Journal of Applied Psychology
2016, Vol. 101, No. 7, 958–975

0021-9010/16/$12.00 http://dx.doi.org/10.1037/apl0000108

Initial Investigation Into Computer Scoring of Candidate Essays for Personnel Selection

Michael C. Campion
University of South Carolina

Michael A. Campion
Purdue University

Emily D. Campion
University at Buffalo, The State University of New York

Matthew H. Reider
Campion Services, Inc., West Lafayette, Indiana

Emerging advancements including the exponentially growing availability of computer-collected data and increasingly sophisticated statistical software have led to a "Big Data Movement" wherein organizations have begun attempting to use large-scale data analysis to improve their effectiveness. Yet, little is known regarding how organizations can leverage these advancements to develop more effective personnel selection procedures, especially when the data are unstructured (text-based). Drawing on literature on natural language processing, we critically examine the possibility of leveraging advances in text mining and predictive modeling computer software programs as a surrogate for human raters in a selection context. We explain how to "train" a computer program to emulate a human rater when scoring accomplishment records. We then examine the reliability of the computer's scores, provide preliminary evidence of their construct validity, demonstrate that this practice does not produce scores that disadvantage minority groups, illustrate the positive financial impact of adopting this practice in an organization ($N \sim$ 46,000 candidates), and discuss implementation issues. Finally, we discuss the potential implications of using computer scoring to address the adverse impact-validity dilemma. We suggest that it may provide a cost-effective means of using predictors that have comparable validity but have previously been too expensive for large-scale screening.

Keywords: adverse impact, Big Data, personnel selection, statistics, test scoring

Recent shifts in technology have staged selection scholarship on the precipice of a revolution. On the one hand, enhanced computing power and increased use of technology in organizations has afforded researchers access to an exponentially greater amount of information on candidates. At the same time, technological innovations such as the development of new software packages have emerged enabling researchers to develop more sophisticated methods of analyzing such massive amounts of data. These shifts, in combination, have led to the "Big Data Movement" and created conditions under which previously agreed upon assumptions can be reexamined and, perhaps, overturned (Kuhn, 1962). This trend is just beginning to be recognized in Industrial and Organizational (I/O) Psychology (Guzzo & Carlisle, 2014; Poeppelman, Blacksmith, & Yang, 2013).

One example of such an assumption is that which underlies a long-standing issue in selection research: the adverse impact-validity dilemma (Ployhart & Holtz, 2008; Pyburn, Ployhart, & Kravitz, 2008). While recognizing that a number of conditions can lead to adverse impact (e.g., when there are differences in predictor true scores by protected class membership), this dilemma is also, in part, attributable to the assumption that there are substantial tradeoffs in costs associated with selection procedures. On the one hand, cost-effective procedures with high criterion-related validity (e.g., cognitive ability tests) exhibit high levels of adverse impact. On the other hand, alternatives that also exhibit high levels of validity but less adverse impact (e.g., structured interviews, assessment centers, work samples, and accomplishment records) are too costly to administer to large numbers of candidates.

The purpose of this paper is to critically investigate whether recent advances in text mining and predictive modeling software can be used as a more economic surrogate for human raters to score candidate responses to essays used for hiring. Drawing on prior research on natural language processing (NLP), we pose a number of research questions, which we then examine using a sample of accomplishment records (ARs; Hough, 1984) from nearly 46,000 candidates as part of an operational assessment battery in a large organization. If successful, this research may have implications for reducing the cost-effectiveness tradeoff underlying the adverse impact-validity dilemma.

This paper contributes to the literature on personnel selection and, more broadly, scholarship in applied psychology in two primary ways. First, we show how recent advances in computer

This article was published Online First April 14, 2016.

Michael C. Campion, Moore School of Business, University of South Carolina; Michael A. Campion, Krannert School of Business, Purdue University; Emily D. Campion, School of Management, University at Buffalo, The State University of New York; Matthew H. Reider, Campion Services, Inc., West Lafayette, Indiana.

