2020

Is the Reliability of Objective Originality Scores Confounded by Elaboration?

Shannon Marie Maio
University of Denver

Is the Reliability of Objective Originality Scores Confounded by Elaboration?

A Thesis

Presented to

the Faculty of the Morgridge College of Education

University of Denver

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

---

by

Shannon M. Maio

June 2020

Advisor: Denis Dumas

Author: Shannon M. Maio
Title: Is the Reliability of Objective Originality Scores Confounded by Elaboration?
Advisor: Denis Dumas
Degree Date: June 2020

ABSTRACT

The increased use of text-mining models as a scoring mechanism for divergent thinking (DT) tasks has sparked concerns about the ways in which automated Originality scores may be influenced by other dimensions of DT, especially Elaboration. The debate centers around the question of whether too much variance in automated Originality scores is accounted for by the number of words a participant uses in a response (i.e., Elaboration), and, thus, how the influence of Elaboration can affect the reliability of Originality scores. Here, a partial correlation analysis, in conjunction with text-mining and psychometric modeling, is conducted to test the degree to which the reliability of Originality scores produced via a freely-available text-mining system is dependent on the variance explained by Elaboration. Findings reveal that, when modern methodological recommendations for text-mining Originality scoring are applied, the reliability of Originality scores estimated by the *GloVe 840B* text-mining system is not meaningfully confounded by Elaboration. I conclude that, even when the variance attributed to Elaboration is partialled out, this method is capable of providing reliable Originality scores.

Keywords: Divergent thinking, Originality, Elaboration, text-mining model, reliability

ACKNOWLEDGMENTS

These past two years at DU have afforded me an educational experience far above anything I could have hoped for. I was able to dive deep into my interests in psychometrics, pursue my own research, and even first author a contribution to the larger literature through, what I hope will soon be, a publication in the *Creativity Research Journal.* As I reflect back on these times at DU, I feel a tremendous sense of appreciation for several individuals. First, I would like to thank my advisor, mentor, and thesis director Dr. Denis Dumas for his constant support. He collaborated with me on research, pushed me to continually improve my writing, and helped me pursue my interest in psychometrics. Thanks to his support, I have always felt challenged and my education has skyrocketed. Next, I would like to thank Dr. Peter Organisciak, who has impacted much of my research and success through his collaboration on research, his willingness to offer support, and his contributions on my thesis committee. Also, Dr. Jesse Owen, who was willing to jump in as the thesis committee chair with such late notice and has been extremely helpful ever since. Further, Garrett Roberts, who so generously agreed to join my thesis committee and support me in my final weeks here at DU. Each of these individuals has played an important role in my graduate education and for that I am enormously grateful.

TABLE OF CONTENTS

CHAPTER ONE: INTRODUCTION

The study of human creativity has been of interest for decades (e.g., Guilford, 1950; Hudson, 1968; Torrance, 1972, 1995; Simonton, 2018). In the words of Mednick (1962), the process of creative thinking is "the forming of associative elements into new combinations which either meet specific requirements or are in some way useful. The more mutually remote the elements of the new combination, the more creative the process or solution" (p. 221). Under this definition, those who are creative have greater access to remote associative thoughts, that are also useful, which lead to their greater ability to form creative solutions. Similarly, Guilford (1956, 1968) tied the idea of *divergent production*— or the generation of multiple different solutions for a problem— to creative ability and emphasized the importance of divergent thinking for creative production. Somewhat more recently, Runco and Jaeger (2012) pointed to the *standard definition* of creativity, which recognizes Originality, or novelty as a core component of creativity. Both, the standard definition of creativity and Guilford's early conceptualization of divergent thinking are exemplified in much of the scholarly work concerning divergent thinking (DT) today because DT often leads to Originality.

Divergent Thinking as an Indicator of Creative Potential

Divergent thinking (DT) tasks seek to tap the human capacity to generate ideas in many different and unique directions (Runco, 1999; Kaufman et al., 2008; Acar & Runco, 2015). While idea generation is only one of the many processes that make up the

full creative process (Guilford, 1950; Reiter-Palmon, 2018), it is critical to creativity across domains. DT often leads to Originality (unique and uncommon ideas) and Originality is central to creativity (Torrance, 1995; Runco, 2007; Runco & Jaeger, 2012). For this reason, DT tasks have dominated creativity testing, proving to be useful and reliable estimates of human creative potential (Davis, 1989; Plucker, 1999; Acar & Runco, 2019). As such, DT tasks are today the most often utilized measures in the creativity research literature (Plucker & Makel, 2010; Reiter-Palmon et al., 2019).

DT tasks are characterized by open-ended response formats, which prompt the participant to produce several ideas in response to a stimulus (e.g., an abstract picture or a common word). Since the theorizing of Guilford (1956) and Torrance's (1962, 1970) emphasis on supporting DT in students, participant data has been used by researchers to indicate a number of specific dimensions of DT, including Originality (average novelty of ideas; Runco, 1999), Fluency (quantity of ideas; Hocevar, 1980), Flexibility (diversity of ideas; Guilford, 1968), and Elaboration (the extent to which an idea is explained; Torrance, 1988). The Originality of a given response typically represents the relative *remoteness* or *uncommonness* of an idea, such that an idea that is more unusual or uncommon in relation to a given prompt or context would be considered to be more original. Scoring a response for Fluency reflects the aspect of DT concerned with productivity, meaning the more ideas produced in a response, the higher the Fluency score would be. Truly divergent thought, a term coined by Acar and Runco (2015), is not only characterized by uncommon or unique ideas, but also by the generation of ideas in many different directions (Taylor, 1988), as implied by the concept of divergence. The extent to which an individual generates responses that fall into different categories, rather

than responses that follow one cognitive pathway, is a reflection of Flexibility. Therefore, the more diverse each response is from the others, the greater the Flexibility score. Refining or Elaborating on a specific idea has been considered to improve the quality of an idea (Runco & Pritzker, 1999), meaning the level of detail with which a response is explored (Besemer & O'Quin, 1987) may be an indicator of quality. Among these four dimensions, DT tests are most often scored for Originality and Fluency (Hornberg & Reiter-Palmon, 2017).

Divergent thinking abilities are typically measured using both verbal and figural DT tasks (Kuhn & Holling, 2009). Figural DT tasks ask the respondent to participate in activities, such as picture construction and picture completion, or meaning extraction from patterns, lines, or circles. In verbal DT tasks, individuals are presented with a textual problem or stimulus and asked to generate as many written or verbal responses as possible, such as unique uses for an object. Several different verbal and figural DT measures can be found in the creativity literature, such as the widely used Torrance Tests of Creative Thinking (TTCT; Torrance 1998), encompassing both verbal and figural components, the Remote Associates Test (RAT; Mednick, 1968), the Evaluation of Creative Potential (EPoC; Lubart et al., 2011) battery, the Similarities Test and Instances Test (Wallach & Kogan, 1965), and Test of Creative Thinking – Drawing Production (TCT-DP; Jellen & Urban, 1996). Among the various DT measures available, the Alternate Uses Task, (AUT; Guilford, 1967) a verbal DT measure, has dominated DT measurement for many years and remains the most frequently utilized task within creativity literature (Puryear et al., 2017).

The AUT presents participants with a common object (e.g., book) and instructs participants to generate as many creative uses for it as possible. For example, if participants are presented with the prompt "brick", they may respond with something along the lines of "build a home" or "weapon against predators" for possible uses. Within verbal DT tasks, such as the commonly administered AUT, responses are typically scored on three dimensions: Originality, Fluency, and Flexibility (Guilford 1968; Torrance, 1995). Originality is particularly important and represents the most often quantified DT index because the standard definition of creativity points to Originality as the prerequisite for all creativity (Runco & Jaeger, 2012). Other DT indices, specifically Elaboration, are traditionally more commonly explored in Figural tasks, if at all.

Despite the broad use of the AUT in creativity research, accessibility to efficient scoring methods is limited. Additionally, a thorough psychometric understanding of participant scores from the AUT is currently lacking in comparison to other areas of psychology. These shortcomings of DT measurement are largely due to the open-ended and ill-structured nature of responses on which it relies. Unlike most other psychological attributes that can be quantified using closed-ended items, such as numerical rating scales (e.g., personality attributes), reliable estimates of DT cannot be attained unless participants are allowed the freedom to generate their own ideas through open-ended response sets. However, most psychometric modeling frameworks that evaluate the reliability of scoring models (e.g., item response theory; Lord, 2012) were designed for closed-ended response sets, making it very challenging to assess the reliability of latent scores from the AUT. Not only does the AUT have an open-ended response format (i.e., participants must verbally respond or fill in their responses), it also encourages ill-

structured responses (i.e., participants are not restricted in terms of the number of responses they can supply or the length of those responses.) This response format is incompatible with most psychometric scoring methods, requiring more subjective methods to be utilized when scoring the AUT. To this end, the Originality of participant responses to DT tasks, such as the AUT is often quantified through subjective human judges (e.g., Harrington, 1975; Silvia et al., 2008; Zarnegar et al., 1988). However, inter-rater reliability of judge-based Originality scores can be low, and even when reliability is achieved, human judgment is time- and resource-intensive. Also, the reliability and validity of human-rated Originality scores is subject to measurement error due to the raters' implicit beliefs and biases. As a result, the subjectivity of scores derived by human-judges has been criticized in the literature (Gwet, 2014), but no alternative for open-ended and ill-structured response types have historically existed.

