Creativity Evaluation through Latent Semantic Analysis

Eve A. Forster (eve.forster@utoronto.ca)
University of Toronto Scarborough, Department of Psychology
1265 Military Trail, Toronto, ON M1C 1A4 Canada

Kevin N. Dunbar (dunbar@utsc.utoronto.ca)
University of Toronto Scarborough, Department of Psychology
1265 Military Trail, Toronto, ON M1C 1A4 Canada

Abstract

The Uses of Objects Task is a widely used creativity test. The test is usually scored by humans, which introduces subjectivity and individual variance into creativity scores. Here, we present a new computational method for scoring creativity: Latent Semantic Analysis (LSA), a tool used to measure semantic distance between words. 33 participants provided creative uses for 20 separate objects. We compared both human judges and LSA scores and found that LSA methods produced a better model of the underlying semantic originality of responses than traditional measures.

Keywords: latent semantic analysis; creativity; natural language processing

Creativity research has had a short, but interesting history in the Cognitive Sciences. Beginning with Guilford's presidential address to the American Psychological Association in 1950, researchers have sought ways of discovering creative individuals (Guilford, 1947) that provide an alternative to the long and laborious methods used by the Gestalt psychologists.

Gestalt research methods often consisted of extensive interviews with creative individuals (such as Albert Einstein; Wertheimer, 1945), which offered fascinating accounts of creative moments of some creative people, but was not amenable to discovering vast numbers of creative individuals. Guilford advocated the use of the psychometric approach for this purpose, and over the subsequent decade a number of new creativity tests were devised. By the mid-1960s, the Guilford Alternate Uses test (Guilford, 1967) and the Torrance Test of Creative Thinking (Torrance, 1998) were widely used measures of creativity across the world. On the surface, these tests were ideal; they were easy to administer and quick to score: the more responses made, the more creative the individual.

Researchers soon discovered that there was more to creativity than number of responses. New measures were proposed that counted number of categories employed, and measured response elaboration and novelty. These measures brought new problems to creativity assessment: they are inherently subjective, have large variances in coding, and take a considerable amount of time to score.

Sternberg and Lubart (1992) write that creativity is a function of six factors: intelligence, knowledge, thinking style, personality, motivation and environmental context. Each of these can fluctuate from day to day due to changes in a person's internal and external environment, causing different subscales such as drawing or writing to fluctuate in different ways. The psychometric approach accommodates these multiple factors by administering a large battery of short tests, to encapsulate all aspects of creativity. Most of the tests require people to generate or manipulate a large number of ideas. Guilford and Hoepfner (1966) provide 57 tasks that ask participants to do things such as grouping and regrouping objects according to common properties, listing the consequences of unlikely situations, and the UoO Task.

While it is almost 60 years since Guilford's original address to APA, the scoring of creativity tasks still remains problematic. One way of addressing these problems would be to use an automated measurement tool that uses underlying semantic knowledge to assess creativity. Although initially developed to model language learning, Latent Semantic Analysis (LSA) has since proven itself as a flexible tool with a variety of sophisticated uses. In this article we test the hypothesis that it can be used as a consistent and completely automated creativity scoring method. Here, we use LSA to score creativity of participants who perform the Uses of Objects task.

The Uses of Objects Task

The UoO Task is a psychometric test that requires people to generate multiple, original uses for a given object. Quantitative scores count the number of ideas (a measure of fluency) or number of words per response (elaboration), and subjective scores judge creativity and category switching. The task is widely used (Dunbar, 2008; Guilford, 1967; Guilford & Hoepfner, 1966; Hudson, 1968; Torrance, 1998). Scoring of the UoO task can be easily automated, but doing so strips the responses of their meaning. The only two scoring options at present are meaningful but subjective and slow, or consistent and fast but meaningless. The ideal scoring method should be meaningful, consistent and completely automated; such a method may be devised by combining a traditional elaboration measure with a novel assessment of originality.

The need for consistent measurement

Popular scoring systems such as the Torrance Test of Creative Thinking (Torrance, 1998) require a trained person to assess productions, but this option is not always practical. Such assessment is slow, expensive, and subjective (and

therefore potentially biased). Creativity ratings can vary substantially from one judge to another, and depend on the individual interpretations and knowledge of judges. Judges may use particular heuristics to save time (awarding high ratings to longer sentences or unusual words), and may adjust their rating methods over time as they learn which responses are common. This results in a highly inconsistent assessment, which is why such testing is still controversial.