Correspondence concerning this article should be addressed to Michael C. Campion, Moore School of Business, University of South Carolina, Columbia, SC 29208. E-mail: michael.campion@grad.moore.sc.edu

technology can be applied within the domain of I/O Psychology. Specifically, we explain how to program or "train" a computer to emulate a human when scoring ARs. The focal topic of examination in this paper lies at the intersection of psychology and advances in predictive modeling computer analytics. As such, examining this process is important as it allows us to shed light on how information gathered regarding a judgment process that is largely cognitive and behavioral is assigned structure and meaning by a computer program and used to produce scores for future applicants. It also enables us to explain how scholars can develop more sophisticated methods of analyzing data that includes a qualitative (text-based) component.

Second, we identify the advantages and potential disadvantages associated with attempting to use text mining and a predictive modeling computer program to score ARs, which have been historically shown to exhibit low levels of adverse impact but high validity. For example, we demonstrate that the computer program can exhibit a level of reliability comparable to that of a human rater when scoring the content of ARs. We also provide preliminary evidence of construct validity for the computer scores by relating them to a wide range of other variables (e.g., other tests). Then, we demonstrate that this practice does not produce scores that disadvantage any minority demographic group, and we illustrate the positive financial impact of adopting this practice in the organizational context under investigation. In terms of potential disadvantages, we investigate likely reasons for scores we refer to as "extreme mis-predictions," and discuss practical issues with implementation, such as what to do if AR questions change.

Prior Research on the Computer Scoring of Essays

The literature on computer scoring of essays is considered part of a research stream that investigates what is referred to as either Automated Essay Scoring (AES) or Computer-Automated Scoring (CAS). Many of the systems currently used to score essays were crafted in the late 1990s and early 2000s (Dikli, 2006; Valenti, Neri, & Cucchiarelli, 2003). The computer scoring of essays has been widely researched and used in educational contexts to more efficiently score large numbers of essays that assess the writing skills of students for a range of purposes, such as feedback in the classroom (Dikli, 2006), evaluation of educational systems (Ben-Simon & Bennett, 2007; Leacock & Chodorow, 2003), and selection into college and graduate school (Attali, Lewis, & Steier, 2013; Dikli, 2006; Yang, Buckendahl, Juszkiewicz, & Bhola, 2002). One example illustrating the level of interest on this topic is the Hewlett Foundation's recent sponsorship of a competition for software developers, enlisting them to improve automated scoring of student essays (https://www.kaggle.com/c/asap-aes).

Most AES programs have focused on writing quality and essay structure using NLP techniques such as latent semantic analysis (LSA; Attali, Bridgeman, & Trapani, 2010), Bayesian networks (Rudner, & Liang, 2002), and simpler systems that use descriptive statistics and multiple regression (Valenti et al., 2003). Dikli (2006) and Shermis, Burstein, Higgins, and Zechner (2010) provide reviews of the various proprietary AES approaches.

AES programs can score writing skill at a level of reliability comparable to that of a human grader, but they are often limited in their ability to evaluate content and other higher order skills (Attali et al., 2013; Deane, 2013; Ramineni & Williamson, 2013; Shermis et al., 2010). As such, researchers have begun searching for ways to create programs that can score the content of essays. However, this has proved to be more difficult as vocabulary and sentence structure are not proxies for the quality of the content (Attali et al., 2013).

Historical Approaches to Text Mining and Natural Language Processing (NLP)

Perhaps the greatest hindrance to using computers to score the narrative responses of candidates thus far has been a lack of software development. Although computing power has vastly improved over the past decades, programs have, until recently, not been developed that are capable of leveraging it to transform the large quantity of text-based data available into meaningful (i.e., reliable and valid) scores. In the context of narrative responses, a critical factor that initially constrained this effort was the "information retrieval problem" (Dumais, 2004). The information retrieval problem refers to the fact that precise lexical matching between words in a user's query and words used within documents often does not exist. For example, a rater might be interested in inducing a candidate's leadership skill. However, the terms "leader" or "leadership" may not have been used by the candidate. Instead, he or she may have used the term "manager." This mismatch occurs as a result of fundamental characteristics of human word usage. Specifically, humans often use a variety of words to describe the same concept or object (i.e., synonymy) or the same word to refer to different things (i.e., polysemy; Dumais, 2004).