Computational Psychometrics

However, recent advances in text-mining methodology have offered some solutions for the reliable quantification of complex mental attributes like DT that cannot be assessed via closed-ended measures. These solutions are based on a new area of research, called *computational psychometrics* (von Davier, 2017), which combines psychometric methodology and text-mining to produce a method for quantifying language (i.e., mapping words into numbers; e.g., Neuman & Cohen, 2014; Park et al., 2014; Kern et al., 2016). These computational psychometric methods are based on text-mining algorithms that are trained on a massive corpus of text meant to represent the semantic structure of a given language.

Text-mining models can determine the semantic meaning of words by the context in which the specific words appear in the corpus of text. For example, the text-mining model most often used in the psychological literature today is Latent Semantic Analysis (LSA; Landauer et al., 1998), which is a dimensionality reduction technique that derives high-dimensional vector representations of word meaning from the distribution of words in a text corpus. The LSA modeling approach is centered around the distributional hypothesis that the meaning of a word is reflected in its use in language (i.e., the context in which the word is used) and thus in its distributional pattern (Harris, 1954; Lenci, 2008). LSA represents texts in a word × document matrix of word co-occurrence counts, in which the rows represent words, the columns represent documents, and how often each word occurs in each document is specified in the cells. The semantic meaning of each word in the matrix is represented as a numerical vector, and semantic similarity among words can be estimated by taking the cosine of the angle between the word vectors (see Landauer & Dumais, 1997 for a more detailed description). Weighting schemes are often applied to the word-document counts to decrease the extent to which word vectors are impacted by word frequency and to identify informative word-document co-occurrences (Martin & Berry, 2007). Additionally, singular value decomposition (SVD; Golub & Kahan, 1965) is applied to reduce the matrix of word-document counts to a much smaller representation of documents. SVD reduces noise in the data, and thereby reveals the latent structure of the text corpus. The resulting high dimensional structure is referred to as a semantic space.

The capability of text-mining models to represent the semantic structure of a given language has provided psychometricians the opportunity to begin to attempt the

objective (i.e., not human-rated) and automatic (i.e., computer-generated) quantification of complex psychological constructs from open-ended data sources. In fact, some applications of text-mining methodology are experiencing relatively broad use. For example, some complex psychological constructs that are now frequently quantified via text-mining models include, depression (i.e., through textual analysis of patient interview data; Kjell et al., 2019), writing ability (i.e., via computerized essay scoring systems; Rehder et al., 1998; Foltz et al., 2013), and collaborative ability (i.e., examination of semantic structure of participant collaboration efforts on an online chat; He et al., 2017).

In recognition of the broad capabilities of computational psychometrics in various types of knowledge representation, creativity researchers have begun adopting computerized methods, especially text-mining methods, to supplement their own assessment procedures. Evidently, the application of text-mining methodology to creativity assessment is rather valuable, given the often open-ended and ill-structured nature of response data collected in that area of research. In effect, the use of text-mining methods has extended to the assessment of DT, specifically as a scoring mechanism for verbal DT tasks.

Semantic Distance as a Measure of Originality

Since Guilford's (1956, 1968) initial conceptualization of divergent thinking (i.e., divergent production), DT has been a commonly researched construct, resulting in a considerable number of theoretical advances, and the continuous refinement of the methods used to assess DT. Although his early work recognized four DT indices (i.e., Originality, Flexibility, Fluency, and Elaboration), Originality has received noticeable attention, due to its recognition as a central component of creativity (Runco & Jaeger,

2012; Simonton, 2018). As such, the feasibility of sophisticated scoring procedures for participant Originality has been an increasing concern across the creativity literature.

The Originality of participant responses to creative thinking tasks are measured by indicators of *uncommonness*, *cleverness*, and *remoteness* (Wilson et al., 1953). Thus, original ideas or responses are remote in associative terms. The consideration of Originality in terms of *associative distance* – a viewpoint first conceived by Mednick (1962) – has allowed creativity researchers to conceptualize Originality from DT data in a way that can be represented through text-mining methodology as *semantic distance*. For example, in terms of associative distance, the Originality of an idea is determined by its distal-relatedness from the context (e.g., a given prompt or stimulus) from which it arose, such that an idea that is normally associated with the given context would be determined as low in terms of Originality and an idea that is uncommon or remote in the given context would be determined more original. Generally, the idea of associative distance described here can be further conceptualized as the semantic distance between a response and a DT prompt. Therefore, rather than basing the Originality of a participant's response on the remoteness or uniqueness of their idea, it can be determined by the semantic distance between their response and the prompt from which it arose. A greater semantic distance would indicate less similarity between the participant's response and the DT prompt, signifying a more remote or original response. A smaller semantic distance would represent greater similarity between the response and prompt, indicating a less original response. The ability of text-mining models to represent language and derive semantic relations among responses to DT tasks (e.g., AUT) warrants the application of computational psychometric techniques (e.g., text-mining algorithms such as LSA) to DT

task data. Indeed, such techniques can quantify the semantic distances among DT responses and prompts to generate semantic distance scores, which can then be used to operationalize the associative distance of the responses, and thus the Originality of ideas. Originality scores among ideas can then be psychometrically aggregated to yield the latent Originality of participants.

Text-Mining Model Use in Creativity Research

Over the last handful of years, some creativity researchers have been engaged with the creation of automated and objective Originality scoring systems based on text-mining models (Dumas & Dunbar, 2014; Forster & Dunbar, 2009; White & Shah, 2016). These models are trained to identify common patterns of word co-occurrence in a language (e.g., English), and then can be applied by researchers to identify instances where participant responses to a DT task utilize language in an uncommon or Original way (Acar & Runco, 2014; Heinen & Johnson, 2018). Previous work with text-mining model-based Originality scores have shown the method to be useful both in saving the time and money it takes to score DT tasks, but also in uncovering psychologically relevant relations among DT scores and other creativity related measures (Dumas et al., 2020; Gray et al., 2019; Harbison & Haarmann, 2014).

The text-mining algorithm that has dominated the creativity research literature thus far is LSA (Acar & Runco, 2019). LSA has been demonstrated to be an appropriate and effective tool for modeling meaning representation from text to estimate the semantic distances among word and phrases to quantify Originality (Dumas & Dunbar, 2014). The semantic meaning of words are represented as vectors in a geometrically represented space and the cosine of the angle between the word vectors represents the semantic

similarity or distance among words, hence the associative distance among those words (Deerwester et al., 1990). Therefore, when estimates of Originality are determined for a response to an AUT prompt, the cosine of the angle between the prompt and response would be calculated to represent the semantic similarity. The semantic similarity score is then subtracted from 1 to yield a semantic distance score (i.e., Originality score) for each response.

Beginning with the work of Forster and Dunbar (2009), the usefulness of LSA based Originality scores has been explored in various contexts and LSA has been utilized to answer a range of questions in the creativity research literature. Forster and Dunbar (2009) were, to the best of my knowledge, the first to demonstrate the feasibility and usefulness of the application of LSA as a tool for scoring DT tasks. They found that LSA-based Originality scores were better predictors of judged creativity than were the traditional DT indices (i.e., Fluency and Elaboration). This study represents one of few investigations using verbal DT tasks that scored responses for Elaboration. In 2014, Dumas and Dunbar explored the psychometric properties of LSA-based Originality scores and their relation to Fluency scores from the AUT. They demonstrated that LSA generated Originality scores have discriminant validity from Fluency scores. Dumas and Runco (2018) explored the relation between Fluency and LSA based Originality scores even further using a partial correlation approach, revealing that Originality scores generated by the LSA model exhibit a high level of reliability, even after the variance explained by Fluency was partialled out.

In a slightly different line of work, Hass (2017) applied LSA as a method to calculate *local similarity* (semantic relation between adjacent responses) and *global*

similarity (semantic proximity between prompt and each response). He found that judged creativity was negatively associated with both local and global similarity. Hass (2017) also discovered that participants took less time to generate a similar response than a dissimilar response. Additionally, LSA has been used for its objectivity in quantifying creative potential as a method to compare with self-rated creativity and other-rated creativity. Harbison and Haarmann (2014), noticed that LSA-based indices exhibited a very different relation with self-rated creativity than they did other-rated creativity. Insights about other psychological phenomena, such as Attention-Deficit/Hyperactivity disorder (ADHD), have been facilitated by the use of LSA. Specifically, White and Shah (2016) demonstrated that semantic distances between word association pairs could explain observed advantages of individuals with ADHD on a measure of DT, which motivated the hypothesis that DT may be assisted by ADHD because of the wider range of semantic activation in individuals with ADHD. Further, LSA-based Originality scores have demonstrated predictive validity in terms of the method's ability to significantly predict malevolence, a construct theoretically relevant to creativity. LSA-based Originality scores were used by Dumas and Strickland (2018) to explore malevolent responses on the AUT. They found a significantly positive relation between malevolence, as indicated by the generation of a violent response to a prompt (e.g., responded to the prompt "brick" with the use "kill someone") and participant Originality scores.