It is difficult to include creative thinking goals in school curricula, as proposed scoring methods are not consistent enough for evaluation in results-oriented systems such as those in North America and the UK. If used in hiring, testers must worry about consistency to avoid discrimination claims. In both industry and education, efficiency, consistency and cost are extremely important factors, and ordinarily one of those factors would suffer for the sake of others. The proposed method allows all of the above factors to be maximized. It allows for a consistent and meaningful assessment of creativity without sacrificing efficiency.

Subjective measures can be used, requiring independent judges to score the creativity or originality of a response or count the number of changes in the category of use between responses (referred to as the category switch score). Psychometric assessment methods are less subjective, scoring people by number of responses or reaction times (Guilford & Hoepfner, 1966).

Scoring creativity by elaboration and originality

Despite disagreements at finer grains of detail, there is a general consensus that originality and practicality are two of the most important components of creativity (Runco & Pritzker, 1999). In other words, creativity depends on generation of diverse ideas and their subsequent pruning.

Runco and Pritzker (1999) suggest that refining and elaborating on a particular idea improves the quality of the idea, so level of elaboration may be an indicator of quality. It may also mean that the idea is more practical or applicable; one requirement of elaboration is that the idea be elaborate-able in the first place, which is only possible if there is a tangible association between the object and its use. Because the responses will not be seen or edited by a human, there is also the danger that some ideas may not be legitimate uses at all; highly elaborated statements may correspond to more appropriate responses.

A word count can be used to approximate elaboration, but an originality judgment requires much greater complexity. It must involve an understanding of the meaning of a response, or failing that, the capability to determine the conceptual distances between two responses or a response and its prompt. This may be possible with the use of LSA.

A New Method of Originality Assessment: LSA

LSA is a model of language learning in which word meaning is inferred from statistical analyses of large batches of text. Word relatedness is inferred according to which words often co-occur, and lack of relatedness according to which words are rarely together. By reducing the noise in the dataset, singular, local relationships between words are amplified into consistent, global word associations.

The process operates by Singular Value Decomposition of a large lexical co-occurrence matrix. The frequency of each word within each passage of text is stored in the matrix, with each row corresponding to a word and each column to a passage. The frequencies are scaled by an inverse entropy measure to reflect the probabilities of those frequency-context associations, estimating the importance of the word to the overall meaning of the passage. The matrix is then decomposed into 3 matrices ($X = W \times S \times P^T$), where S is a diagonal matrix of scaling coefficients or *singular values*. To reduce the dimensionality of the original matrix to rank k (typically, $k = 300$), all but the highest k singular values are deleted. The matrix is then reconstructed with the new singular value matrix, transforming the original matrix of word-document associations into a matrix relating words to abstract contexts. The dimensionality reduction reduces the noise in the data, allowing the latent relationships between words to be revealed. Words are represented as vectors in this high dimensional space, and word similarity can be calculated by taking the cosine of the angle between vectors. See Landauer and Dumais (1997) for more details.

Applications LSA has been shown to behave like humans with respect to category membership (Laham, 1997), word-word and passage-word priming (Landauer, Foltz & Laham, 1998), vocabulary growth and performance on the TOEFL synonym test (Landauer & Dumais, 1997), and metaphor comprehension (Kintsch, 2000). It has been repurposed as a method of indexing and retrieving documents (Latent Semantic Indexing; Deerwester et al., 1990), grading essays (Rehder et al., 1998), and distinguishing between humorous and non-humorous texts (Mihalcea & Strapparava, 2006). Variations on the system have also been used to uncover the latent relationships between chemicals and between genes (Hull et al., 2001; Kim, Park & Drake, 2007).

LSA's successful employment in essay evaluation and language modeling—as well as a wide assortment of additional knowledge representation mechanisms—suggests great potential for use in creativity evaluation. The following study pits LSA against a large group of human judges, who—although potentially inconsistent in the smaller numbers used for traditional studies—represent a more consistent assessment due to the higher numbers used. To determine whether use of LSA would be an adequate replacement for human assessment, two groups of people were instructed to produce either creative or uncreative uses of objects and their scores according to LSA and human assessments were compared. LSA was also evaluated for internal consistency and its ability to predict human judgments, and the human judges were assessed in their susceptibility to bias. Because of the potential subjectivity of the judges' ratings, it was important to determine how and by which dimensions the judges were assessing creativity; originality and practicality were measured independently to isolate those dimensions, and several LSA measures were calculated to model suspected evaluation strategies.