Over the years, numerous information retrieval software programs have been developed attempting to solve this problem. For example, software has been developed that uses "stemming," which essentially breaks down one's search query into its root form. Thus, using the example above, the term "leadership" would be converted to the term "leader." Similarly, software has been developed that uses "controlled vocabularies," which reduce the query and index terms to a specific set of terms. As an example, scholarly journals often use these programs, which supply a list of keywords to be used for articles to reflect their content.

As the quantity of text-based data available became greater, the demands increased for better text mining software. As a result of this, the field of NLP expanded its focus to not only include issues related to information retrieval, but also information extraction. Presently, information extraction is the primary emphasis in text mining (Gaizauskas & Wilks, 1998). Indeed, *text mining* is generally defined as the extraction of meaningful (and previously unknown) information or knowledge structures from unstructured text-based data (see Chowdhury, 2003).

As text mining has become increasingly emphasized in scholarship on NLP, numerous programs have been developed that are capable of utilizing this extracted information for a range of practical purposes (e.g., prediction of attitudes, beliefs, sentiments, and other higher order characteristics; Attali et al., 2013; Liu, Lieberman, & Selker, 2003; Liu & Maes, 2004; Padmaja & Fatima, 2013). In addition to varying in terms of their purposes, software programs also tend to vary widely in terms of the amount of "supervised learning" they require. For example, programs using LSA tend more toward requiring very little training as they rely entirely on probabilistic and/or statistical techniques to mine text (Dumais, 2004; Hofmann, 2001). Meanwhile, others, such as the program used in the present study are considered "semi-supervised" in the manner in which they learn. They make use of

previously humanly constructed resources such as dictionaries and thesauri and provide the user with the option of making additional manipulations to the knowledge structures they extract.

Using Text Mining and a Predictive Modeling Computer Software Program to Score Accomplishment Records

Accomplishment Records

ARs are a personnel selection technique that requires job candidates to provide a narrative description of an accomplishment they have achieved in the past that demonstrates they have a competency necessary to perform the job for which they have applied (Hough, 1984). For example, ARs may ask a candidate to "Describe an accomplishment that demonstrates you have leadership skills." The candidate is then required to respond with a 200-word description of such an accomplishment. They are also often required to provide the name and contact information of someone who can verify the accomplishment (e.g., past supervisor or coworker). The verifying reference may or may not be contacted by the organization. Meta-analytic summaries support the high validity and low adverse impact of ARs. For example, McDaniel, Schmidt, and Hunter (1988) report an uncorrected validity of .24 and corrected validity of .45 in their summary of 15 studies (total $n = 1{,}148$). They described the AR as a "point method (behavioral consistency type)" because that was the term used by the Office of Personnel Management, which used the method extensively. Hough, Oswald, and Ployhart (2001) report a *d*-score difference between Minorities and Whites of .24 for ARs, which was much lower than those they reported for mental ability tests for Blacks and Whites (1.0) and Hispanics and Whites (.50).

Translating Accomplishment Record Data into Statistical Relationships and Output

Raters, being human, conform to principles of NLP in that they automatically recognize and reduce synonymy and polysemy within text-based responses. This, coupled with the fact that they are generally trained in some manner, renders them capable of not only understanding candidates' responses to narrative questions such as ARs, but also drawing inferences useful in making judgments regarding their characteristics. Over time, data regarding candidate responses and raters' scores can be accumulated in organizations. Text mining programs that use NLP might be used to assign meaning to this data. First, they impose structure on documents and words through analyzing large pools of text or *corpuses* (Dumais, 2004; Recchia & Jones, 2009). These corpuses, in the context of the present study, refer to text-based responses to AR questions. In analyzing these corpuses, these programs statistically model the relationships among documents (such as AR responses) comprising the corpus based on the words within them while simultaneously modeling relationships between words within the documents based on their occurrence (e.g., frequency, proximity to other words).