Limitations of Text-Mining Models

As the use of text-mining models in the creativity research literature continued to grow, concerns related the ability of these models to produce both reliable and valid estimates from DT tasks began to develop. The reliability and validity of Originality

scores estimated using text-mining models are dependent on a various factors, but for the purpose of this study, two factors are highlighted: a) the text corpus on which the model was trained and b) the methodological choices made to handle common words (e.g., "and", "the", "is") and to minimize the extent to which text-mining model-based scores are influenced by varying amounts of words in a response (i.e., varying amounts of Elaboration).

Training corpora. The most commonly used training corpus for LSA in the creativity literature has been the Touchstone Applied Science Associates (TASA) corpus (Landauer & Dumais, 1997; Kintsch & Bowles, 2002). The TASA corpus is meant to represent the average reading experience of an English-speaking undergraduate student. Despite the frequency of the TASA trained LSA model in the creativity literature compared to other text corpora (e.g., Forster & Dunbar, 2009; Beaty et al., 2014; Prabhakaran et al., 2014) , it is arguably outdated, as it was originally created in 1997 and has not been updated since the early 2000's. It is possible that the TASA corpus is no longer an accurate enough representation of language use to be a reliable corpus for a LSA model to be based. In substitution of the TASA corpus, some creativity researchers have turned to another text corpus to use in their applications of LSA as a possibly more accurate representation of the semantic structure of language. The English 100k (EN 100k; Günther et al., 2015) corpus has been utilized very little in the creativity literature thus far (e.g., Forthmann et al., 2019; Dumas et al., 2020), but shows perhaps greater potential for the reliable quantification of Originality from DT task data than the TASA LSA. This notion is based on the fact that EN 100k is trained on multiple corpora of text including a 2009 Wikipedia dump, the British National Corpus, and web crawl corpus.

Thus, EN 100k represents a more general and newer corpus of text than TASA, possibly making it a better representation of true semantic relations among words.

The confounding influence of elaboration on originality. One of the inherent drawbacks of text-mining models when applied to DT responses is the influence that the open-ended nature of DT responses has on the models' estimations of Originality. Since participants can vary in the level of detail they use to express an idea in response to a DT prompt (i.e., vary in Elaboration), their responses will likely vary in terms of the number of words used to express an idea. This variation in number of words across responses has been found to be a problem for LSA based estimations of Originality because semantic similarity (i.e., cosine of the angle between the two-word vectors) is dependent on the number of words that enter the LSA model (Graesser et al., 2013; Penumatsa et al., 2006). In more technical terms, when a participant's response contains more than one word (they almost always do), the LSA model will sum the vectors of each word included in that response to obtain the meaning of the entire phrase, which is then used to determine the semantic distance of the response from the DT prompt, and thus an estimate of Originality for that response (Landauer et al., 1997). Since some words are frequently used in responses, known as *stop words* (e.g., "and", "the", "is"), the use of these common words decreases the semantic distance of the response from the prompt, and thus lowers the Originality of the response. For example, if the item "hammer" was administered as the stimulus item in the AUT, text-mining models would certainly give Originality scores to the very different responses "use it as weapon" and "smash the head of a zombie that was attacking you in the apocalypse." The concern centers around the question of whether these Originality score differences are psychologically valid.

Because automatic Originality scores analyze the specific words used by participants, they are (perhaps unduly) sensitive to the number of words used (Forthmann et al., 2019; Forthmann et al., 2020). In effect, because text-mining model results are influenced not only by the Originality of an idea itself, but also by the specific words and the number of words used to express each idea, the degree to which participants elaborate their DT responses can influence their Originality scores. Consequently, concerns have been raised about the ways in which these automatic and objective Originality scores may be influenced by other dimensions of DT, especially Elaboration. Specifically, these concerns are directed towards Elaboration operationalized as word count within a participants' verbal (or written) response to a DT task.

Traditionally, Elaboration has not been a commonly explored DT index in verbal DT tasks. However, Forster and Dunbar (2009) assessed Elaboration as one of the verbal DT indices in a comparison between LSA and human judged creativity on the AUT. They reported a relatively large negative correlation of LSA based Originality and Flexibility with Elaboration (indicated by the average number of characters in a response). Later, Forthmann et al., (2017) observed a very similar relation between Originality and Elaboration. More recently, in a simulation performed by Forthmann and colleagues (2019), they investigated this phenomenon (which they refer to as elaboration-bias) known to be caused by the vector addition method utilized by LSA to represent text. Results from Forthmann and colleagues' (2019) study revealed that Elaboration (defined as word count) confounds LSA-based Originality estimates from the TASA corpus (Landauer & Dumais, 1997) and the EN 100k corpus (Günther et al., 2015), such that as more words were used in response to an item, the smaller the semantic distance between

a prompt and a given response was expected to be (decreasing Originality scores). Once stop-word removal and a simulation-based bias correction were applied, semantic distances from both text-mining models were left uncorrelated with Elaboration. Overall, the slightly larger and newer EN 100k corpus was less affected by Elaboration and was determined more valid than the older TASA corpus, which is slightly smaller in size. Based on these findings, Forthmann and colleagues recommended that LSA and other vector models undergo stop-word removal and a bias correction in future investigations to limit the Elaboration confound. Also, they advise the use of a larger corpus when applying LSA to DT task data in the future to help maximize validity. This demonstration informs the focus of the current study, as I seek to address the recent concerns surrounding the confounding influence of Elaboration, defined as word count, on the reliability of Originality scores using recent recommendations for text-mining methodology.

Here, I report a test of the degree to which the reliability of Originality scores produced via a freely-available text-mining model is dependent on the variance accounted for by Elaboration. The text-mining model employed in this study has very recently been identified as a more valid and reliable model for the assessment of divergent thinking compared to LSA (Dumas et al., 2020). The intention here is not to demonstrate with finality the full reliability and validity of automatic Originality scores. Important psychometric issues surrounding these scoring systems will need to be addressed in the future. The particular goal here is to provide specific details about the current feasibility of automatic Originality scoring, taking Elaboration into account and statistically eliminating it as a possible confound.

CHAPTER TWO: METHODOLOGY

This is a reanalysis of data collected by Dumas et al. 2020. The goal of that study was to assess the reliability and predictive validity of computer-generated Originality scores produced by four major text-mining systems in comparison to human-rated scores. Correlations among Originality and Elaboration were computed, but the confounding effect of Elaboration on the internal consistency of Originality was not part of that earlier study. The goal of this study is to build off of the correlations between Originality and Elaboration found by Dumas and colleagues (2020) and the Elaboration confound demonstrated by Forthmann and colleagues (2019) to investigate this relation in a deeper manner using recent recommendations for text-mining methodology.

Participants

Participants were recruited via Amazon Mechanical Turk, an online crowdsourcing marketplace widely used in psychology research. Only self-reported fluent English speakers over 18 years of age were selected to participate. The final sample included 92 (53 female; 57.6%) participants, ranging from 21 to 68 years of age (M = 37, SD = 10.58). The majority of participants (n = 68; 73.9%) identified as White, with the remainder identifying as Black (n = 6; 6.5%), Asian (n = 9; 9.8%), Latinx (n = 5; 5.4%), or multiple ethnicities (n = 4; 4.2%). A compensation of $3.00 was given for participation.

Measures

The Alternate Uses Task (AUT; Guilford, 1967) is the most frequently employed psychometric measure of divergent thinking and creative potential (e.g., Dumas, 2018; Forthmann et al., 2016; Runco & Acar, 2012). Examinees are presented with a common object and instructed to generate as many creative uses for it within a given time (i.e., 2 minutes per object in this case). Participants were shown the following object names in random order: book, rope, fork, table, pants, bottle, brick, tire, shovel, and shoe.

Scoring the AUT: Elaboration

Concerns related to the confounding of Elaboration with automated Originality scores refer specifically to Elaboration defined as word count. As such, Elaboration for each AUT item (e.g., book) was determined by counting the number of words used by a participant in each response to that item, and then those response-level counts were summed within each of the 10 AUT items. The item-level word counts were then averaged across the 10 items to yield an Elaboration score for each participant. For the purpose of this study, it is critical that the data show sufficient variability in Elaboration across the 10 items due to the importance of this condition for the Elaboration confound to emerge. A low variance in this regard would likely prevent the Elaboration confound from appearing, giving our continued investigation of this phenomenon with these data little purpose. However, Table 1 reveals that the data do in fact display adequate variability in Elaboration, supporting the appropriateness of these data for the purpose of this investigation. Within this analytical sample, on average, participants responded with 13.03 words per AUT item (SD = 10.46).

