanding school type and the *Real*- and *Hauptschule* (middle and low track, respectively) less difficult, respectively. The sample of this study consists of students attending all levels of the educational system. Because the BIS-HB was designed to provide a reliable measure for the identification of profoundly gifted students, a significant number of students visiting schools with special curricula for fostering giftedness were also included in the standardization sample. The corresponding school type is called *Gymnasium select* (gifted track) in this study. Table 1 gives an overview of the data structure. It is apparent that students attending the top track and the gifted track represent the majority of the sample. The latter school form is also visited by more boys than girls, which is a consequence of the fact that many of these select schools have an emphasis on natural sciences, which seem to attract boys more often than girls. In addition, gifted boys seem to attract the attention of parents and teachers more often than girls, and their talents are consequently identified and fostered more often. Overall, though, the proportions of girls and boys are not too discrepant, with girls representing 46% of the total sample. The students attended grades 7 to 10, and their mean age was 14.5 years, with a standard deviation of 1.07 years. The smallest classroom size was 6 students, and the largest was 30 students. The average classroom size was 19 students.

TABLE 1
Distribution of the Sample Data Across the Different School Types

School Type	*n*	*Girls*	*Boys*	*No. of Classrooms j*	*Percentage n*	*Range n_j*
Low track	181	86	95	11	16.0	8–20
Middle track	180	94	86	8	15.9	11–30
Top track	433	225	208	20	38.2	16–28
Gifted track	339	118	221	21	29.9	6–25
Total	1133	523	610	60	100.0	6–30

Measures

Students supplied data on their scholastic achievement, expressed through the grade point average (GPA). The mean GPA for the whole sample was 2.69, with a standard deviation of .66 and a range from 1.00 to 4.58. Unlike in the United States, lower GPAs in the German school system indicate higher achievement, and a GPA of 1.00 is the best possible average a student can obtain. Furthermore, the students had scores on the BIS-HB scales of creativity and reasoning ability. Although scores on the other BIS-HB scales were also available, in the present study the focus was on creativity and reasoning ability as the main competing constructs for predicting scholastic achievement, and, therefore, the other scales were not taken into consideration. The reliability of the creativity and reasoning scores were $\alpha = .81$ and $\alpha = .93$, respectively (both coefficients taken from the BIS-HB manual). Although reasoning was measured with greater precision, an α of .81 for creativity shows that the creativity tasks of the BIS-HB can be scored with sufficient reliability.

All measures (GPA, creativity, and reasoning ability) were normalized in order to circumvent any problems resulting from scores not being normally distributed across the sample.

RESULTS

The predicted criterion used in this analysis was student GPA. To show the merits of a multilevel over a single-level analysis, the variance partition coefficient (VPC) was computed. The VPC quantifies the extent to which the individual scores in the same group resemble each other, as compared to those from individuals in the other groups and is computed as $\sigma^2_{u0}/(\sigma^2_{u0} + \sigma^2_e)$, where σ^2_{u0} is the between-group variance (in a multilevel model, this is the second level), and σ^2_e is the individual variance (level 1 in a multilevel model). It can, therefore, be interpreted handily as the proportion of the total residual variation attributable to between-group differences (Goldstein, 2003). In the present context, that refers to differences in overall scholastic achievement between groups.

The between-students variance in the singlelevel model amounted to .997, with a standard error of .042.[3]

[3]In the following, the standard errors of the parameter estimates are in brackets.