Method

Participants

Creative Responders 61 participants began the study,[1] and of these 33 participants completed all 20 objects (21 women and 12 men). Online participants had a mean age of 18.24 (*SD* = .60) and in-lab participants a mean age of 18.75 (*SD* = .89). All were first year psychology students at the University of Toronto at Scarborough, taking part in return for course credit. 24 took part in the lab (8 of which completed the experiment), in a bare room devoid of inspiration. 37 took part through an on-line interface outside the lab (which 25 completed). Only those who completed all 20 objects were included in the analyses. To determine any bias due to exclusion, analyses were conducted on the 33 included participants and subsequently with the addition of a subset (33) of the excluded participants. There were no differences between the results of the two analyses. The authors elected to report data from completed participants.[2]

Common Use Responders 28 participants (23 female, 3 male, 2 declining to answer) were specifically instructed to provide common uses for all 40 objects in the study. This group had two purposes: not only could their responses be used as a non-creative control for comparison to the creative responders, but their responses were also used in the calculation of the Common Use LSA score (described in the Data Analysis section below). All participated through the online interface, and all but one resided in either the United States of America or Canada. All spoke English as their first language. Their mean age was 35.25 (*SD* = 11.44).

Creative Response Judges 26 participants (7 male, 19 female) judged the individual responses provided by the Creative Responders. 14 judges judged responses by their creativity, 7 by their originality, and 5 by their practicality. The mean age of judges was 21.09 (*SD* = 2.15).

Apparatus

All interaction with participants (except Creative Response Judges) occurred through a web browser. In the lab, data was collected on a 20-inch 2.4GHz Apple iMac. Creative Response Judges typed their responses into an Excel spreadsheet rather than the online system.

Stimuli

40 objects were chosen to maximize variety in word frequency (in the British National Corpus; Burnard, 2000), distinctiveness (using feature production norms in McRae et al., 2005) and homogeneity. Six object categories were included (e.g. *furniture* and *vehicles*). Stimuli were divided into two groups of 20, and participants were randomly assigned to view either group. 16 participants were assigned to Stimulus Group (SG) 1 and 17 to SG 2.

Procedure

Participants were encouraged to generate as many original uses as they could for each object. They were told that they could manipulate the objects any way they wanted, including taking the objects apart, and that they could fall back on obvious uses if they were unable to generate an original use. They were instructed to fixate on an image of crosshairs, which was displayed for 2 seconds, followed by an object slide. An image of each object was displayed on a white background, with the word for the object below it.[3] To the right of the object was a text box where participants could enter their responses. They were given 2 minutes for each object, after which time the text box disappeared and they were directed to continue to the next object. In-lab participants performed the task in a room that was bare except for a large desk, coat rack and filing cabinet. Online participants were instructed to perform the task wherever they were comfortable.

The Common Use Judges were asked to state the most common way they would use each object. They were instructed to give only a single answer for each object. Responses were coded into broad general categories by the author (e.g. "play with my daughter", "to play in", and "play in it" were all coded as "play"). Of the most common use category for each object, the least specific expression (e.g. "play") was recorded as the object's most common use.

The Creative Response Judges were then asked to judge each response by the creative and common use responders on a scale from 1 to 3 (where 1 = uncreative/ unoriginal/ impractical, 2 = somewhat creative/etc. and 3 = very creative/etc.). The judges were given 1 hour to evaluate as many as possible. 7 judges were instructed to evaluate responses (for objects from SG 1) by creativity; 4 others evaluated the originality of responses and 2 evaluated practicality. The remaining judges (7 creativity, 3 originality and 3 practicality) rated the responses to objects from SG 2. The order of objects was counterbalanced.

Data Analysis

Traditional Response Scoring

Several scores were calculated for each participant's responses. *Elaboration* was measured by counting the

[1] An additional 112 participants were recruited through the Internet. All took part in an online version of the task. From this group only 13 participants completed all 20 tasks, and due to the high level of attrition, these participants were not included in the analyses. Results were the same regardless of whether these subjects were included.

[2] The lab room was kept purposefully bare, which may have adversely affected the creativity of the in-lab participants. This may in turn have resulted in a higher number of responses left blank, and thus a lower number of in-lab participants completing all 20 objects.

[3] In some cases, objects had to be depicted as a pile of "stuff," but no containers were used. Object categories were depicted using several pictures, each a possible object within the group.

number of words in each response, and *fluency* was measured by counting the number of responses for each object. Scores by the Creative Response Judges were averaged for each response's *subjective creativity score*, as well as an *originality* and *practicality* score.