In this way, these programs do not depend on lexical matching. Rather, they reduce the dimensionality of documents attributable to synonymy and polysemy. They build what is referred to as a term-document matrix, which shows the frequencies of occurrence of terms in the documents. Then, they attempt to discover similarity structures (also referred to as semantic similarities; Foltz, 1996) to aid the user in information extraction by identifying which terms often occur together (Liu & Singh, 2004). Finally, the subset of terms and collections of commonly co-occurring terms from the large number extracted are selected based on their ability to predict some criteria (e.g., human raters' scores), which assigns meaning (regression weights and levels of statistical significance) to them (Dumais, 2004; Landauer & Dumais, 1997). The scoring of future documents is then based on counting and weighting the terms in those documents.

Note that this is not meant to imply that these software programs truly "understand" what the candidate has written and are making inferences regarding his or her characteristics based on this information. Rather, they model inferences made previously by raters based on the terms and relationships among terms from responses to AR questions and the scores the raters assigned to the responses. In the end, one of the primary advantages of these computer programs is that they can be used to infer characteristics of individuals such as leadership skill even when the term "leader" is not used within the document (or AR response). For example, the document may refer to taking initiative and organizing people.

Software programs using NLP are becoming increasingly popular as a way to assess student essays in low-stakes classroom settings as well as high-stakes entrance exam settings (e.g., GRE, GMAT; Dikli, 2006). In terms of their use in high-stakes settings, studies have shown that scores produced by these programs (e.g., E-Rater in Powers, Burstein, Chodorow, Fowles, & Kukich, 2000; Intelligent Essay Assessor in Landauer, Laham, & Foltz, 2003) tend to correlate with human raters at levels between .70 and .86 (also see Foltz, Laham, & Landauer, 1999).

However, rather than being scored for content of what is said, essays on exams such as the GRE and GMAT are primarily scored based on quality of writing (e.g., Does the essay include proper discourse elements such as an introduction, a thesis, main ideas, and supporting arguments? Does the writer use active voice? Is he or she not overly repetitious in word use? Is the essay written in proper English exhibiting, e.g., proper grammar, spelling, sentence structure, and subject-verb agreement?; Attali & Burstein, 2006; Rudner, Garcia, & Welch, 2005). Conversely, a written narrative response in a selection context is generally not meant to enable the organization to assess only the applicants' quality of writing, if at all. Rather, it may be more important to extract information regarding an applicant's standing on a variety of job-related, latent constructs (e.g., leadership, managerial skill, critical thinking). This is derived by assessing an applicant's quality of work history based on past accomplishments and comparing that to the job requirements. Thus, the program examined in the present study is used for an altogether different purpose. It is used to assess what is being stated, rather than how it is said.

Thus, it is not entirely clear (and not yet demonstrated) whether software programs using NLP techniques to text mine and model data can produce reliable and construct-valid scores in selection contexts where the content of narratives is scored. Further, the implications of using these computer scores to make selection decisions are also unclear. For these reasons, we propose the following research questions:

Research Question 1: Can a computer program be trained to generate scores that demonstrate a level of reliability that is comparable with scores of a human rater?

Research Question 2: What is the construct validity of computer scores? Specifically, (a) how will they correlate with human scores, (b) what are the subcomponents of computer scores, (c) will computer scores exhibit the same correlations with the other selection procedures as human scores, and (d) what is the nature of any extreme differences between computer and human scores?

Research Question 3: Will scores generated by the computer be associated with the same lack of subgroup differences as those observed with human raters?

Research Question 4: What are the potential cost savings associated with using text mining and predictive modeling as a surrogate for human raters?

Research Question 5: What are the potential implementation issues associated with using computer scoring as a surrogate for human raters?