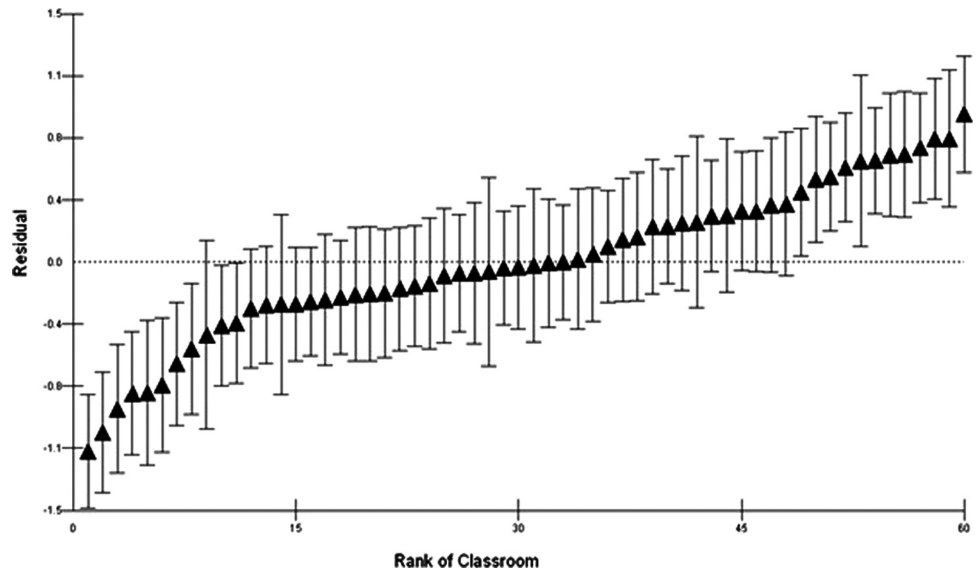

FIGURE 4 GPA residuals–confidence intervals.

The multilevel model yielded a level 1 variance of .742 (.032), and a level 2 variance of .246 (.053). The VPC was, therefore, .249, indicating that about 25% of the total variance in GPA may be due to differences between classrooms. The application of a single-level analysis would not take into account these structural differences between classrooms. A multilevel analysis, on the other hand, serves to do just that and, hence, is the obvious choice in this case. Figure 4 is a caterpillar plot of the 60 classrooms residuals, ordered by rank. Looking at the confidence intervals around the residuals, one can see that a total of 23 classrooms at the lower and upper end of the plot possess confidence intervals showing no overlap with zero. Because these residuals represent classroom distances from the overall mean GPA predicted by the fixed intercept only, it makes sense to conclude that these classrooms differ significantly from the mean at the 5% level.

As explanatory variables, the following predictors were considered in the full model: (normalized) creativity and reasoning ability, the interaction of creativity and reasoning ability, gender, and school type. The model can thus be written as:

$$y_{ij} = \sum_{h=0}^{4} \beta_h x_{hij} + \sum_{h=5}^{7} \beta_h x_{hj} + \sum_{h=0}^{3} u_{hj} x_{hij} + \sum_{h=0}^{2} e_{hij} x_{hij}$$

In this model, i refers to student and j refers to school. The subscript 0 always refers to the regression constant (equal to 1), the subscript 1 refers to creativity, 2 is for reasoning ability, 3 is for the interaction term, and 4 is for gender, which is translated into the difference between boys and girls, with girls being the reference category. The subscripts 5 through 7 are for the three school level defined variables listed in Table 2. Here, the school type *Hauptschule* (the low track) is always the reference category. The explanatory variables defined at the level of the individual student are, therefore, contained in the first summation, and the second summation refers to the explanatory variables defined at the classroom level. The random part of the model defining the variation at the classroom level is expressed through the third summation, and the random part defining the variation at the student level is in the fourth summation.

The complete model specified above was estimated using the software MLwiN 2.02 (Rasbash, Browne, Healy, Cameron, & Charlton, 2005), and the method of Iterative Generalized Least Squares (IGLS) was employed for parameter estimation. The results of the regression analysis are shown in Table 2. The deviance of the total model was 2660.578. This was found to be a significantly better fit than other models omitting any of the included or adding any further variables yielded. In the fixed part of the model, Table 2 shows the mean effects of the jointly fitted explanatory variables specified above. The effects of creativity and reasoning ability are both highly significant, as is the effect for the interaction term (at the 5% level), but the effect of reasoning ability is largest. In general, girls outperform boys, and there are differences between the school types. To fully understand the meaning of these contextual effects, one can look at the combined effects of gender and school type. Girls attending the low track are the reference group (because girls were coded with 0, and boys with 1), and it was found that boys in the low track do .297 standard deviations worse than girls in the low track. All other factors held constant, the grades of girls attending the middle track are

TABLE 2
Multilevel Analysis of GPA

Fixed Part	*Estimate*	*SE*
Intercept	−.232*	.107
Creativity	−.203**	.035
Reasoning	−.368**	.044
Creativity × reasoning	−.051*	.028
Boys – girls	.297**	.048
Middle – low track	.381**	.124
Top – low track	.210*	.117
Gifted – low track	−.069	.139