Developing Scores using Latent Semantic Analysis

Similarity measures were calculated using 300 factors (the most typical number used in LSA), with a corpus consisting of the expected reading experience of a 1st year college student (*general-reading-up-to-the-first-year-in-college*).

Responses were spellchecked against a list of words in the corpus. Very common words such as "the" and "and" were eliminated, as well as common phrases such as "I would use it to…." Not only did this mean that noisy responses such as "use it to throw food" and "throw food" were treated as the same response, but it also allowed the two responses to be handled simultaneously, reducing total processing time. Several scores were calculated with the help of LSA:

Category Switch The similarity scores between successive response pairs were averaged for each object. This was intended to be similar to a category switch score.

Variety The similarity scores between every single pair of responses for an object were also averaged, as a measure of the variety of responses produced by each person.

Originality For each response, 25 responses produced by other people for the same object were selected at random and the similarities between the participant's response and each of the other 25 responses were averaged.[4] This provided a measure of originality compared to responses of others.

Pruned Originality Of the originality scores calculated previously, the three highest scoring responses were averaged for each object, to simulate an originality score if the participant were asked to report their three most creative uses. While this did not necessarily prune ideas the way the participant would, it allows an estimation of how original they would be if they were more selective.

Common Use Each response was compared to the most common use of the corresponding object (collected previously from Common Use Judges).

Each of these calculations produced a similarity score, which was then subtracted from 1.0 to produce a corresponding novelty score.

Results

Consistency of Creativity Scores

To determine how consistently each type of score was able to gauge the creativity of the people generating ideas (independent of the object being shown to them), the reliability of each score was calculated with each object as a subscale (as shown in Table 1). Additional reliability calculations were performed for the human scores with each individual judge as a subscale, to determine how much consensus there was between human judges (Table 2). SG 2 reliabilities were very similar to those for SG 1, so reliability was reported only for that group.

Table 1: Consistency of each scale over 20 objects.

Measure	Cronbach's α	Cases	N
Quantitative			
Elaboration	.93	16	20
Fluency	.91	16	20
LSA			
Category Switch	.86	16	20
Variety	.92	16	20
Originality	.90	16	20
Pruned Originality	.85	16	20
Common Use	.87	16	20
Human Judges			
Creativity	.90	16	20
Originality	.87	16	20
Practicality	.85	16	20

Table 2: Consistency of each human judge.

Measure	Cronbach's α	Cases	N
Creativity (SG 1)	.61	16	7
Originality (SG 1)	.84	16	4
Practicality (SG 2)	.94	17	3

Scores by most of the individual judges correlated significantly with the average subjective creativity score ($.58 < rs < .92$, $ps < .05$; mean $r = .62$, $SD = .35$), although four did not ($.01 < rs < .34$, $ps > .13$). These four were removed from subsequent calculations of the average subjective creativity score. This did not affect the significance of any of the following statistics.

Relationships among Traditional and LSA Scores

Significant correlations between LSA and traditional scores are shown in Table 3. Although there was no connection between elaboration and fluency averaged across participants, there was a significant correlation when scores were examined for each object ($r(852) = .13$, $p < .001$).

Table 3: Correlations between LSA scores and traditional measures.

	Com-mon	Pruned	Origin-ality	Switch	Variety
Fluency	.41**	.66**	.31*		
	46	46	46		
Elaboration		–.42**	–.46**	–.60**	–.63**
		46	46	46	46

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

[4] Scores composed of comparisons to 25 responses and scores compared to every single response produced (both for the *scotch tape* object) were very strongly correlated ($r = .98$, $p < .001$).

Validity of LSA Scores

Of the LSA scores, only the comparison to common use score had a significant correlation with the human-judged creativity score ($r(33) = .60$, $p < .001$). Both fluency ($r(33) = -.07$, $p > .60$) and elaboration ($r(33) = .22$, $p > .20$) were unrelated to the subjective creativity measure.

T-tests comparing the creativity responders in the lab to outside the lab were not significant for both human scores and the LSA comparison to common score. However, when online creativity responders were compared to online common use responders, the differences were equally significant for both scores (LSA: $t(51) = 10.79$, $p < .001$; Human: $t(51) = 11.12$, $p < .001$).

