Psychology Science Quarterly, Volume 50, 2008 (2), pp. 295-307

Something new in the garden: Assessing creativity in academic domains

ELENA L. GRIGORENKO[1], LINDA JARVIN[2], MEI TAN[1] & ROBERT J. STERNBERG[2]

Abstract

Although not generally included in the classroom curricula of the past, creativity and creative thinking are emerging as important skills that can and should be taught. While considering current research and definitions of these constructs, this paper proposes an understanding of the construct of creativity as a skill that can be actively and constructively nurtured within the context of the classroom. Presenting empirical work in which analytical assessment and creative assessment data are collected from 1120 students within the same test of reading comprehension, and the psychometric properties of each are evaluated and compared using item-response theory (IRT), we share results that reflect the differentiation of skills and variations in analytical and creative competencies. By this means, we also pose a way of developing proficiency scales for creativity, as well as ways to naturally integrate the encouragement of creativity in the classroom.

Key words: creativity, competency, proficiency, assessment, reading comprehension

[1] Elena L. Grigorenko, Child Study Center, Department of Psychology, Department of Epidemiology & Public Health, Yale University, 230 South Frontage Road, New Haven, CT 06519-1124; email: elena.grigorenko@yale.edu
[2] Tufts University

The concepts of creativity and creative thinking are springing up in the educational literature, more and more often, like the incremental signs of an emerging life form whose time has come and whose flower has blossomed. In the past, assessing creativity was associated primarily with the identification of talented and gifted students (TaG) (Kaufman & Baer, 2006). Similarly, not so many years ago, teaching for creativity, if it was on the agenda at all, was directed only toward educating TaG students (Treffinger & Isaksen, 2005). Now the situation is changing rapidly, with the tasks of teaching for creativity and creative thinking placed on the agendas of primary (Maker, Jo, & Muammar, in press), secondary (Mouchiroud & Bernoussi, in press), tertiary (Ogunleye, 2006), and further (Ogunleye, 2006) education. Contrary to past ideas of what can be nurtured academically in the minds of students, we consider now, in research and in the literature, skills that were once not considered to be of a common garden variety.

This new focus of a number of educational systems on teaching for creativity [for examples and detailed illustrations, see (Grigorenko & Tan, in press)] poses many new tasks both for researchers who study creativity and for educators who teach for it. These tasks are numerous and important, but two of them are of particular importance, specifically: (1) defining creativity and creative thinking within the context of changing curricula worldwide; and (2) developing recommendations on how creativity and creative thinking can be (a) taught and (b) assessed within the context of education in school.

The literature on creativity is voluminous, with most of it devoted to definitions of concepts of creativity and creative thinking, and to correspondent theories. The body of empirical research on creativity is substantially smaller. And the body of educational and pedagogical literature with specific research-supported suggestions on teaching for creativity is smaller yet. Here, we contribute by providing a succinct overview of these literatures to define the context for our empirical work, whose purpose is to illustrate how the environments and mechanisms for the nurturing of creativity and creative thinking may already exist within the world of formal education. We argue that now creativity should be regarded not as a delicate hot-house species, but rather, as a strong and hardy plant, whose roots are already firmly established. Driven by the demands of the labor market, we need to develop the tools and techniques to help it thrive; however, these tools and techniques, from our point of view, should be introduced and promoted within a particular domain, rather than as abstract skills.

Creativity as a skill

Broadly speaking, one can distinguish ***C***reativity from ***c***reativity. This distinction implicitly and explicitly separates two kinds of creativity. One encompasses the high-impact creative activities that result in the creation of new cultural (e.g., music, literature, film), scientific (e.g., scientific discoveries), and social (e.g., religious, political, and societal) products – big-C ***C***reativity. The other is everyday creativity, which results in finding solutions to small local problems (e.g., creating a new outfit, offering a new interpretation of a situation, coming up with a new dish to eat, finding a novel solution to a problem, developing an incremental advancement at a work place) – small-c ***c***reativity. There are multiple fundamental accounts presenting definitions and theories of creativity [for reviews, see Kaufman & Sternberg (2006)]; we contextualize our work within the larger arena of creativity research

by defining *c*reativity as a skill used in approaching novel tasks, problems, and situations. In this interpretation, *c*reativity is a skill that can be taught and developed to the point of competency (e.g., along with specific levels of proficiency). Also, from this point of view, *c*reativity can be taught within one or more academic domains.