Covariance Matrix	*Intercept*	*Creativity*	*Reasoning*	*C × R*
Random–between classrooms				
Intercept	.081** (.023)			
Creativity	.003 (.012)	.021† (.013)		
Reasoning ability	.023† (.014)	−.013 (.012)	.026† (.017)	
Creativity × reasoning	−.014 (.011)	.010 (0.07)	−.010 (.008)	.009 (.008)
Random–between students				
Intercept	.541** (.036)			
Creativity	−.002 (.015)	−.011 (.027)		
Reasoning	.026† (.017)	−.027 (.024)	.058* (.033)	

† $p < .06$; * $p < .05$; ** $p < .01$.

.381 standard deviations worse than those of the reference groups, girls in the top track score .210 standard deviations worse, and the grades of girls attending the gifted track are .069 standard deviations better. However, although these results suggest that students in the lower track to receive better grades, this is misleading because the mean differences in ability must still be taken into account. Looking only at the results as reported, then, it becomes clear that a typical member of the reference category (i.e., a girl attending a school in the low track) would receive worse grades if she attended a school in one of the higher tracks. The mean difference in reasoning ability alone between students in the gifted and in the low track is about 2 standard deviations, which corresponds to 30 IQ points. For creativity, the difference between these two school types is about 1.3, corresponding to approximately 20 IQ points. Hence, these differences in ability need to be added to the parameter estimates when a meaningful comparison between school types is to be made.

A look at Figure 5 shows that when predicting GPA with creativity (Panel A) and reasoning ability (Panel B) only, respectively, there is significant variation in the intercepts and slopes among the classrooms. A single-level analysis would totally ignore these differences and estimate a single regression line. Closer inspection of the slopes for the different classrooms reveals that for creativity, there is no systematic trend for school type. Among the classrooms with the largest negative slope values (where higher ability indicates better grades), there are representants of the school types of the middle, top, and gifted track, and the only differences between them are in their intercepts, where there is a trend according to the rank-order of the school types, with students attending the gifted track doing better than their counterparts in the top, middle, or low track, respectively. However, there are even a few classrooms where higher creativity may lead to worse school grades, as can be seen by the positive or near zero slopes in Panel A of Figure 5. These classrooms can be located in the top, as well as in the middle, track, so there seems to be no trend according to school type. For reasoning ability, there are, in general, some large differences in the slope values for the different classrooms, but everywhere the same pattern holds: Brighter students receive better marks. Interestingly, though, the slope values for classrooms in the gifted track are smallest, although the intake of ability is highest. This shows that, although these students receive rather good marks in general, the relationship between ability and achievement is not very close.

One can also look at the level 1 and level 2 contributions of the predictors creativity and reasoning ability to the GPA variance. The level 1 contribution to the variance is $\sigma^2_{e0} + 2\sigma_{e01}x + \sigma^2_{e1}x^2$, and changing the *e*s to *u*s gives the level 2 contribution. In both cases, the variance function was calculated for models incorporating only the explanatory variables creativity and reasoning ability, respectively. Panel A of Figure 6 shows that for creativity (dots) the variance is maximal at a value of about