Scoring the AUT: Originality

Originality scores were generated using the freely available Global Vectors for Word Representation *840B* system (GloVe; Pennington et al., 2014). Trained on a corpus of 840 billion words scraped from a variety of online sources such as Wikipedia and Twitter, GloVe quantifies the semantic relation between two words or phrases in a geometric space. GloVe was used to estimate the semantic distance between each AUT prompt (e.g., book) and response (e.g., read for fun) by calculating the cosine of the angle between the word vectors (see Rakib et al., 2018 for a technical treatment of this procedure). GloVe is conceptually very similar to other more commonly used models for the objective quantification of Originality (e.g., Latent Semantic Analysis; LSA; Landauer et al., 1998; Forthmann et al., 2017) and was recently identified as the most reliable and valid text-mining system for use in divergent thinking research (Dumas et al., 2020). As a result of its massive corpus and probabilistic modeling approach, GloVe *840B* may be more appropriate for the measurement of Original thinking than the smaller corpora and traditional parametric approach on which most LSA text-mining models utilized in creativity research (e.g., Beaty et al., 2014; Dumas, 2018; Prabhakaran et al., 2014) are based.

In addition, to best address methodological concerns within the creativity literature related to the confounding of Originality with Elaboration (e.g., Forthmann et al., 2019), an inverse document frequency (IDF) weighting scheme was applied to the GloVe system, in which individual words within AUT responses were weighted depending on their commonness within the 840 billion word GloVe corpus, with common words weighted weakly and uncommon words weighted more strongly (see

Organisciak, 2016 for a technical description of IDF weighting). The intent behind the application of the IDF weighting scheme on the GloVe system was to minimize the influence of word count of an individual response on Originality scores. With the use of the IDF weighting scheme, we hope to maintain better control of the Elaboration confound. Raw cosine-distances (ranging from -1 to 1) from the GloVe system represent semantic similarity, and were subtracted from 1 to yield Originality scores that ranged from 0 to 2 (with higher values indicating more Originality). Originality scores for each response to an AUT item-prompt (e.g., book) were averaged within the item. Item-level Originality scores were averaged across the 10 items to create an Originality score for each participant. Within this sample, participant Originality scores averaged .71 (*SD* = .04) across the 10 AUT items.

Residualized Originality

Partial correlation procedures, similar to those used by Runco and Albert (1985) and Dumas and Runco (2018), were employed. To partial the variance in the Originality scores that could be attributed to Elaboration, 10 separate regression models were fit in which Elaboration scores from each of the 10 AUT items predicted Originality scores associated with the same items. The corresponding residuals from each of the 10 regression analyses were saved (in z-score format) and represented the variance in Originality scores for each AUT item completely independent of Elaboration.

CHAPTER THREE: RESULTS

To illustrate the extent to which Elaboration, or word count in a given response influences the reliability of Originality scores generated by the GloVe *840B* text-mining system, the analysis unfolded in the following stages: (a) a thorough investigation of the level of reliability attained by the GloVe Originality scores; (b) an examination of the correlation between Elaboration and GloVe Originality, as well as an analysis of individual R-squared values from the 10 linear regression models; and (c) a reexamination of the reliability of GloVe Originality scores after Elaboration was partialled out (now residualized Originality scores) in comparison to the nonresidualized Originality scores. An explanation of these stages and results are presented below.

Reliability

The primary concern of this investigation was the reliability of participant Originality scores estimated by the GloVe system both before and after Elaboration was statistically controlled. Accordingly, the reliability of Originality scores, residualized Originality scores, and Elaboration scores from the AUT were examined using both Cronbach's alpha as a measure of internal consistency reliability and Raykov's reliability (*RR*; Raykov, 1997) as an estimate of factor reliability. Due to the commonness of Cronbach's alpha in the literature (Sijtsma, 2009), its use as a reliability estimate in this analysis was to allow for an easier comparison of reliability across studies. However, alpha requires the assumption of essential tau-equivalence, which is unlikely to be met by

the measures used in this study and unnecessary for the measurement of latent Originality. For this reason, estimates of Raykov's factor reliability for congeneric measures (1997) were generated as a more accurate representation of reliability than alpha. Conceived within the framework of Classical Test Theory (CTT), Raykov's (1997) approach to estimating reliability uses structural equation modeling (SEM) methodology to assess the reliability of congeneric measures, without the requirement of essential tau-equivalence—a condition typically violated by psychological measures (McNeish, 2018). When tau-equivalence is violated, alpha has been shown to underestimate the reliability of a measure (Miller, 1995; Graham, 2006). Thus, alpha was used as a lower bound estimate of reliability for the measures in this analysis.

Internal consistency reliability. Based on coefficient alpha (a lower-bound estimate of reliability in this case), participant Elaboration scores across the 10 items of the AUT demonstrated a high level of internal consistency ($\alpha = .959$). The 10 item-level GloVe Originality scores demonstrated a border-line acceptable level of internal consistency ($\alpha = .784$). Once the variance in GloVe Originality scores explained by Elaboration was partialled out through the 10 separate regression models, the new residualized GloVe Originality scores displayed a nearly identical level of internal consistency ($\alpha = .783$) to the non-residualized Originality scores. These findings suggest that the GloVe text-mining system (at worst) generates Originality scores with a level of reliability that is border-line acceptable in the psychological research literature (.80 or above generally acceptable). More importantly, the minimal decrease in the reliability of GloVe Originality after Elaboration was statistically controlled implies that the GloVe Originality scores were not meaningfully confounded by Elaboration.

Confirmatory factor analysis. Three separate unidimensional confirmatory factor analysis (CFA) models were fit to examine the factor structure of the AUT with respect to Elaboration, GloVe Originality, and residualized GloVe Originality, and allow for the estimation of Raykov's factor reliability. Using the summed scores for word count across the 10 AUT items, a single-factor CFA model was fit, in which the 10 item-level Elaboration scores loaded on the Elaboration factor. The subsequent CFA models correspond to the theoretical assumption that all of the AUT prompts (when scored for Originality) indicate a single underlying Originality construct. Respectively, another CFA model was fit, in which the 10 item-level Originality scores generated by the GloVe system were loaded on the Originality factor. Then, after Elaboration was partialled out of the Originality scores, the 10 item-level residualized Originality scores loaded on the residualized Originality factor in the third CFA. Conceptual path diagrams of each of the three CFA models can be found in Figures 1, 2, and 3 and the standardized loadings can be found in Table 3. The CFA models were fit using maximum likelihood estimation in STATA version 15.1. Based on the model root mean square error of approximation (RMSEA), none of the models achieved a level of model-data fit that would be considered ideal in the methodological literature (i.e., below .06; Hu & Bentler, 1999; McNeish et al., 2018). However, the model RMSEA values would be considered more acceptable for the current standards in the creativity literature, where measurement model-data fit is often slightly weaker than in other areas of measurement (e.g., Yoon, 2017). Please see Table 2 for the model fit statistics for each of the three CFA models.

Factor reliability. Based on the three CFA models, Raykov's factor reliability estimates were generated using the "relicoef" module in STATA 15.1 (Mehmetoglu, 2015). Elaboration scores across the 10 items demonstrated a high level of factor reliability (*RR* = .961). The 10 item-level GloVe Originality scores demonstrated a good level of factor reliability (*RR* = .800) that would likely be considered acceptable in the creativity literature. Once Elaboration was statistically controlled, the new residualized Originality scores for each of the 10 object-prompts displayed an identical Raykov's reliability estimate (*RR* = .800) to the non-residualized Originality scores. Based on the *RR* estimates above, which are anticipated to be more accurate than alpha, the GloVe text-mining system is capable of producing reliable Originality scores that are unaffected by Elaboration, at least within the decimal places included in this calculation.

Influence of Elaboration on Glove Originality

The bivariate correlation between Elaboration and Originality for the full AUT was small ($r = .13$, $p = .22$). This finding reveals that Elaboration, or number of words used in a response, and response Originality (generated by the GloVe system) did not share a lot of variance and therefore, do not appear to have a meaningful relation in this study. This small relation corroborates the conclusion that Elaboration did not meaningfully confound GloVe Originality scores in this investigation.

After statistically controlling Elaboration by partialling out the variance in Originality scores attributed to Elaboration via the 10 regression models (one regression model for each AUT prompt), the corresponding R-squared statistics were analyzed. Each of the 10 regression analyses produced small R-squared statistics, ranging from <.001 to .057, although most of the R-squared values were substantially lower than .057

(see Table 1 for exact R-squared values). The small R-squared statistics reveal that Elaboration explained as much as 5.7% of the variance in participant Originality scores and as little as .1%. Moreover, the variance in Elaboration, or number of words used in a response, had very little influence on the Originality scores generated by the GloVe system.

CHAPTER FOUR: DISCUSSION AND FUTURE DIRECTIONS

The results from this analysis address the important concern that too much variance in Originality scores is accounted for by the number of words used in DT responses (i.e., Elaboration). The results revealed that—at least when modern methodological recommendations for text-mining Originality scoring are applied (Forthmann et al., 2019)—the reliability of automatic Originality scores generated by the GloVe *840B* system is not meaningfully confounded by Elaboration.