Method

Sample and Context

The organization is a large and popular federal government employer that often receives up to 15,000 or more applicants each year for professional jobs in a range of fields (e.g., management, public affairs, economics). Receiving a large number of applicants for a relatively small number of job openings (a few hundred), the need to have merit-based hiring in the government, and the fact that selection procedures with high validity that can be used with large numbers of candidates (such as employment tests) may have adverse impact, especially with very low selection ratios, means that adverse impact is a constant concern. Having a selection procedure with high validity and low adverse impact (such as an AR) is essential but had heretofore come at the high cost of rater time to score.

The sample used to program the computer included 41,429 candidates who had completed narrative responses and all the other data relevant to the selection process at this organization over a 6-year period. Note that the data reported in this study were part of a larger data collection effort. We used the sample of 41,429 candidates to both train the computer program and to test (i.e., cross-validate) the program. In addition, we further cross-validated the program on two subsequent waves of applicants (of 2,300 and 2,198 candidates). Sample sizes may vary slightly downward in various analyses due to missing data. Written ARs are part of a selection procedure in the hiring process at the organization examined. Each candidate writes a 200-word narrative response to each of six AR questions. The AR questions ask candidates to describe their past accomplishments in terms of how they relate to six competencies. In addition to the ARs, the selection procedure also includes an evaluation of the candidate's education, work experience, and other skills and background on the application for employment as well as a review of his or her scores on a test battery. A panel of three raters reviews all the information and each rater independently rates the candidate on the six competencies common to all jobs at the organization. The composite of the six ratings across the three raters becomes the total score used for selecting candidates to move on to the next stage of the hiring process. The raters were highly trained fulltime hiring staff, and the ratings were made on detailed anchored rating scales.

Measures

The primary measures used in this study were *human rater scores* and *computer scores* on the following six competencies (on a 5-point scale with 5 being high): communication skill, critical thinking, people skill, leadership skill, managerial skill, and factual knowledge. Scores are the summed composites, thus ranging from 3 to 15 for the competencies summed across the 3 raters and from 18 to 90 across all 6 competencies. Both the rater scores and the computer scores were based on the 200-word ARs for each competency, five other text fields of application information (i.e., job titles of past jobs, work duties of past jobs, past employers, special skills, and other experiences), and 154 quantitative variables (e.g., scores on several employment tests, several variables reflecting years of work experience and level of education, and a large number of dichotomously scored academic majors, which constituted most of the variables).

The computer scores were broken down to reflect the various components based on text mined variables and the component based on the quantitative variables, as described in the results. Data were also collected on the other selection procedures to examine the construct validity of the computer scores. The employment tests used as a previous hurdle in the hiring process included a professional knowledge test ($\alpha = .92$), a biodata instrument ($\alpha = .95$), an English test ($\alpha = .90$), and a written essay scored for writing skill. The onsite selection procedures used in a subsequent stage of the hiring process included a leaderless group discussion (interrater reliability = .91), a management case exercise (interrater reliability = .91), and a structured interview (interrater reliability = .92). Total scores were also computed for both the employment tests and the onsite selection procedures based on equal weighting.

Programming the Software (Conducting the Text Mining)

Programs used for information extraction are becoming increasingly available. Some are available for free, but they usually lack technical documentation, making them less usable by the ordinary researcher, or they are very simplistic in nature rendering them less valuable. However, recently, two statistical software providers familiar to researchers in I/O psychology, SPSS and SAS, have come out with packages that not only identify key terms within text, but also construct models on the relationships among the terms in order to infer higher order characteristics or constructs in a candidate (e.g., leadership skill). The present study used the SPSS-IBM Premium Modeler package (IBM, 2012). Thus, this description of the text mining process is based on this software, which might be different from other software programs.