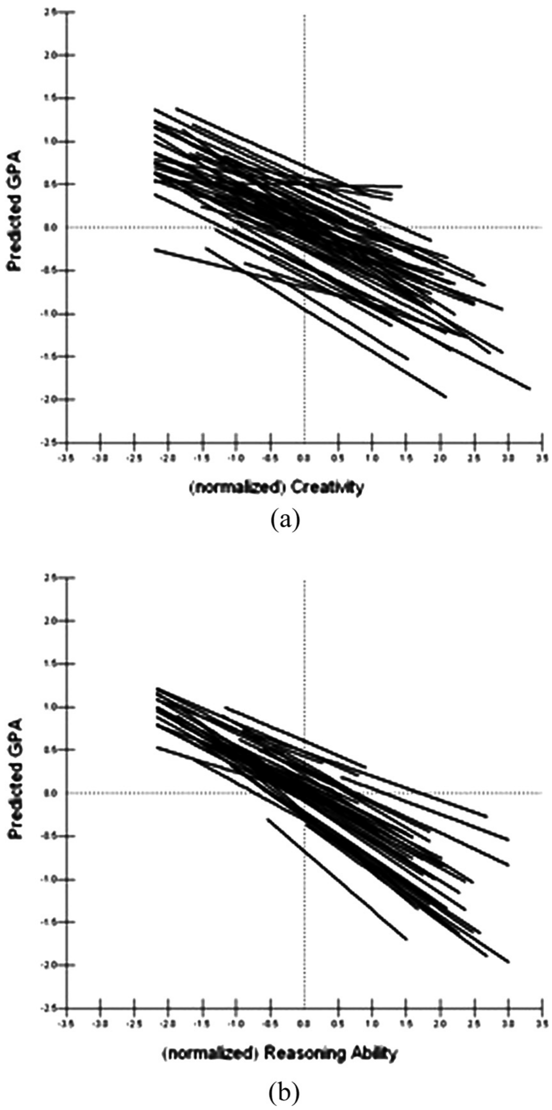

FIGURE 5 Creativity and reasoning ability against predicted GPA.

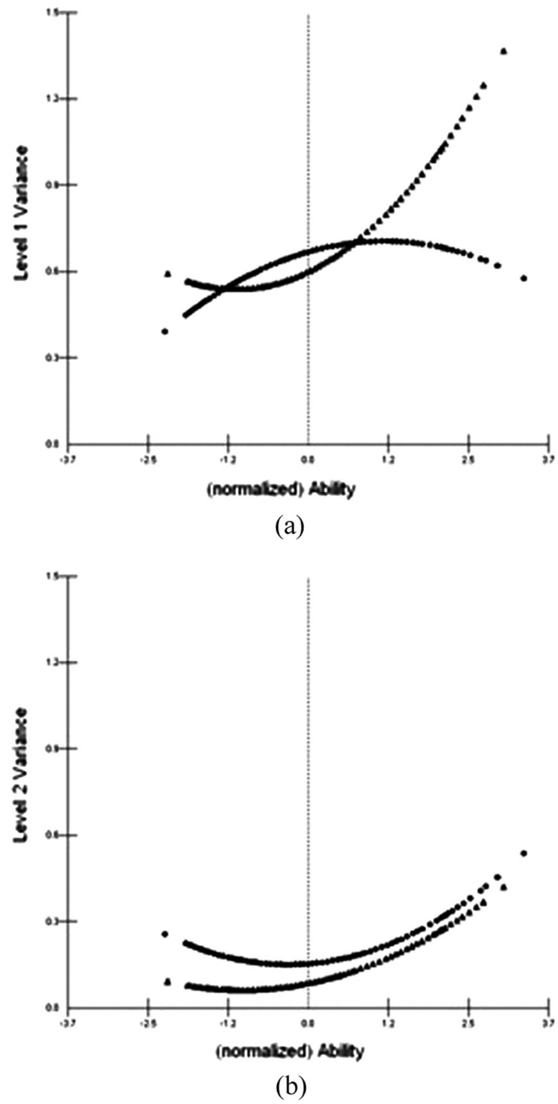

FIGURE 6 Variance functions of creativity and reasoning ability at levels 1 (Panel A) and 2 (Panel B); dots represent creativity, diamonds represent reasoning ability.

0.7 standard deviations above the mean, yet it is lower for extreme values of creativity. Above a certain creativity threshold, school performance does not seem to vary very much. However, for reasoning ability (diamonds), a different picture emerges. Here, the variance in obtained GPA rises dramatically with increasing ability, indicating that there is much more variation in scholastic achievement among students with higher ability.

Panel B shows the level 2 variance. The graphs of the functions for creativity (dots again) and reasoning ability (diamonds) are very similar at this level. Naturally, the between-classroom variance is lower than the between-students variance. However, it is interesting to observe that on both levels, a similar variance function exists for reasoning ability, whereas for creativity, there are completely different functions. At the classroom level, creativity behaves almost exactly like reasoning ability, with only an additive shift in the amount of observed variance, indicating more variability in GPA as a function of creativity, especially at the extremes.