Glove is a Useful Text-Mining Model for Scoring Originality

Most of the existing DT research utilizing text-mining models apply the TASA LSA (e.g., Dumas & Dunbar, 2016; Hass, 2017), which as mentioned earlier in this work, has not been updated since the early 2000s. A very small portion of the research on DT has applied the EN 100K LSA (Forthmann et al., 2019; Dumas et al., 2020). These LSA models may yield less accurate Originality estimates compared to other models, such as GloVe, because the LSA space is unable to update its semantic representations in response to humans' continual accumulation of language. In other words, language is always evolving and adapting as humans evolve, but the LSA space cannot integrate new documents after it has been built without recomputing the original matrix. Doing so

would require the LSA space to be rebuilt entirely. This fact prevents the LSA space from adapting to the dynamic nature of human language, calling into question the validity of this roughly 20-year-old system (Recchia & Jones, 2009; Lemaire & Denhière, 2004).

To the best of my knowledge, the GloVe *840B* text-mining model has been used in the creativity literature only once prior to this investigation (Dumas et al., 2020), but based on the findings from both analyses, creativity researchers might reconsider their use of the TASA LSA and EN 100K and instead apply the GloVe *840B* system to quantify Originality in future investigations using DT task data. Since tau-equivalence, a condition required for accurate estimations of Cronbach's alpha, is not typically met by the measures used in this study (Dumas & Dunbar, 2014), I point toward the Raykov's factor reliability (*RR*) estimates as the more reasonable representation of the reliability of the Originality scores, and the alpha estimates as a lower bound. With this in mind, note that when estimating the reliability of GloVe Originality scores using Raykov's method, the factor scores demonstrate a good level of reliability, which is unaffected by the potential confounding of Elaboration, at least within the decimal places included in this calculation, and within the statistical power of this investigation given the number of items administered and the sample size ($RR_{Originality}$ = .800, $RR_{Residualized\ Originality}$ = .800).

Please note, the relation between automated Originality scores and Elaboration found in this study is noticeably different than that of the study conducted by Forthmann and colleagues (2019). It appears that certain methodological choices made in this demonstration allowed for a better control of what Forthmann and colleagues called the *elaboration-bias*. This study's better control of the elaboration-bias compared to that of Forthmann and colleagues (2019) is possibly due to: (a) GloVe's larger corpus size

compared to that of the TASA LSA and EN 100K LSA, (b) the modeling framework on which GloVe's semantic space is created (i.e., modern log-bilinear modeling approach), and (c) the application of the IDF weighting scheme to the GloVe model. Demonstrations have shown these components to be advantageous in effectively detecting the semantic meaning of words and phrases, often leading to greater validity in GloVe estimations compared to those from other models, such as LSA (e.g., Pennington et al., 2014; Naili et al., 2017; Dumas et al., 2020).

Generally, a larger training corpus may be more representative of real-word language use, and therefore more capable of reproducing the semantic structure of a given language (Reccia & Jones, 2009). It is my belief that the substantial increase in word coverage achieved by the GloVe system compared to that of the models used by Forthmann and colleagues (2019) may have led to more psychologically valid and reliable Originality scores, resulting in a better control of the elaboration-bias in this study. Additionally, the GloVe system combines the main benefits of matrix factorization, count-based methods (e.g., LSA) that can exploit global statistical information with more modern, log-bilinear predictive methods to produce a vector space with meaningful linear substructures (see Pennington et al., 2014 for a more detailed explanation). More specifically, GloVe examines the semantic relationship between words by studying the ratio of their co-occurrence probabilities with various probe words, while LSA simply examines the raw probabilities. Compared to raw probabilities, ratios help to reduce noise because they can better discriminate between two very similar words and distinguish between relevant and irrelevant words (Pennington et al., 2014). Therefore, it is likely that ratios can more effectively estimate semantic distance.

To further address concerns related to the elaboration-bias an inverse document frequency (IDF) weighting scheme was applied to the GloVe system, in which words were weighted based on their commonness within the word corpus (see Organisciak, 2016). This technique was intended to minimize the influence of word count in an individual response, or Elaboration, on Originality scores derived by GloVe, which I presume helped to more effectively avoid the elaboration-bias.

This blend of approaches that make up the GloVe framework in this study have been shown to be more effective in detecting the semantic meaning of words and phrases than estimations based on LSA (e.g., Naili et al., 2017; Dumas et al., 2020; Pennington et al., 2014). For example, Naili and colleagues (2017) found that the traditional count-based method employed by LSA yielded the largest error rate in learning word vector representations compared to the more efficient count-based prediction method utilized by GloVe. In the recent comparison of the reliability and validity of four text-mining models (EN 100k LSA, TASA LSA, Word2Vec, and GloVe *840B*; Dumas et al., 2020) on which this study is based, GloVe emerged as the best model to use for the estimation of Originality. In another demonstration by Pennington and his colleagues (2014), GloVe significantly outperformed other models trained on corpora of the same size, and often outperformed models trained on substantially larger corpora on word analogy, word similarity, and named entity recognition tasks. This suggests that GloVe may have an advantageous modeling approach that is contributing to its better performance regardless of corpus size. Thus, it is plausible that the use of the GloVe model instead of the LSA model used by Forthmann and colleagues (2019) elicited more psychologically relevant and valid results, contributing to a better control of the elaboration-bias.

In my estimation, the combination of these factors allowed for a more valid representation of semantic relations compared to the TASA LSA and EN 100k LSA, and thus more psychologically valid and reliable Originality scores. As a result, it is possible that this approach limited the emergence of the problematic relation often found between Originality and Elaboration (e.g., Forthmann et al., 2019).

Limitations and Future Directions

Since this analysis was based solely on one DT measure (i.e., AUT) and it has been shown that DT scores for the same participant can vary across different DT tasks (Runco et al., 2016), it would be beneficial for future studies to assess how GloVe Originality scores perform when they are based on other well accepted DT measures, such as the TTCT (Torrance, 1999). It is possible that some DT tasks are more prone to Elaboration, or word count biasing Originality scores than others. Also, since these results are entirely based on a native English-speaking population, it would be interesting to investigate the relation between text-mining model-based Originality and Elaboration in a non-native English-speaking sample. The relation between word count and Originality could look very different in a sample with a different cultural background. Similarly, future investigations of divergent thinking using text-mining models should consider using GloVe to assess the correlation between Elaboration and Originality in a sample of children. According to the standard definition of creativity (Runco & Jaeger, 2012), an original idea is not creative unless it is useful or effective. As far as I am aware, creativity researchers have used text-mining models to estimate the Originality of an idea, but not the usefulness of an idea. A meaningful future direction might attempt the use of a text-mining model to assess the Originality and usefulness of an idea in response to a

measure such as the AUT. Then, aggregate the Originality and usefulness scores to suggest how creative an idea is. The combination of objective Originality and usefulness estimates would inform an individual's level of creativity, rather than just creative potential, which is all that DT tasks can measure without taking into account usefulness. Through the objective quantification of usefulness via text-mining models, creativity researchers would avoid potential reliability issues that can sometimes arise when subjective judge-based scoring methods are used.

Issues with quantifying elaboration. It is important to note that Elaboration and word count are not synonymous, but rather word count is one operationalization of Elaboration (Forster & Dunbar, 2009). Elaboration is an inherently subjective DT index, however defining Elaboration as word count allows for this index to be scored in a more objective and efficient manner. Objectively quantifying Elaboration has some drawbacks. One particular issue with quantifying Elaboration as word count is that a 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" (Forster & Dunbar, 2009, p. 603). When Elaboration scores are derived via count-based methods on the computer, there is no way of knowing if participant responses are coherent or meaningful. Thus, defining Elaboration as word count is not guaranteed to be informative of divergent thinking or creative potential, but rather useful as a means of controlling for the potential confounding influence that the vector addition method can have on automated Originality scores.

Recommendations of bias-corrections for elaboration. Despite the chosen text-mining system for scoring DT tasks, creativity researchers should actively try to avoid the

elaboration-bias that can emerge from the vector addition method used by semantic algorithms to represent text. This bias is likely to emerge in other word vector models too, as they all employ very similar methods when representing the semantic structure of language (Jones et al., 2015; Mandera et al., 2017). It is vital that a techniques, such as stop-word removal, a simulation-based correction (Hass, 2016a,b; Forthmann et al., 2019), or an IDF term-weighting correction (Organisciak, 2016; Dumas et al., 2020) are applied to the automated Originality scores to help control for the likely confounding of Elaboration with Originality. Forthmann and colleagues (2019) have demonstrated that the removal of stop words decreases the variation in the number of words used to express ideas that can be seen as synonymous. The residual variance that still remains after stop-word removal can then be accounted for by the simulation-based correction, reducing the elaboration-bias even further.

Closing Comments

Considering the frequent concerns regarding the confounding of Elaboration, or word count with automated Originality scores, it is my hope that creativity researchers can apply the methodological approaches used in this study (i.e., IDF weighted GloVe scores) to their own work to maintain better control of the influence word count often has on automated Originality scores. There is undoubtedly further psychometric work to be done regarding automatic Originality scoring procedures, but the evidence here suggests that even when the variance attributed to Elaboration is partialled out, this method is capable of providing reliable Originality scores. Although the evidence presented here is not enough proof to guarantee that the methods used in this study will diminish the elaboration-confound in all contexts in the future, it is my estimation that this information

is meaningful and can help push creativity researchers in the right direction while trying to achieve reliable Originality scores from text-mining models.