Figure 1 shows the overall steps in the computer modeling analysis. In Step 1, the text mining occurs, producing a computer

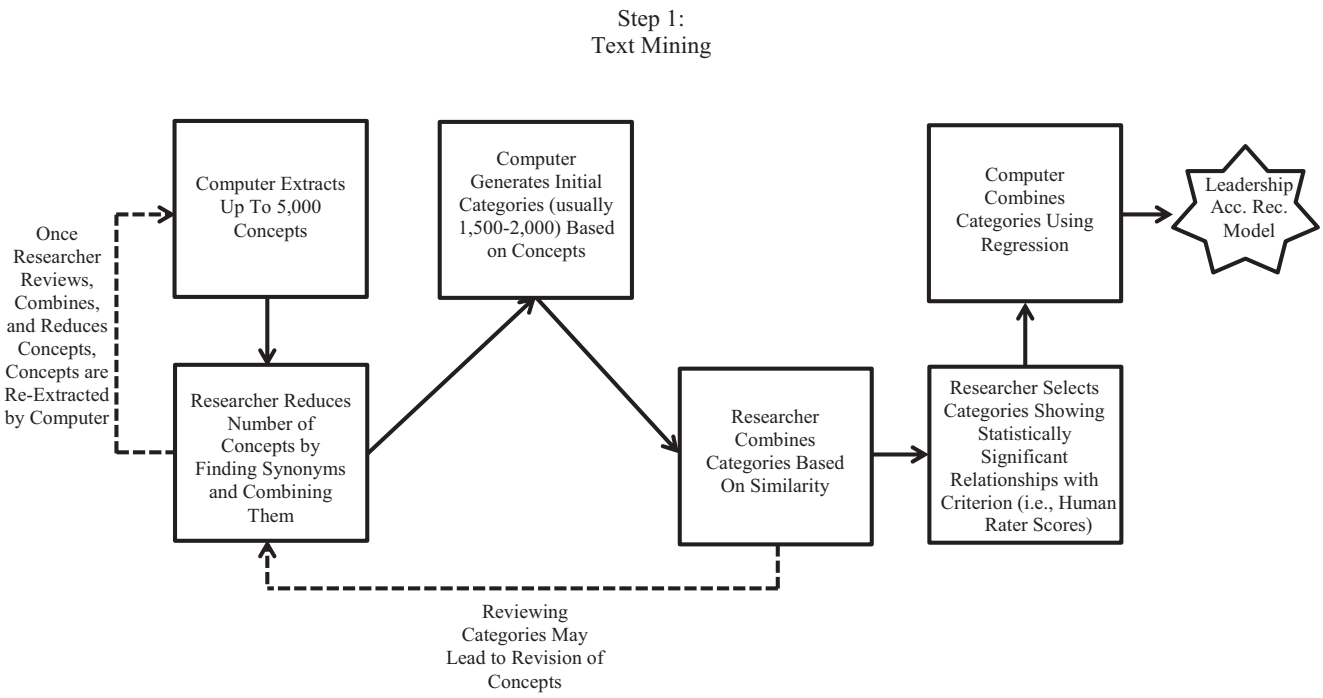

Figure 1. Overall steps in the computer modeling analysis process. Circles represent input data, hexagons represent manipulations made by the researcher, squares represent output data, and stars represent models. Content boxes include the relevant text field and all the quantitative variables.

model for each text-mined field. With this software, the programming (or "training") of the computer to read the ARs and other text-based information in the current selection procedure occurred in four substeps.

In the first substep, the program extracts the features of the essay to score. The extraction of features includes simple approaches such as surface features (e.g., specific words, number of verbs, frequency of articles, essay length, etc.), but recent advances in "text mining" using NLP allows deeper features to be identified (e.g., ordinal relationships among words, frequency of words occurring together, phrases, syntax or structure, etc.). According to the manual,

> Linguistics-based text mining . . . applies the principles of natural language processing—the computer-assisted analysis of human languages—to the analysis of words, phrases, and syntax, or structure, of text. A system that incorporates NLP can intelligently extract concepts, including compound phrases. Moreover, knowledge of the underlying language (by the researcher) allows classification of concepts into related groups, such as products, organizations, or people, using meaning and context. (IBM, 2012, p. 4, parenthetical added)