DISCUSSION

One purpose of this study was to test the possibility that creativity is, in fact, related to scholastic achievement. Also, the role of creativity at school was to be compared with that of reasoning ability, and the multilevel character of the analyses conducted should allow for a more complete investigation of the impact of hierarchically clustering the respondents into aggregate groups and, hence, of the variable relationships between the different predictors and GPA among classrooms. The sample of

1,133 students ensured that a representative blend of students covering the entire ability spectrum was present in the data set. Sample sizes of this magnitude and beyond are necessary for the multilevel analyses conducted in this study. With a smaller sample of fewer students and consequently fewer classrooms, these analyses would not have been feasible because the standard errors of the estimates would have been too large (cf. Maas & Hox, 2005).

The results obtained within the multilevel framework suggest a rather strong effect for reasoning ability and, to a somewhat lesser degree, also for creativity when predicting GPA. The interaction term of these two predictors attained significance as well, but nowhere near the impact the single constructs reached. There is considerable between-classroom variation for both single predictors, which is also visualized in Figures 5 and 6. The full utility of the multilevel analysis becomes obvious when the estimates are compared with those yielded by an OLS method, where the cluster structure of the data is ignored. The OLS estimates for the effects of creativity and reasoning ability vary considerably from the IGLS estimates. For creativity, the OLS estimate is −.193 (.030; vs. the IGLS-pendant of −.203 [.035]), and for reasoning ability the estimates are −.370 (.038; OLS) vs. −.368 (.044). As can be seen from Figure 6, there is higher variation in GPA for both predictors especially at the upper end of the ability distribution, which means that among classrooms with higher levels of creativity and reasoning ability, there will be larger differences in the predictive power of these two constructs. The OLS estimates for the effects of gender and school type also deviate from the IGLS estimates, of course, but they remain fixed effects, so no more attention is paid to the differences observed here. It is not very surprising to see that girls tend to perform better than boys, as this is a rather common finding (e.g., Deary, Strand, Smith, & Fernandes, 2007; Freeman, 2004; Halpern, 2000; Hosenfeld, Köller, & Baumert, 1999). This difference cannot be explained by ability differences, and other factors, like motivation and self-discipline (Duckworth & Seligman, 2006) play a major role here. However, this finding is not of primary interest in this study and, therefore, no possible explanations for it are offered here.

What is interesting is the fact that students (especially girls) attending the low track did so well in comparison with their peers attending the other school forms, even after taking into account the mean group differences in ability. Here, however, it needs to be acknowledged that there are no standardized achievement tests that are administered to all students at all schools. Results from the Programme for International Student Assessment (PISA) study for Germany do point out to achievement gaps between the school tracks (e.g., Baumert et al., 2003), but they are not mirrored in the grades because for each track, different curricula exist. In addition, schools have their own standards, and because the sample of schools (not classrooms!) is, even in this study, not very large, it may well be the case that schools in the low track included in the data collection are at the top of the achievement spectrum for this school type. Of course, it would have been nice to incorporate a third cluster level, namely schools, in the multilevel analyses, but the number of different schools that the classrooms were located in was simply too small for this. But an analysis of the contextual effects of school type was conducted in the present study, and the results are revealing in that regard (see above).

In sum, the application of a multilevel framework yielded a more fine-grained picture of the relationship between creativity, reasoning ability, and scholastic achievement. The predictive utility of creativity and reasoning ability was rather high, and it was comparable to the findings reported in other studies. However, it could be shown here that there can be quite different relationships between these variables in distinct classrooms, so the picture is by no means constant for all units within the educational system. Some teachers simply seem to reward creativity more than others do. Because creativity is such a highly-valued ability, it could be fruitful to look more closely at the teaching concepts of those teachers who seem to appreciate creativity more than their colleagues. Recall here that Westby and Dawson (1995) found that teachers often tend to devalue creative behaviors exhibited by their students, even if they valued the creativity construct highly. It may be hypothesized that these teachers do not like (or are not able to cope well with) the independence, autonomy, and, sometimes, nonconformity needed to produce creative output in their students too much when they address a whole class of often up to 30, or even more, children respectively adolescents. In what ways teaching styles differ would be an interesting question to investigate. This is not possible with the current data set, but it could be a design element that should be considered in future studies. Finally, it really would be interesting to see if the results obtained within the German school system can be replicated in other systems that do not group students according to ability, like the United States, for example.

REFERENCES

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society A*, *149*, 1–43.

Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U. et al. (2003). *PISA 2000—Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* [PISA 2000—A differentiated view on the states of the FRG]. Opladen, Germany: Leske & Budrich.