The level of *c*reativity exerted within a given domain is determined both by the demonstrated mastery of the skill of creativity (i.e., creativity proficiency level) as well as by the demonstrated mastery of the content of the domain (i.e., the knowledge of literature or math). Thus, the development of *c*reativity as a skill requires an analysis of the levels of creative skill that already exist within particular academic domains, so that a scale for acquiring competency or mastering levels of proficiency in creative thinking can be established. Just as educators develop memory and analytical skills while teaching content knowledge within traditional educational frameworks, they can also foster creative skills while teaching mathematics, science, language arts, or any other content area. The trick is to develop tools that can calibrate the growth of the skill of creativity to the level of competency. How can such calibrations for a proficiency scale be developed?

In fact, there is a large body of educational literature that can be of assistance here. Specifically, a massive amount of information has been accumulated within large-scale cross-national/societal studies of various competencies [e.g., Trends in International Mathematics and Science Study TIMSS, Programme for International Student Assessment (PISA), International Adult Literacy Survey (IALS), US National Assessment of Adult Literacy (NAAL) and National Adult Literacy Survey (NALS)]. This information can be of help in developing approaches to teaching and assessing creativity within educational contexts. These assessments have generated a great deal of data and commentaries pertaining to the theoretical representations of various skill-based competencies and the translation of these ideas into assessment items. It is then possible empirically to validate these items as they are mapped onto proficiency scales corresponding to various competencies (Beaton & Allen, 1992; Messick, Beaton, & Lord, 1983). Capitalizing on this literature, we explore empirically the application of proficiency scales for assessing creativity as a skill measurable within specific academic domains.

In these attempts to link performance on items of particular levels of difficulty with specific levels of competency (or specific proficiency levels), each item/task is associated with a particular point on a proficiency scale; individuals functioning at that corresponding level of competency can address the associated item/task successfully. In this context, competency (proficiency) is quantitatively distributed, along with the difficulty level of the items/tasks, so that the closer the match between the two distributions, the higher the probability of a successful outcome in approaching a particular item/task. Describing this relationship, Kirsch (Kirsch, 2001) uses the following analogy: "The relationship between task difficulty and individual proficiency is much like the high-jump event in track and field, in which an athlete tries to jump over a bar that is placed at increasing heights. Each high jumper has a height at which he or she is proficient—that is, the jumper can clear the bar at that height with a high probability of success, and can clear the bar at lower heights almost every time. When the bar is higher than the athlete's level of proficiency, however, it is expected that the athlete will be unable to clear the bar consistently" (p. 26).

The continuity of a proficiency scale assumes that every subsequent level of proficiency can be reached only when the previous level of proficiency is mastered. This idea has been utilized in so-called cumulative scales in psychometrics (Gutman, 1950). This continuity,

however, does not assume that all items/tasks sampling from a level higher than the current level of proficiency will be failed; on the contrary, some of them can be and will be approached successfully, but the performance on such items/tasks is expected to be inconsistent. Thus, assuming a 5-level proficiency scale, a person whose proficiency is at level 3 is expected to approach items/tasks whose levels of difficulty match levels 1 and 2 of the proficiency scale with a success rate of at least 80%. Of course, they may also deal with some level 3 items/tasks successfully, but the corresponding rate of success will be lower. Such an approach underscores the importance of developing creative skills within a particular domain, assuming, at least for pedagogical purposes, its relevant domain unidimensionality.

Below we describe our own empirical work with respect to the possibility of encouraging the growth of creative skills and assessing them within the existing environment of the classroom, while teaching a particular academic subject. A proficiency scale can be developed when we begin to isolate what we intend to measure in individuals and then plot its variations. Below we exemplify such an attempt to develop creativity in the domain of language arts. The purpose of this work is not to apply or validate any particular theory of creativity. On the contrary, we are focused here on an infrastructure that permits the building of a proficiency scale that might be instructive for the pedagogical applications of teaching language arts and developing creative thinking skills.