The software automatically extracts a maximum of 5,000 features it calls "concepts," from the corpus. Concepts are nouns, other terms, or phrases that occur most commonly in the ARs. This process of creating concepts involves reducing the dimensionality of the semantic space created by the usage of all terms within all candidates' ARs. *Reducing dimensionality* refers to the process by which pair-wise occurrences among terms across windows of discourse (e.g., phrases, sentences, paragraphs, documents) are used to generate vectors (i.e., to assign meaning) for terms. In this way, the program is essentially attempting to construct its own "vocabulary" for "understanding" AR responses, the result of which is a list of concepts. Extraction settings can be customized for the data. Table 1 shows the beginning of the list of concepts initially extracted for the leadership ARs. Table 1 also shows the number of times each concept appears across all documents and the number of documents in which the concept appears.

In the second substep, the researcher groups the concepts together into synonyms, including actual synonyms and terms that generally mean the same thing. The computer does not know the semantics of the concepts. In addition, its dimension reduction process that resulted in the creation of these concepts is influenced by failures of words and phrases that mean the same thing semantically to co-occur within windows of discourse within the corpus. Thus, the computer program often lists concepts separately that could be synonyms (e.g., "teamwork" and "working in teams"). The researcher reviews the concepts and groups them manually in three ways. First, the researcher groups the concepts directly. Second, the program produces "concept maps" that show the concepts that frequently occur near each other. Figure 2 shows an illustrative map for teamwork. How near concepts must be can be specified by the researcher, such as the default of within five words. Third, the researcher reads samples of specific essays of the concepts in context to make determinations of synonyms.

The process is iterative. Once many concepts have been combined, the researcher saves the data and then reextracts the concepts. The revised concepts allow for more nuanced concepts to be created. The researcher repeats the process until the set of concepts is satisfactory. Whereas the number of extracted and reextracted concepts will be up