As Guilford (1968) pointed out, “Most of our problem solving in everyday life involves divergent thinking” (p. 8), likely because divergent thinking often leads to Originality, and thus creative ideas that find solutions to those problems. Similarly, Kaufman and colleagues (2008) explained, “When problem solving, the divergent thinker simply has a fuller cognitive toolbox from which to pull potential solutions, which from a statistical perspective suggests a greater chance of solving a problem than someone with fewer, less original ideas” (p. 17). This idea suggests that DT and Originality are critical for creative thinking and advantageous for the cognitive processes that facilitate an individual’s ability to innovate around obstacles and challenges. Therefore, it is important that creativity researchers develop an empirical understanding of characteristics such as these and establish an optimal approach to measuring them. Once this is achieved, it will be easier to foster creativity among students, employees, and many others, so they are more equipped to pose innovative and effective solutions to problems. As a result, humans will be more prepared and capable to overcome unprecedented challenges that may arise in the future.

REFERENCES

Acar, S., & Runco, M. A. (2014). Assessing associative distance among ideas elicited by tests of divergent thinking. *Creativity Research Journal, 26*(2), 229–238. https://doi.org/10.1080/10400419.2014.901095

Acar, S., & Runco, M. A. (2015). Thinking in multiple directions: Hyperspace categories in divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts, 9*, 41–53. http://doi.org/10.1037/a0038501

Acar, S., & Runco, M. A. (2019). Divergent thinking: New methods, recent research, and extended theory. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 153–158. https://doi.org/10.1037/aca0000231

Beaty, R. E., Silvia, P. J., Nusbaum, E. C., Jauk, E., & Benedek, M. (2014). The roles of associative and executive processes in creative cognition. *Memory and Cognition, 42*, 1186–1197. https://doi.org/10.3758/s13421-014-0428-8.

Besemer, S. P., & O'Quin, K. (1987). Creative product analysis: Testing a model by developing a judging instrument. In S.G. Isaksen (Ed.) *Frontiers of creativity research: beyond the basics*. Buffalo: Bearly Limited.

Davis, G. A. (1989). Testing for creative potential. *Contemporary Educational Psychology, 14*, 257–274. http://doi.org/10.1016/0361-476X(89)90014-3

Dumas, D. (2018). Relational reasoning and divergent thinking: An examination of the threshold hypothesis with quantile regression. *Contemporary Educational Psychology, 53*, 1–14. https://doi.org/10.1016/j.cedpsych.2018.01.003

Dumas, D., & Dunbar, K. N. (2014). Understanding fluency and originality: A latent variable perspective. *Thinking Skills and Creativity, 14*, 56–67. https://doi.org/10.1016/j.tsc.2014.09.003

Dumas, D., & Dunbar, K. N. (2016). The creative stereotype effect. *PLOS ONE, 11*(2), https://doi.org/10.1371/journal.pone.0142567

Dumas, D., Organisciak, P., & Doherty, M. (2020). Measuring divergent thinking originality with human raters and semantic network algorithms: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts.*

Dumas, D., & Runco, M. (2018). Objectively scoring divergent thinking tests for originality: A re-analysis and extension. *Creativity Research Journal, 30*(4), 466–468.

Dumas, D., & Strickland, A. L. (2018). From book to bludgeon: A closer look at unsolicited malevolent responses on the alternate uses task. *Creativity Research Journal, 30*(4), 439–450. https://doi.org/10.1080/10400419.2018.1535790

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the intelligent essay assessor. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions*. (pp. 68–88). New York, NY: Routledge/Taylor & Francis Group. (2013-15323-005).

Forster, E. A., & Dunbar, K. N. (2009). Creativity evaluation through latent semantic analysis. In N. A. Taatgen, & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 602–607). Austin, TX: Cognitive Science Society.

Forthmann, B., Gerwig, A., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2016). The be-creative effect in divergent thinking: The interplay of instruction and object frequency. *Intelligence, 57*, 25–32. https://doi.org/10.1016/j.intell.2016.03.005.

Forthmann, B., Holling, H., Çelik, P., Lubart, T., & Storme, M. (2017). Typing speed as a confounding variable and the measurement of quality in divergent thinking. *Creativity Research Journal, 29*(3), 257–269. http://doi.org.du.idm.oclc.org/10.1080/10400419.2017.1360059

Forthmann, B., Oyebade, O., Ojo, A., Günther, F., & Holling, H. (2019). Application of latent semantic analysis to divergent thinking is biased by elaboration. *The Journal of Creative Behavior, 53*(4), 559–575. http://doi.org.du.idm.oclc.org/10.1002/jocb.240

Forthmann, B., Szardenings, C., & Holling, H. (2020). Understanding the confounding effect of fluency in divergent thinking scores: Revisiting average scores to quantify artifactual correlation. *Psychology of Aesthetics, Creativity, and the Arts, 14*(1), 94–112. https://doi-org.du.idm.oclc.org/10.1037/aca0000196.supp

Golub, G. H. & Kahan, W. M. (1965). Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics, 2*, 205–224. http://dx.doi.org/10.1137/0702016

Graesser, A. C., Hu, X., & McNamara, D. S. (2013). Computerized learning environments that incorporate research in discourse psychology, cognitive science, and computational linguistics. In A.F. Healy (Ed.), *Experimental cognitive psychology and its applications: Festschrift in Honor of Lyle Bourne,*

Walter Kintsch, and Thomas Landauer. Washington, DC: American Psychological Association.

Graham J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability what they are and how to use them. *Educ. Psychol. Meas. 66*, 930–944. 10.1177/0013164406288165

Gray, K., Anderson, S., Chen, E. E., Kelly, J. M., Christian, M. S., Patrick, J., Huang, L., Kenett, Y. N., & Lewis, K. (2019). "Forward flow": A new measure to quantify free thought and predict creativity. *American Psychologist 74*(5) 539–554. https://doi.org/10.1037/amp0000391

Guilford, J. P. (1950). Creativity. *American Psychologist, 5*, 444–454. https://doi.org/10.1037/h0063487.

Guilford, J. P. (1956). The structure of intellect, *Psychological Bulletin 53*(4), 267–293. https://doi-org.du.idm.oclc.org/10.1037/h0040755

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

Guilford, J. P. (1968). *Creativity, intelligence, and their educational implications*. San Diego, CA: EDITS/Knapp.

Günther, F., Dudschig, C., & Kaup, B. (2015). LSAfun-An R package for computations based on Latent Semantic Analysis. *Behavior Research Methods, 47*(4), 930–944. https://doi.org/10.3758/s13428-014-0529-0

Gwet, K. L. (2014). Intrarater Reliability. Wiley StatsRef: Statistics Reference Online.

Hass, R. W. (2016a). Conceptual expansion during divergent thinking. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th annual*

conference of the cognitive science society (pp. 996–1001). Austin, TX: Cognitive Science Society.

Hass, R.W. (2016b). Tracking the dynamics of divergent thinking via semantic distance: Analytic methods and theoretical implications. *Memory and Cognition, 45*, 233–244 Advanced online publication. https://doi.org/10.3758/s13421-016-0659-y

Hass, R. W. (2017). Semantic search during divergent thinking. *Cognition, 166*, 344–357. http://dx.doi.org/10.1016/j.cogntion.2017.05.039.

Harbison, J. I., & Haarmann, H. (2014). Automated scoring of originality using semantic representations. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society (COGSCI)* (pp. 2337–2332). College Park, MD. Retrieved from https://mindmodeling.org/cogsci2014/papers/405/paper405.pdf

Harrington, D. M. (1975). Effects of explicit instructions to "be creative" on the psychological meaning of divergent thinking test scores. *Journal of Personality, 43*(3), 434–454. https://doi-org.du.idm.oclc.org/10.1111/j.1467-6494.1975.tb00715.x

Harris, Z. (1954). Distributional structure. *Word, 10*, 146–162. https://doi.org/10.1080/00437956.1954.11659520.

Heinen, D. J. P., & Johnson, D. R. (2018). Semantic distance: An automated measure of creativity that is novel and appropriate. *Psychology of Aesthetics, Creativity, and the Arts, 12*(2), 144–156. https://doi.org/10.1037/aca0000125

Henrich, J., Heine, S., & Norenzayan, A. (2010). Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral and Brain Sciences, 33*, 2-3, 111–135. doi:10.1017/S0140525X10000725

He, Q., von Davier, M., Greiff, S., Steinhauer, E. W., & Borysewicz, P. B. (2017). Collaborative problem solving measures in the Programme for International Student Assessment (PISA). In A. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), *Innovative assessment of collaboration*. (pp. 95–111). https://doi.org/10.1007/978-3-319-33261-1_7

Hocevar, D. (1980). Intelligence, divergent thinking, and creativity. *Intelligence, 4*, 25–40.