Methods

Participants

Participants in this study were 1120 teenagers, enrolled in grades 7-12 (aged 13-18 years); ~40% of the participants were girls and ~60% were boys.

Materials

Each participant in the study received 5 printed paragraphs, each of which had 1-2 traditional and 1-2 creative comprehension questions. Paragraphs sampled both narrative and expository texts and differed in their levels of difficulty, based on the anticipated reading levels for the participants' grades. There were 16 paragraphs altogether, and the collections of paragraphs for each grade partially overlapped, so that there were paragraphs that were received in common by students from adjacent grades. The paragraphs for each grade were chosen in such a way that the distribution of students' ability in each grade was covered as well as possible. Thus, we implemented a so-called anchor-test design (Petersen, Kolen, & Hoover, 1989): although the items were tailored to particular grade levels, adjacent grades had items in common (an anchor set), guaranteeing stability for the equating procedure. The selection of overlapping paragraphs was guided by recommendations in the literature (Cook, Eignor, & Taft, 1988; Kolen & Brennan, 1995; Notenboom & Reitsma, 2003).

As indicated above, each paragraph was followed by two types of questions assessing comprehension: traditional (e.g., understanding, memory, and/or analysis-based) and creative comprehension questions. Traditional questions required the respondents to repeat memorized facts or to analyze the content of the reading passage, while creative questions required

a novel response. There were 23 traditional and 22 creative questions. Both the analytical and creative questions were presented in mixed formats, utilizing multiple-choice and short response formats. Short response format questions were scored as "right" or "wrong," on a 0 to 1 scale. Below is an example of a paragraph and its corresponding questions.

"Currently NASA is devoting serious study to human space-flight beyond the Earth's orbit. At this time, the most likely target is Mars. When the possibility of Mars expeditions was first seriously considered in 1989, NASA estimated that the necessary industrialization of space would take 30 years and cost a staggering $450 billion. This was an utterly unrealistic goal. Reconsidered in 1998, the costs of a two-and-a-half year mission to Mars have now fallen to as low as $20 billion, with the expectation that such an expedition could take place early in the 21st century. This drastic change is explained by the fact that all technological advancements are immediately incorporated into the space industry. Of all the known planets, Mars is thought to be the most like Earth. The astronauts may not have to change their sleep cycles, because a Martian day is only 37 minutes longer than an Earth day, but while awake they'll notice plenty of differences between the two planets. Here is what it will be like on Mars. *Length of year:* 687 days. Mars has four seasons, but they last twice as long as the Earth's. *Temperatures:* They range from minus 220°F to plus 81°F. *Winds*: They can reach several hundred miles per hour, whipping up dust storms that last for weeks. *Atmosphere*: Very thin and mostly carbon dioxide (93.5%). *Gravity*: If you weigh 200 pounds on Earth, you'll weigh only 78 pounds on the surface of Mars. *Surface features:* Mars has the tallest mountain and the deepest canyon in the solar system. *Sky*: Pink during the day, blue at sunrise and sunset. Mars has two moons, Deimos and Phobos.

Traditional (analytical) question:
What types of companies contribute the most to the industrialization of space?
Companies that
(a) issue patents for scientific discoveries
(b) apply technological advancements to space exploration
(c) develop anti-aging medication
(d) insure space-crafts

Creative question:
Suppose you are a realistic painter capturing Martian scenery at sunset. Imagine and describe, in writing, your painting."

Procedure

Students received 5 paragraphs each with ~4 comprehension questions during one of their regular class periods. The assessment was administered at schools and scored by research staff at Yale University.