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.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, NY: Cambridge University Press.

Cline, V. B., Richards, J. M., Jr., & Needham, W. E. (1963). Creativity tests and achievement in high school science. *Journal of Applied Psychology*, *47*, 184–189.

Cortis, G. A. (1968). Predicting student performance in colleges of education. *British Journal of Educational Psychology*, *38*, 115–122.

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

Duckworth, A. L., & Seligman, M. E. P. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology*, *98*, 198–208.

Freeman, J. (2004). Cultural influences on gifted gender achievement. *High Ability Studies*, *15*, 7–23.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement*, *8*, 369–395.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Arnold.

Guilford, J. P. (1967). *The nature of human intelligence*. New York: McGraw-Hill.

Halpern, D. F. (2000). *Sex differences in cognitive abilities*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hocevar, D. (1981). Measurement of creativity: Review and critique. *Journal of Personality Assessment*, *45*, 450–464.

Hosenfeld, I., Köller, O., & Baumert, J. (1999). Why sex differences in mathematics achievement disappear in German secondary schools: A reanalysis of the German TIMSS-data. *Studies in Educational Evaluation*, *25*, 143–161.

Jäger, A. O. (1967). *Dimensionen der Intelligenz* [Dimensions of intelligence]. Göttingen, Germany: Hogrefe.

Jäger, A. O. (1982). Mehrmodale Klassifikation von Intelligenzleistungen. Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells [Multimodal classification of intelligence achievement: Experimentally controlled, further development of a descriptive intelligence structure model]. *Diagnostica*, *28*, 195–226.

Jäger, A. O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven [Structural research on 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. (2006). *Berliner Intelligenzstruktur-Test für Jugendliche: Begabungs- und Hochbegabungsdiagnostik (BIS-HB)*. [Berlin structure of intelligence test for youth: Assessment of talent and giftedness]. Göttingen, German: Hogrefe.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

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

MacKinnon, D. W. (1978). *In search of human effectiveness: Identifying and developing creativity*. Buffalo, NY: Creative Education Foundation.

Milgram, R. M., & Hong, E. (1993). Creative thinking and creative performance in adolescents as predictors of creative attainments in adults: A follow-up study after 18 years. *Roeper Review*, *15*, 135–139.

Niaz, M., de Nunez, G. S., & de Pineda, I. R. (2000). Academic performance of high school students as a function of mental capacity, cognitive style, mobility-fixity dimension, and creativity. *Journal of Creative Behavior*, *34*, 18–29.

Ones, D. S., Viswesvaran, C., & Dilchert, S. (2005). Cognitive ability in selection decisions. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence*. Thousand Oaks, CA: Sage Publications.

Rasbash, J., Browne, W., Healy, M., Cameron, B., & Charlton, C. (2005). *MLwiN 2.02* [Computer Software], Bristol, UK.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Rindermann, H., & Neubauer, A. C. (2004). Processing speed, intelligence, creativity, and school performance: Testing of causal hypotheses using structural equation models. *Intelligence*, *32*, 573–589.

Runco, M. A., Plucker, J. A., & Lim, W. (2000). Development and psychometric integrity of a measure of ideational behavior. *Creativity Research Journal*, *13*, 393–400.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 262–274.

Sternberg, R. J. (1999). *Handbook of creativity*. Cambridge, NY: Cambridge University Press.

Sternberg, R. J., & Lubart, T. I. (1999). The concept of creativity: Prospects and paradigms. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 3–15). Cambridge, NY: Cambridge University Press.

Süß, H.-M. (2001). Prädiktive Validität der Intelligenz im schulischen und außerschulischen Bereich [Predictive validity of intelligence in school-related and external domains]. In E. Stern & J. Guthke (Eds.), *Perspektiven der Intelligenzforschung [Perspectives on intelligence research]* (pp. 109–135). Lengerich, Germany: Pabst.

Torrance, E. P. (1966). *Torrance Tests of Creative Thinking*. Lexington, MA: Personnel Press.

Westby, E. L., & Dawson, V. L. (1995). Creativity: Asset or burden in the classroom? *Creativity Research Journal*, *8*, 1–10.

Williams, R. L. (2002). Creative performance in the classroom. *Inquiry: Critical Thinking Across the Disciplines*, *22*, 7–20.