Hornberg, J., & Reiter-Palmon, R. (2017). Creativity and the big five personality traits: Is the relationship dependent on the creativity measure? In G. Feist, R. Reiter-Palmon, & J. Kaufman (Eds.), *The Cambridge handbook of creativity and personality research* (pp. 275–293). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/ 9781316228036.015

Hudson, L. (1968). *Frames of mind: Ability, perception and self-perception in the arts and sciences*. Oxford England: W. W. Norton.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi-org.du.idm.oclc.org/10.1080/10705519909540118

Jellen, H. G., & Urban, K. K. (1986). The TCT-DP (Test for Creative Thinking-Drawing Production): An instrument that can be applied to most age and ability groups. *Creative Child & Adult Quarterly, 11*(3), 138–155.

Jones, M. N., Willits, J. A., & Dennis, S. (2015). Models of semantic memory. In J. R. Busemeyer & J. T. Townsend (Eds.), *Oxford handbook of mathematical and computational psychology* (pp. 232–254). New York, NY: Oxford University Press.

Kaufman, J. C., Plucker, J. A., & Baer, J. (2008). *Essentials of creativity assessment*. https://ebookcentral.proquest.com

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological Methods, 21*(4), 507–525. http://dx.doi.org/10.1037/met0000091

Kintsch, W., & Bowles, A. R. (2002). Metaphor comprehension: What makes a metaphor difficult to understand? *Metaphor and Symbol, 17*(4), 249–262. https://doi.org/10.1207/S15327868MS1704_1

Kjell, O. N. E., Kjell, K., Garcia, D., & Sikström, S. (2019). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods, 24*(1), 92–115. https://doi.org/10.1037/met0000191

Kuhn, J. T., & Holling, H. (2009). Measurement invariance of divergent thinking across gender, age, and school forms. *European Journal of Psychological Assessment*, *25*(1), 1–7. https://doi.org/10.1027/1015-5759.25.1.1

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. *Proceedings of the 19th Annual Meeting of the Cognitive Science Society*, 412–417.

Lemaire, B., & Denhière, G. (2004). Incremental construction of an associative network from a corpus. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (pp. 825-830). Austin, TX: Cognitive Science Society.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics, 20*, 1–31.

Lord, F. M. (2012). *Applications of item response theory to practical testing problems.* Routledge.

Lubart, T. I., Besançon, M., & Barbot, B. (2011). Evaluation du potentiel créatif (EPoC). Editions Hogrefe France, Paris.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and

counting: A review and empirical validation. *Journal of Memory and Language, 92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Martin, D. I., & Berry, M. W. (2007). Mathematical Foundations Behind Latent Semantic Analysis. In T.K. Landauer, D.S. McNamara, S. Dennis & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 35–56). Mahwah, NJ: Erlbaum.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. https://doi.org/10.1037/met0000144

McNeish, D., An, J., & Hancock, G. R. (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. *Journal of Personality Assessment, 100*(1), 43–52. https://doi.org/10.1080/00223891.2017.1281286

Mednick, S. A. (1962). The associative basis of the creative process. Psychological Review, 69, 220–232.

Mednick, S. A., (1968). Remote associates test. *Journal of Creative Behavior, 2*, 213–214. https://doi.org/10.1002/j.2162-6057.1968.tb00104.x

Mehmetoglu, M. (2015). STATA 15.1: "Relicoef" module [STATA software]. http://fmwww.bc.edu/RePEc/bocode/r

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling, 2*, 255–273.

Naili, M., Chaibi, A., & Ben Ghezala, H. (2017). Comparative study of word embedding methods in topic segmentation. *Procedia Computer Science, 112*, 340-349. https://doi.org/10.1016/j.procs.2017.08.009

Neuman, Y., & Cohen, Y. (2014). A vectorial semantics approach to personality assessment. Scientific Reports, 4, 4761. http://dx.doi.org/10.1038/srep04761

Organisciak, P. (2016). Term Weights for 235k Language and Literature Texts [Data set]. Retrieved from https://www.ideals.illinois.edu/handle/2142/89691

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. P. (2014). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934-952. DOI: 10.1037/pspp0000020

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Stanford, CA. Retrieved from https://nlp.stanford.edu/pubs/glove.pdf

Penumatsa, P., Ventura, M., Graesser, A. C., Louwerse, M., Hu, X., Cai, Z., & Franceschetti, D. R., The Tutoring Research Group (2006). The right threshold value: What is the right threshold of cosine measure when using latent semantic analysis for evaluating student answers*? International Journal of Artificial Intelligence, 15*, 767–777.

Plucker, J. A. (1999). Is the proof in the pudding? Re-analyses of Torrance's (1958 to present) longitudinal data. *Creativity Research Journal, 12*(2), 103–114. https://doi.org/10.1207/s15326934crj1202_3

Plucker, J. A., & Makel, M. C. (2010). Assessment of creativity. In J. C. Kaufman & R. J. Sternberg (Eds.), *The Cambridge handbook of creativity*. (pp. 48–73). New York,

NY US: Cambridge University Press. Retrieved from https://ebookcentral.proquest.com/lib/du/reader.action?docID=585336

Prabhakaran, R., Green, A. E. & Gray, J. R. (2014). Thin slices of creativity: Using single-word utterances to assess creative cognition. *Behavior Research Methods 46*, 641–659. https://doi.org/10.3758/s13428-013-0401-7

Purser (Eds.), *Social creativity* (Vol. 1, pp. 237–264). Cresskill, NJ: Hampton.

Puryear, J. S., Kettler, T., & Rinn, A. N. (2017). Relationships of personality to differential conceptions of creativity: A systematic review. *Psychology of Aesthetics, Creativity, and the Arts, 11*(1), 59–68. https://doi.org/10.1037/aca0000079

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*(2), 173-184. https://doi-org.du.idm.oclc.org/10.1177/01466216970212006

Rakib, R. H., Islam, A., & Milios, E. (2018). Improving text relatedness by incorporating phrase relatedness with word relatedness. *Computational Intelligence, 34*(3), 939–966. http://doi.org.du.idm.oclc.org/10.1111/coin.12152

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods, 41*, 647–656. https://doi.org/10.3758/BRM.41.3.647

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.

Reiter-Palmon, R. (2018). Creative cognition at the individual and team level: What happens before and after idea generation. In R. Sternberg & J. Kaufman (Eds.), *The nature of human creativity* (pp. 184–208). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/ 9781108185936.015

Reiter-Palmon, R., Forthmann, B., & Barbot, B. (2019). Scoring divergent thinking tests: A review and systematic framework. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 144–152. http://doi.org/10.1037/aca0000227

Runco, M. A. (1999). Creativity need not be social. In A. Montuori & R.

Runco, M. A. (2007). *Creativity: Theories and themes: Research, development, and practice.* Academic Press.

Runco, M. A., Abdulla, A. M., & Paek, S.-H. (2016). Which test of divergent thinking is best? Creativity: *Theories-Research-Applications, 3*, 4–18. https://doi.org/10.1515/ctra-2016-0001

Runco, M. A. & Acar, S. (2012). Divergent thinking as an indicator of creative potential. *Creativity Research Journal, 24*(1), 66–75. https://doi.org/10.1080/10400419.2012.652929

Runco, M. A. & Albert, R. S. (1985). The reliability and validity of ideational originality in the divergent thinking of academically gifted and nongifted children. *Educational and Psychological Measurement, 45*(3), 483–501. https://doi.org/10.1177/001316448504500306

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal, 24*(1), 92–96. https://doi.org/10. 1080/10400419.2012.650092

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

Sijtsma K. (2009). On the use, the misuse, and the very limited usefulness of cronbach's alpha. *Psychometrika, 74*(1), 107–120. https://doi.org/10.1007/s11336-008-91010

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., … Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts, 2*(2), 68–85. https://doi-org.du.idm.oclc.org/10.1037/1931-3896.2.2.68

Simonton, D. K. (2018). Defining creativity: Don't we also need to define what is not creative? The Journal of Creative Behavior, 52(1), 80–90. https://doi.org/10.1002/jocb.137

Taylor, C. W. (1988). Various approaches to and definitions of creativity. In R. J. Sternberg (Ed.), *The nature of creativity: Contemporary psychological perspectives* (pp. 99–121). England: Cambridge

Torrance, E. P. (1962). *Guiding creative talent*. Englewood Cliffs, NJ: Prentice-Hall.

Torrance, E. P. (1970). *Encouraging creativity in the classroom*. Dubuque, IA: William C. Brown Company Publishers.

Torrance, E. P. (1972). Predictive validity of the Torrance Tests of Creative Thinking. *The Journal of Creative Behavior, 6*(4), 236–252.

Torrance, E. P. (1988). The nature of creativity as manifest in its testing. In Sternberg, R. J. (Ed.), *The Nature of Creativity*, 43–75. New York, NY: Cambridge University Press.

Torrance, E. P. (1995). *Why fly?* New York, NY: Greenwood Publishing Group.