Data analyses

To investigate the comparative properties of the traditional and creative subscales of this comprehension assessment, we used BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). This software permits the utilization of dichotomous scores (e.g., right-wrong and forced-choice), applying to them item-response theory (IRT), in which each item is modeled separately and the log-odds of the probability of a correct response for examinees is conceived as a function of a latent variable θ (ability) and one or more additional item parameters. Correspondingly, models are defined through the number of parameters utilized. Specifically, the one-parameter logistic (1PL) model includes one estimable item parameter, the item difficulty parameter, β. The two-parameter logistic (2PL) model contains β and an additional parameter, an item discrimination parameter, α. The three-parameter logistic (3PL) includes β, α, and a guessing parameter, γ. Thus, the models vary in the number of parameters required to explain the probability of answering an item correctly at every level of examinee ability (i.e., at every level of the underlying θ). A representation of the probability of answering correctly utilizing the IRT parameters is captured by a stretched-out S-shape curve that increases in slope until it reaches a point of inflection (β), at which point the rate of increase slows down and changes to a decrease. Accordingly, item characteristic curves (ICC) generated by the 1PL-model tend to look alike and run parallel to each other, differing only in their relative locations on the ability or trait continuum. The 2PL ICCs differ not only in relative placement along the ability continuum, but also in the degree to which the probability of answering an item correctly increases steeply with an increase in level of ability (θ). ICCs differ in their slopes (α), so that some are characterized by steep slopes and others by a flat slope before hitting the point of inflection. ICCs with steep slopes indicate that corresponding items have more power to discriminate between people who are above or below that particular level of ability (θ), where the point of inflection for a given ICC lies. Finally, the 3PL ICCs differ in the lower asymptote (bottom left corner of the S shape) of their curve, capturing the guessing (γ) parameter. This parameter indicates the degree to which individuals with a lower level of ability (θ) will answer an item (or guess) correctly.

Typically, in IRT analyses, the best fitting model (1, 2, or 3PL model) is selected first by comparing the fitness of indices, then by analyzing the ICCs, and, finally, by assessing the properties of the scale as a whole.

Results

Results of the model comparison are shown in Table 1. The preferred model, for both the traditional and creative items, is the 2PL model, including the difficulty and discrimination parameters. This conclusion is supported by both the analysis of the -2 Log Likelihood indicators, which show only negligible improvement in the fit for the 3PL model for the traditional items, and the worsening of the fit of the creative items for the 3PL model, as compared with the 2PL model. Similarly, indices of reliability are the best for the 2PL model for both the traditional (.845) and creative (.872) scales. Consequently, the item analysis was carried out within the context of the 2PL model.

Table 1:
Results of Model-Fitting for 1PL, 2PL, and 3PL Models

	Traditional Items		Creative Items	
	-2LL	RI	-2LL	RI
1 PL Model	11925.733	.770	10239.664	.808
2 PL Model	11792.866	.845	10119.753	.872
Incremental -2LL (1PL vs. 2PL)	132.867		119.911	
3 PL Model	11791.802	.839	10160.834	.849
Incremental -2LL (2PL vs. 3PL)	1.064		-41.081	

Notes: -2LL: -2 Log Likelihood; RI: Reliability Index

Patterns of the observed data generated by each item were compared to the patterns predicted using the 2PL model and the associated estimated parameters; the degree of deviance from the predicted pattern was estimated by means of likelihood-ratio χ^2 values. Among the traditional items (n = 23), there were 8 misfit items (3 with p-values of <.05 and 5 with p-values of <.001); among the creative items (n = 22), there were 6 misfit items (all with p-values of <.001).

Table 2 presents the parameter estimates for item difficulty and discrimination (β and α) for both the traditional and creative items. As shown in this table, the administered items cover a range of ability and discriminate participants at different levels of ability as well. Of note here is that the creative items appear to be sampling from different points on the ability scale as effectively as the traditional items do. The content analyses of both the traditional and creative items, and an examination of their levels of difficulty and discrimination, resulted in the generation of Table 3. Table 3 (A and B) captures the proficiency levels needed for dealing effectively with traditional (memory-analytical) and creative items. Thus, these data illustrate that, similar to traditional tasks, creative comprehension items can be constructed so that they differ in levels of difficulty and discrimination power, and that these items can be mapped on a scale, capturing a stepwise progression in the development of creative skills.