Four backwards regressions were performed, to determine which variables best approximated the subjective creativity scores. Potential factors in the regressions were: 1) the traditional quantitative measures, elaboration and fluency, 2) the LSA measures, 3) a combination of traditional quantitative measures and LSA measures, and 4) the subjective measures, originality and practicality.

The model generated in a backwards regression on the two traditional measures was non-significant.[5] The model generated using the LSA measures had greater significance ($F(2,43) = 14.66$, $p < .001$). Common use and pruned originality were most significant to the model, which accounted for 43% of the variation in creativity ($R^2_{adj} = .39$).

Combining the traditional and LSA measures improved the model further ($F(3,29) = 19.99$, $p < .001$), this time including the measures of elaboration, fluency and common use. This accounted for 67% of the variation ($R^2_{adj} = .64$).

Practicality was not significant to the model of creativity; thus, a regression model of subjective score included only originality as the main factor. This was the most significant model ($F(1,19) = 245.44$, $p < .001$), and accounted for 93% of the variance in the creativity scores ($R^2_{adj} = .92$).

Because average scores from approximately 3 judges is a generally acceptable number for more creativity studies, three additional score averages were calculated and used to model the average creativity score, to illustrate the kind of predictive power a randomly chosen group of 3 judges may have. An average score from the 3 judges with the highest correlation with average creativity accounted for 93% of the variance ($R^2_{adj} = .92$). The 3 with the lowest correlation, however, accounted for 6% ($R^2_{adj} = .002$) and a group composed of the highest, middle and lowest correlation with average creativity accounted for 44% ($R^2_{adj} = .43$).

Discussion

LSA scores (the common use score in particular) successfully predicted the average human creativity scores, and were capable of differentiating the uses generated by participants in the creative and common use conditions. The motivation of the responders (whether they were told to generate common or creative uses) had quite a strong effect on both human and LSA creativity scores. This motivation had similarly strong effects on both scores, which may imply that both humans and LSA are equally capable of differentiating creative from uncreative ideas. The prominence of the common use score as the central predictor to the creativity scores also suggests that the humans were assessing responses by comparing responses to the common uses for the objects.

Idea Generation

Fluency and elaboration of ideas correlated strongly with each other, but this was a tendency that disappeared when they were averaged into overall participant scores; the relationship existed within objects but was not characteristic of particular people. Surprisingly, elaboration was negatively related to the pruned, originality, category switch and variety scores, suggesting that people who gave longer responses also had less variety in their responses, did not make significant changes in the themes of their responses, and/or gave similar responses to those of other people. A likely explanation could be that when participants were spending time thinking through an elaborate answer, this hampered their ability to give additional responses of the same quality. It may also have prevented them from giving enough answers to generate a high variety and category switch score. This may have been the reason why these scores did not correlate with the subjective creativity score, and why comparison to a common use was more realistic for this particular style of evaluation. They may still have potential as creativity measures, but would correspond better to subjective category switch measures.

Idea Evaluation

Those who were asked to judge creativity most often reported using originality as a guide to their creativity judgments, which was confirmed by the regression model of creativity on subjective scores. A model including the LSA common use score and elaboration score was the best predictor of human creativity measures, suggesting that the strategies that judges used for deriving the "originality" and "creativity" scores involved comparing ideas to a self-generated prototypical use for the object in question. It is suspected that the contribution of elaboration to the model was indirectly related to practicality; uses that were more practical may have also been more elaborated.

The LSA assessment had no major advantage over human judges in its consistency over objects, but the low consensus between the human judges in their assessments suggested that the LSA measures had a great advantage due to standardization (in that the algorithm used was the same each time). This study used a much higher number of judges than are usually used in creativity studies (generating an average score from judges with the most agreement), in an effort to provide a creativity evaluation that may be more representative of overall human approval. The LSA scores

[5] It should be noted that a model including non-local participants was significant ($F(1,44) = 10.29$, $p < .005$), accounting for 19% of the variation in creativity ($R^2_{adj} = .17$). Elaboration was the only variable significant to the model.

were compared to a creativity score of high judge consensus, rather than a less-consistent score driven by the deviations of a minority. The decision to remove judges did not affect the primary results of the study, but it is argued that it improved the consistency of the human score and allowed for a better estimation of model performance.

A model of LSA measures was more successful in predicting creativity than a traditional scoring method. A combination of traditional and LSA scoring produced the best model of automated measures; it accounted for over two-thirds of the variation in creativity, which was twice the average performance of the judges in the study.