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

Torrance, E.P. (1999*). Torrance Test of Creative Thinking: Norms and technical manual.* Beaconville, IL: Scholastic Testing Services.

Wallach, M. A., & Kogan, N. (1965). *Modes of thinking in young children*. New York: Holt, Rinehart and Winston.

White, H. A., & Shah, P. (2016). Scope of semantic activation and innovative thinking in college students with ADHD. *Creativity Research Journal, 28*(3), 275–282. https://doi.org/10.1080/10400419.2016.1195655

Wilson, R.C., Guilford, J., & Christensen, P.R. (1953). The measurement of individual differences in originality. *Psychological Bulletin, 50*, 362–370. https://doi.org/10.1037/h0060857.

von Davier, A. A. (2017). Computational psychometrics in support of collaborative educational assessments. *Journal of Educational Measurement, 54*(1), 3–11. https://doi.org/10.1111/jedm.12129

Yoon, C. H. (2017). A validation study of the Torrance Tests of Creative Thinking with a sample of Korean elementary school students. *Thinking Skills and Creativity, 26*, 38–50. https://doi.org/10.1016/j.tsc.2017.05.004

Zarnegar, Z., Hocevar, D., & Michael, W. B. (1988). Components of original thinking in gifted children. *Educational and Psychological Measurement, 48*. 5–16. https://doi-org.du.idm.oclc.org/10.1177/001316448804800103

APPENDICES

Appendix A – Tables 1-3

Table 1.
Means and standard deviations for Originality scores and Elaboration on each AUT object-prompt and R-squared coefficients from each of the 10 linear regressions.

AUT Object-Prompt	*M*	*SD*	*M*	*SD*	R^2
Book	.780	.098	13.51	12.07	.005
Rope	.642	.065	13.99	12.63	.017
Fork	.757	.048	12.10	11.32	.001
Table	.681	.069	14.95	13.49	.029
Pants	.650	.086	13.64	12.80	.002
Bottle	.689	.082	15.50	15.50	.043
Brick	.719	.076	12.22	11.03	.001
Tire	.709	.077	13.23	12.33	<.001
Shovel	.696	.068	10.19	10.07	.057
Shoe	.746	.078	11.00	10.24	.001

Note: Means and standard deviations for Originality are presented on the left, followed by the means and standard deviations for Elaboration, or number of words used in response to that item. All 10 regression models were computed with the Elaboration score as the predictor and the Originality score as the outcome.

Table 2.

Model fit statistics from the CFA models for Elaboration, GloVe Originality, and residualized GloVe Originality scores from the AUT.

Model	X^2	*df*	RMSEA	CFI	SRMR
Elaboration	97.356**	35	0.139	0.932	0.042
Originality	65.887*	35	0.104	0.835	0.082
Residualized Originality	64.113*	35	0.101	0.842	0.081

Note: *p < .01, **p < .001.

Table 3.
Alternate Uses Task prompt standardized loadings from the three unidimensional confirmatory factor analyses.

	Elaboration	Originality	Residualized Originality
Book	0.778	0.672	0.689
Rope	0.844	0.192	0.176
Fork	0.807	0.217	0.22
Table	0.905	0.346	0.356
Pants	0.904	0.475	0.463
Bottle	0.815	0.791	0.771
Brick	0.852	0.635	0.644
Tire	0.891	0.488	0.492
Shovel	0.833	0.444	0.443
Shoe	0.796	0.759	0.757

Note: All standardized loadings are significant at $p < .05$ except for Rope and Fork in the Originality and Residualized Originality models.

Appendix B – Figures 1-3

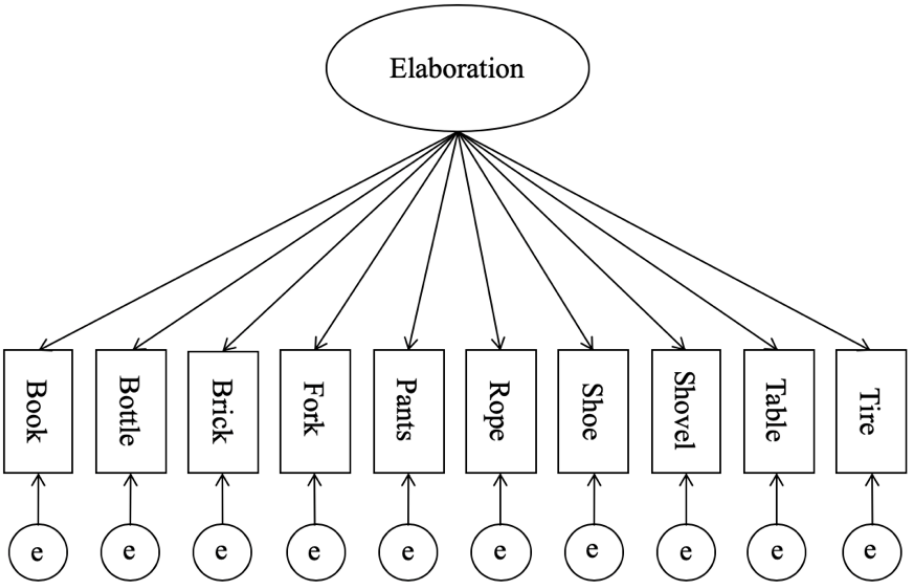

Figure 1. Conceptual path diagram of the latent measurement model used to determine the factor reliability of the Elaboration, or word count, scores from the AUT.

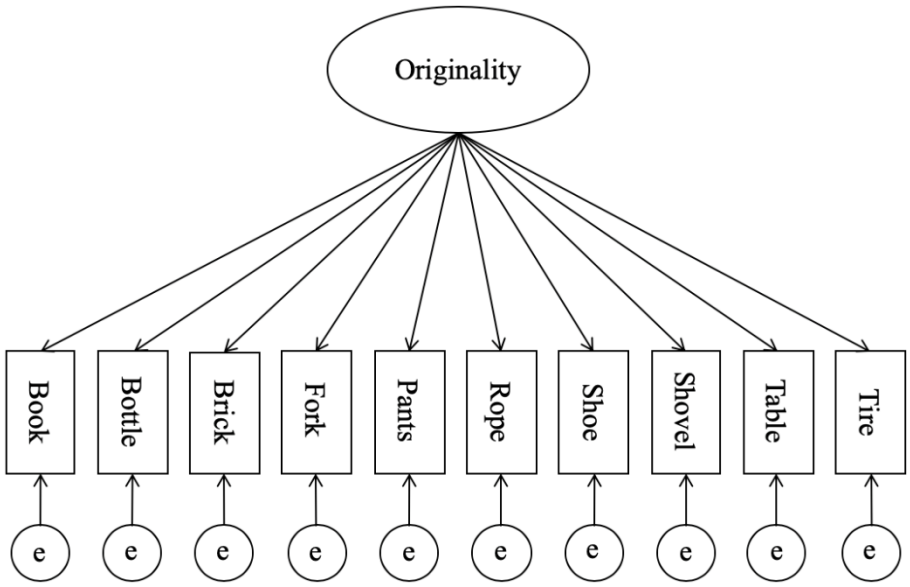

Figure 2. Conceptual path diagram of the latent measurement model used to determine the factor reliability of the GloVe Originality scores from the AUT.

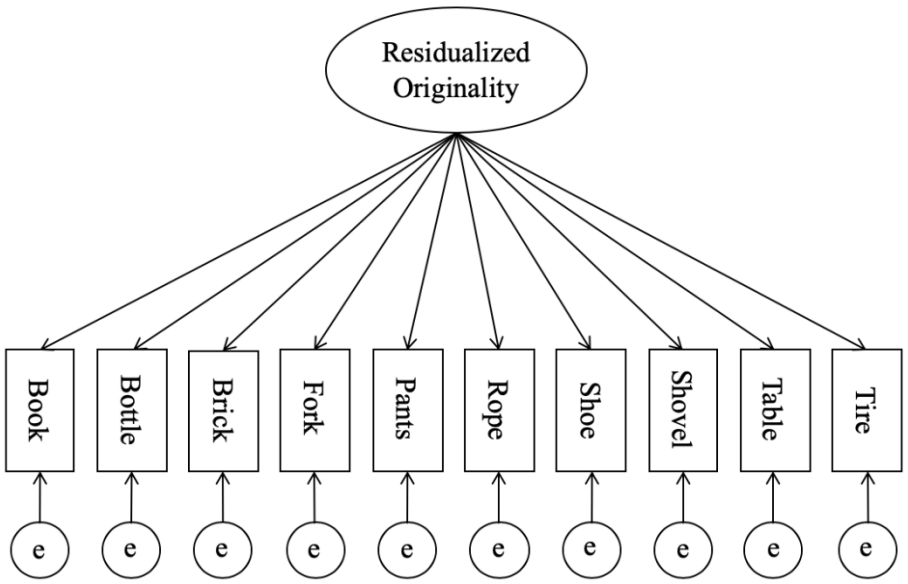

Figure 3. Conceptual path diagram of the latent measurement model used to determine the factor reliability of the residualized GloVe Originality scores from the AUT.