Figure 1 (A and B) presents the information function and standard error for both the traditional and creative comprehension scales. For the scale comprising traditional items, the amount of information peaks at an ability level lower than 0 (i.e., lower than an average level of ability, θ = -.375); for the scale including creative items, the information level peaks at a point greater than 0 (i.e., higher than an average level of ability, θ = .625). Of note also is that the creative scale contains more information than the traditional scale. At the extremes of both scales, the information is lower and the value of the standard error is higher; nevertheless, the average range of scores within our sample of 1120 teenagers is well covered. In other words, creative items in this set of comprehension questions appeared to be more difficult. Creative items discriminated students at higher levels of skills (slightly above average) more effectively; the traditional items provided more information at a level of skill slightly below average. In general, the levels of information and the precision of items for both scales were higher in the middle region of the scales than at the scales' tails.

Table 2:
Item Parameter Estimates for the 2PL Logistic Model for Traditional and Creative Comprehension Items

Traditional Items	Item Difficulty (*β, sd*)	Item Discrimination (*α, sd*)	Creative Items	Item Difficulty (*β, sd*)	Item Discrimination (*α, sd*)
1	.174 (.037)	-2.089 (.486)	1	.403 (.057)	-.290 (.105)
2	.641 (.074)	-.362 (.074)	2	.527 (.088)	.898 (.175)
3	.489 (.085)	.198 (.145)	3	.610 (.078)	1.052 (.122)
4	.709 (.110)	-.157 (.105)	4	.664 (.093)	.741 (.114)
5	.768 (.125)	-.462 (.105)	5	.852 (.124)	.199 (.091)
6	.833 (.134)	.755 (.128)	6	.854 (.120)	.444 (.094)
7	.707 (.096)	-.382 (.098)	7	1.176 (.196)	1.110 (.112)
8	.894 (.132)	.504 (.096)	8	1.038 (.160)	.808 (.093)
9	1.113 (.191)	-.763 (.098)	9	.905 (.135)	.703 (.100)
10	.434 (.079)	.915 (.219)	10	1.064 (.159)	.834 (.096)
11	.583 (.101)	-2.804 (.403)	11	1.699 (.335)	.646 (.068)
12	.513 (.082)	.988 (.154)	12	.595 (.097)	.886 (.138)
13	.649 (.090)	-.525 (.094)	13	.448 (.074)	.586 (.133)
14	.602 (.088)	.828 (.125)	14	.467 (.078)	.769 (.147)
15	.296 (.069)	3.821 (.856)	15	.282 (.059)	1.585 (.346)
16	.833 (.154)	-1.772 (.221)	16	1.211 (.402)	.518 (.137)
17	.488 (.082)	.823 (.150)	17	.446 (.116)	.282 (.264)
18	.820 (.263)	.402 (.237)	18	1.045 (.368)	.309 (.154)
19	.950 (.300)	.051 (.218)	19	.616 (.144)	.226 (.211)
20	.598 (.167)	.470 (.304)	20	.838 (.220)	.615 (.188)
21	.805 (.208)	-.590 (.192)	21	.476 (.119)	.102 (.259)
22	1.239 (.460)	-1.201 (.228)	22	1.061 (.281)	.085 (.156)
23	.533 (.135)	.454 (.257)			
Scale	-.039 (1.310)	.681 (.245)	Scale	.596 (.413)	.785 (.344)

Finally, we calculated the correlation between the two scales, the one comprised of traditional items, and the other creative items. The correlation was .519 ($p < .001$, uncorrected for attenuation), indicating that, although both scales capture the comprehension of written material, they are not redundant and share only ~25% of their variance.

Table 3:
Proficiency Scale for Comprehension Items

A. Traditional Items

Levels of Proficiency	Appropriate Tasks
1	Tasks at this level require the reader to grasp the central idea of the paragraph or to locate a single piece of information that is identical to or synonymous with the information given in the question.
2	Tasks at this level require the reader to integrate two or more pieces of information or to compare or contrast easily available details in the text.
3	Tasks at this level require the reader to encode, compare, or contrast information that is readily available and/or to make low-level inferences and provide basic explanations of their conclusions.
4	This level of tasks requires readers to encode, compare, or contrast on a number of levels, extract the meaning from the text, and to carry out higher-level inference.
5	These tasks call for high-level inferences and the use of specialized background knowledge. Readers are expected to carry out complex analyses of the text based on multi-level encoding, contrasting, and comparing of various pieces of information. Readers must then judge the presented data and, finally, infer and justify their conclusions.