Summary

Measures such as fluency and elaboration may be simple to quantify, but they can sacrifice realism. This study allowed participants to respond with elaboration in a familiar and anonymous environment, and showed that a consistent measurement scheme can be possible with Latent Semantic Analysis. The success of this measurement technique was confirmed with a scale independently judged by humans, and shown to be a better approximation of human responses than traditional measures.

Acknowledgments

This research was funded by two grants from the University of Toronto to both the first and second authors.

References

Amabile, T.M., Conti, R., Coon, H., Lazenby, J. & Herron, M. (1996). Assessing the work environment for creativity. *Acad Manage J*, *39*, 1154–1184.

Burnard, L. (2000). *British national corpus user reference guide V2.0*. Oxford: Oxford University Computing Service. Retrieved fall 2007 from http:// natcorp.ox.ac.uk

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. & Harshman, R. (1990). Indexing by latent semantic analysis. *J Am Soc Inform Sci*, 41, 391–407.

Dunbar, K. (2008). Arts, education, the brain and language. *Learning, Arts, and the Brain. The Dana consortium report on Arts and Cognition*. New York: Dana Press.

Foltz, P.W., Laham, D. & Landauer, T.K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic J of Computer-Enhanced Learning*, 1(2).

Glenberg, A.M. & Robertson, D.A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *J Mem Lang*, *43*, 379–401.

Guilford, J.P. (1947). The discovery of aptitude and achievement variables. *Science, 106*, 279–282.

Guilford, J.P. (1967). *The Nature of Human Intelligence.* New York: McGraw-Hil1.

Guilford, J.P. & Hoepfner, R. (1966). Sixteen divergent-production abilities at the ninth-grade level. *Multivar Behav Res*, *1*(1), 43–66.

Helson, R. (1996). In search of the creative personality. *Creativity Res J*, *9*(4), 295–306.

Hudson, L. (1967). *Contrary imaginations: A psychological study of the English schoolboy*. Harmondsworth, England: Penguin Books.

Hull, R.D., Singh, S.B., Nachbar, R.B., Sheridan, R.P., Kearsley, S.K. & Fluder, E.M. (2001). Latent semantic structure indexing (LaSSI) for defining chemical similarity. *J Med Chem*, *44*, 1177–1184.

Hutchison, K. (2003). Is semantic priming due to association strength or featural overlap? A "micro-analytic" review. *Psychon B Rev*, *12*, 82–87.

Kim, H., Park, H., Drake, B.L. (2007). Extracting unrecognized gene relationships from the biomedical literature via matrix factorizations. *BMC Bioinf*, *8*(Suppl. 9), S6.

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychon B Rev*, *7*, 257–266.

Laham, D. (1997). Latent semantic analysis approaches to categorization. In M.G. Shafto & P. Langley (Eds.) *Proceedings of the 19th annual meeting of the Cognitive Science Society* (p. 979). Mawhwah, NJ: Erlbaum.

Landauer, T.K. & Dumais, S.T. (1997). A solution to Plato's problem: The Latent Semanctic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychol Rev*, *104*, 211–140.

Landauer, T.K., Foltz, P.W. & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Process*, *25*, 259–284.

McCrae, R.R. (1987). Creativity, divergent thinking, and openness to experience. *J Pers Soc Psychol*, *52*, 1258–1265.

McRae, K., Cree, G.S., Seidenberg, M.S., McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behav Res Meth Ins C*, *37*, 547–559.

Mihalcea, R. & Strapparava, C. (2006). Learning to laugh (automatically): Computational models for humor recognition. *Comput Intell*, *22*(2), 126–142.

Plucker, J.A. & Renzulli, J.S. (1999). Psychometric approaches to the study of human creativity. In R.J. Sternberg (Ed.), *Handbook of Creativity*. New York: Cambridge University Press.

Rehder, B., Schreiner, M.E., Wolfe, M.B.W., Laham, D., Landauer, T.K. & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Process*, *25*, 337–354.

Runco, M.A. Pritzker, S.R. (1999). *Encyclopedia of Creativity*. San Diego: Academic Press.

Sternberg, R.J. & Lubart, T.I. (1992). Creativity: Its nature and assessment. *School Psychol Int*, *13*(3), 243–253.

Torrance, E.P. (1998). *Torrance tests of creative thinking: Norms*. Bensenville, IL: Scholastic Testing Service.

Wertheimer, M. (1945). *Productive Thinking*. New York: New York: Harper.