B. Creative Items

Levels of Proficiency	Appropriate Tasks
1	Tasks at this level ask the reader to elaborate on the details presented in the paragraph/poem; the reader is expected to come up with words (concepts, ideas) that are slight modifications of those used in the paragraph.
2	Tasks at this level require the reader to provide a novel variation of the usage of a word (concept, idea, and so on) presented in the paragraph/poem by either changing the word (concept, idea) or the context in which it is presented.
3	These tasks require the reader to suppose that some aspects of the situation presented in the paragraph/poem change; the reader, then, is expected to construct a meaningful modification of the situation.
4	This level of tasks introduces a significant amount of ambiguity by asking the reader to come up with missing pieces of information or extend the paragraph/poem creatively.
5	These tasks ask the reader to devise, create, or originate novel situations (character developments, plot lines) going beyond what was read.

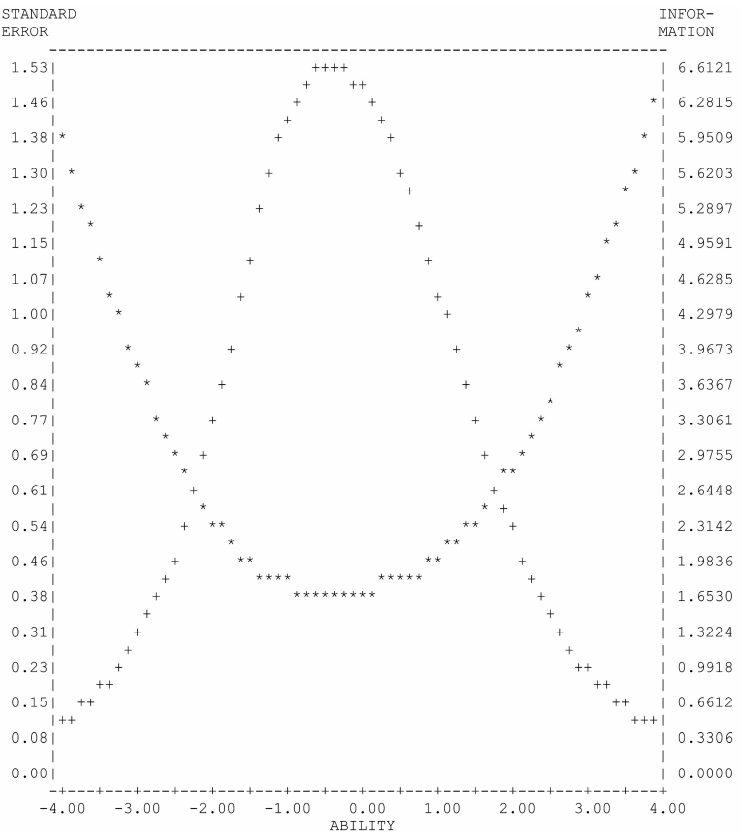

Figure 1A:
Traditional and Creative Scales Information Curves:
Scale Comprised of Traditional Items

Discussion

The main point of this work was to demonstrate how teaching and assessing creativity can be naturally integrated into teaching and assessing domain-specific knowledge, at least in the domain of reading. In this particular case, we generated two types of items capturing different skills, which can be developed while practicing the comprehension of printed materials. These items can be mapped onto scales capturing levels of skill acquisition (i.e., levels of proficiency). When such proficiency scales are constructed, each level can be decomposed into a set of teachable and exercisable skills that can be nurtured and encouraged while acquiring specific content areas within particular academic domains. When exercised multiple times across many content areas and domains, these skills might be expected to crystallize and become transferable. However, further research is needed to substantiate this hypothesis.

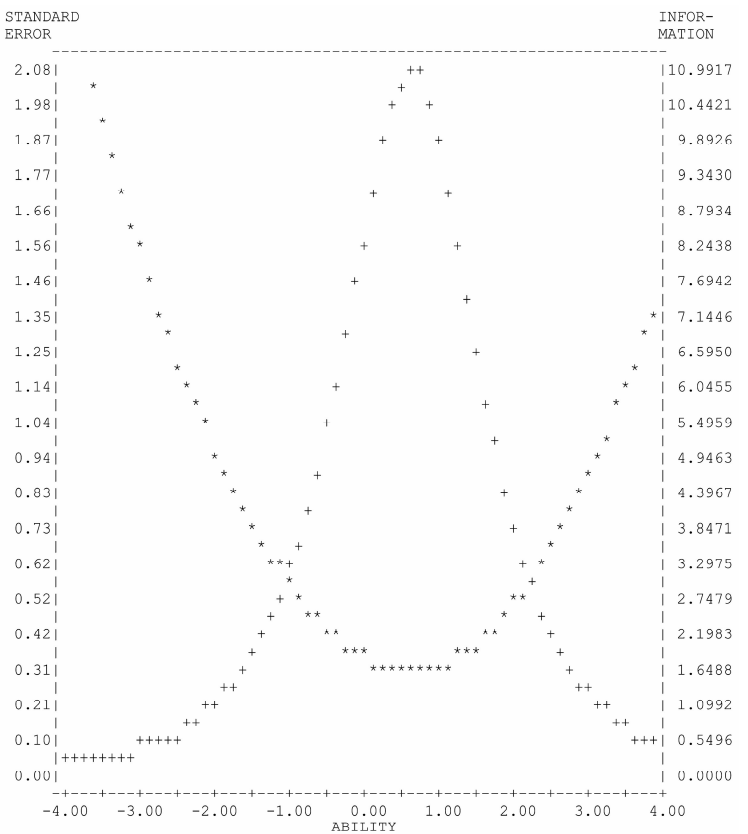

Figure 1B:
Traditional and Creative Scales Information Curves:
Scale Comprised of Creative Items

To illustrate these general points, we utilized the domain of language arts. However, it has been demonstrated that similar proficiency scales can be built for developing analytical and creative skills in a variety of academic domains, for example, mathematics and science (Jarvin, Grigorenko, & Sternberg, in preparation). In fact, the learning of new content in all areas lends itself to creative exercises. For example, when learning new math applications, students can be asked to imagine new, futuristic, or fantastical uses of such an application. Facts learned in a science exercise can be employed in visual creations (pictures and illustrations, e.g., Draw a scenario in which some typical behaviors of a sea anemone are included); in verbal creations (sentences or stories, e.g., Write a sentence or story about sea anemones employing some facts that you have learned); as well as in numerical creations (mathematical equations or relationships, e.g., Come up with some equations that describe quantitative relationships that are important in the life cycle or in the environment of a sea anemone). Formulating such tasks, or items, could constitute fertile soil for growing, nurturing and

ultimately assessing creative thinking. It should also be noted that items that are designed to assess creative rather than traditional skills demonstrate psychometric properties that are comparable to those of conventional items.

In sum, this work contributes to multiple literatures. First, it adds to the literature demonstrating that creativity and creative thinking can be assessed reliably by a variety of means, ranging from standardized tests (Kaufman & Baer, 2006) to consensual assessment techniques which involve experts evaluating student products such as short stories and poems (Baer, 2003; Kaufman, Lee, Baer, & Lee, 2007) or, in this case, responses to items calling for creativity. Second, it contributes to the literatures on assessing creativity in educational settings [e.g., (Brandau et al., 2007; Zaragoza, 2003)]. Here we have attempted to demonstrate that creative skills can be assessed within an existing educational curriculum without engaging specific standardized tests of creativity but rather in the context of assessing domain-specific achievement. Third, it contributes to the educational literature on teaching thinking skills and developing competencies such as creative thinking (Treffinger, Selby, & Isaksen, in press), assuming an easy "move" from proficiency scales to teaching creativity skills in language arts in an incremental fashion. These contributions are relevant to modern teaching and schooling approaches when the demand for creativity in knowledge economies is rapidly growing and calls educational systems to prepare participants for labor markets who master creative skills (Grigorenko & Tan, in press).

In sum, although illustrative in purpose, the study and its findings contribute to the growing literatures on teaching and assessing creativity in school settings.

Acknowledgments

The preparation of this assay was supported by federal funding administered through Grant REC-9979843 from the National Science Foundation and Grant No. R206R000001 under the Javits Act Program as administered by the Institute of Educational Science, U. S. Department of Education. Correspondence regarding this article should be sent to Elena L. Grigorenko at 230 South Frontage Road, New Haven, CT 06519, USA (elena.grigorenko@yale.edu).

References

Baer, J. (2003). The impact of the core knowledge curriculum on creativity. *Creativity Research Journal, 15*, 297-300.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17*, 191–204.

Brandau, H., Daghofer, F., Hollerer, L., Kaschnitz, W., Kellner, K., Kirchmair, G., et al. (2007). The relationship between creativity, teacher ratings on behavior, age, and gender in pupils from seven to ten years. *Journal of Creative Behavior, 41*, 91-113.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement, 25*, 31-45.

Grigorenko, E. L., & Tan, M. (in press). Teaching creativity as a demand-led competency. In O. S. Tan, D. M. McInerney, A. D. Liem & A.-G. Tan (Eds.), *What the West can learn from the*

East: Asian perspectives on the psychology of learning and motivation (Vol. 7). Greenwich, CT: Information Age Press (IAP).

Gutman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer (Ed.), *Measurement and Prediction* (pp. 60-90). Princeton, NJ: Princeton University Press.

Jarvin, L., Grigorenko, E. L., & Sternberg, R. J. (in preparation). Teaching and assessing for creativity in the elementary school classroom. In R. A. Beghetto & J. C. Kaufman (Eds.), *Finding creativity in the classroom: Between chaos and conformity*. New York, NY: Cambridge University Press.

Kaufman, J. C., & Baer, J. (2006). Intelligent testing with Torrance. *Creativity Research Journal, 18*, 99-102.

Kaufman, J. C., Lee, J., Baer, J., & Lee, S. (2007). Captions, consistency, creativity, and the consensual assessment technique: New evidence of reliability. *Thinking Skills and Creativity, 2*, 96-106.

Kaufman, J. C., & Sternberg, R. J. (Eds.). (2006). *The international handbook of creativity*. New York, NY, US: Cambridge University Press.

Kirsch, I. (2001). *The International Adult Literacy Survey (IALS): Understanding what was measured* (No. RR-01-25). Princeton, NJ: ETS.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York, NY: Springer-Verlag.

Maker, C. J., Jo, S., & Muammar, O. M. (in press). Development of creativity: The influence of varying levels of implementation of the DISCOVER curriculum model, a non-traditional pedagogical approach. *Learning and Individual Differences*.

Messick, S., Beaton, A., & Lord, F. (1983). *National assessment of educational progress reconsidered: A new design for a new era* (No. NAEP Report 83–1). Princeton, NJ: National Assessment of Educational Progress.

Mouchiroud, C., & Bernoussi, A. (in press). An empirical study of the construct validity of social creativity. *Learning and Individual Differences*.

Notenboom, A., & Reitsma, P. (2003). Investigating the dimensions of spelling ability. *Educational and Psychological Measurement, 63*, 1039-1059.

Ogunleye, J. (2006). A review and analysis of assessment objectives of academic and vocational qualifications in English further education, with particular reference to creativity. *Journal of Education and Work, 19*, 95-104.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (pp. 221-262). New York, NY: American Council of Education.

Treffinger, D. J., & Isaksen, S. G. (2005). Creative problem solving: The history, development, and implications for gifted education and talent development. *Gifted Child Quarterly, 49*(342-353).

Treffinger, D. J., Selby, E. C., & Isaksen, S. G. (in press). Understanding individual problem-solving style: A key to learning and applying creative problem solving. *Learning and Individual Differences*.

Zaragoza, F. A. M. (2003). Psychometric characteristics of CREA (creative intelligence). A study with Spanish and Argentinian samples. *Revista Iberoamericana de Diagnostico y Evaluacion Psicologica, 16*, 71-83.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *Multi-group IRT analysis and test maintenance for binary items*. Chicago, IL: Scientific Software International.