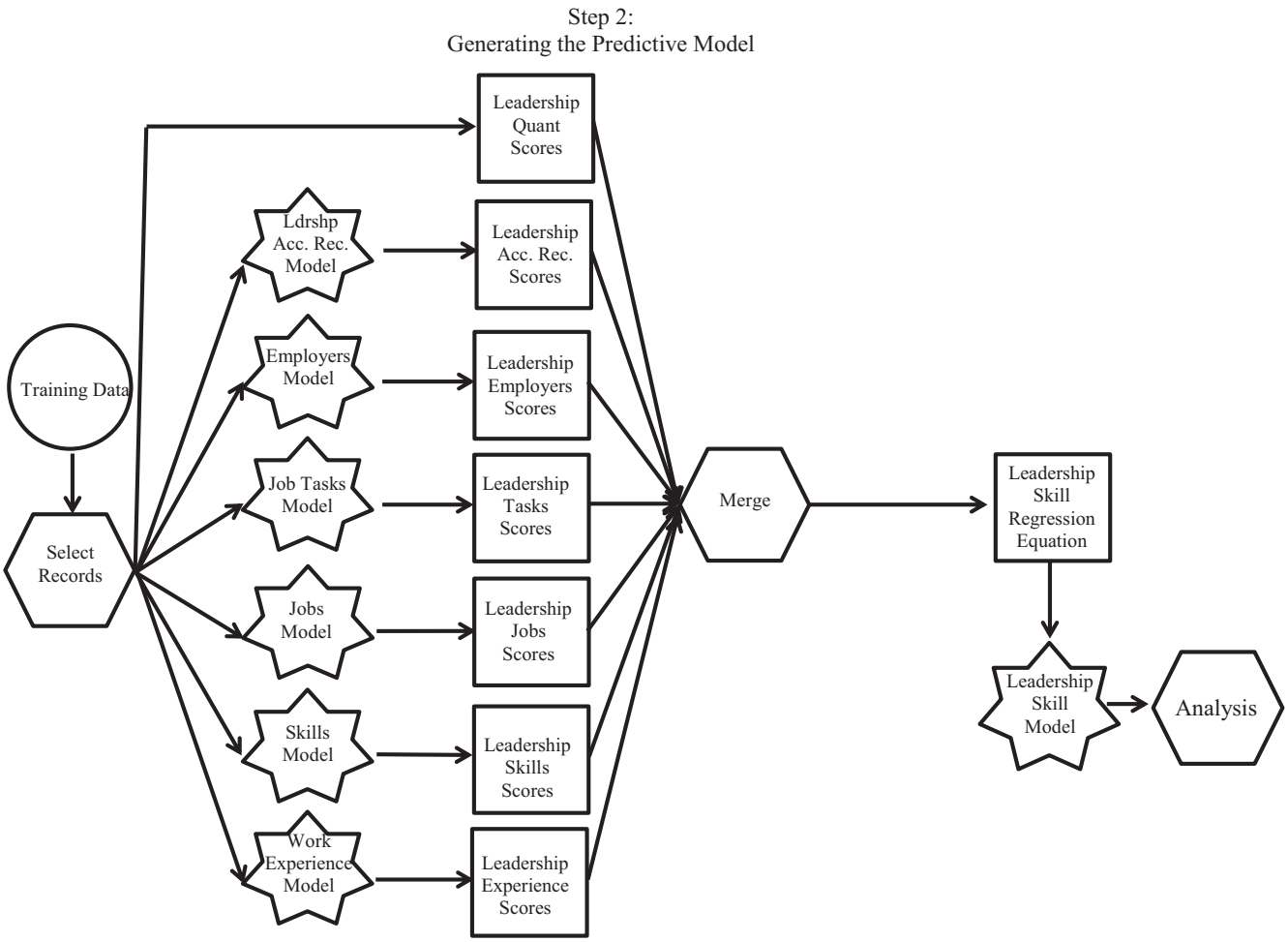

Figure 1. (*continued*).

to 5,000 with this software, the goal of this process is to identify and combine synonyms and related concepts into semantically independent concepts for categorization. Success was defined here as the top 2,000 to 2,500 concepts (those that have the highest frequencies) being nonsynonymous with each other.

In the third substep, the software generates "categories" and "subcategories," which are groups of concepts. Categories are groupings of concepts that commonly occur together and thus may have meaning semantically. The concept maps illustrated in the figure are one source of information the computer uses to help create these categories. The researcher also visually reviews the potential categories proposed by the computer program and then retains, combines, or eliminates them based on their differences, similarities, and meaningfulness, both by direct manipulation of categories and by revising the concepts to influence their relationships. Table 2 shows the beginning of the list of example categories output by the program for leadership.

In the fourth substep, the computer assigns scores to answers based on the categories present in an answer. The categories serve as dummy variables for predicting the ratings made by human raters, that is, they take on values of 1 if present or 0 if not present. Categories receive different weights depending on how well they predict the criterion (the raters' scores) using regression or similar statistical analysis. The goal is to select the best categories that are not too great in number to exceed the capacity of the statistical model (e.g., less than 1,000 here). The program offers a variety of choices of statistical models (e.g., linear regression, discriminant analysis, logistic regression, Cox regression, etc.). Linear regression was used in the current research. Table 2 shows the regression results for the beginning of the list of categories for leadership. The categories significant at $p < .05$ (bolded) were retained.

Note that some researchers have used Bayes's theorem as a statistical model to predict grader scores from essay features by computer (e.g., Frick, 1992; Rudner & Liang, 2002). The goal here is to predict the classification of the examinee, such as minimum competency or levels on a rating scale. The statistical model calculates the conditional probabilities of various scores based on the essay features, usually starting out with equal "prior" probabilities, and then updating the "posterior" probabilities based on the data. The process is iterative, starting with one feature, and then updating the probabilities as each feature is added. Once the final probabilities are computed, they can be used to score future essays.

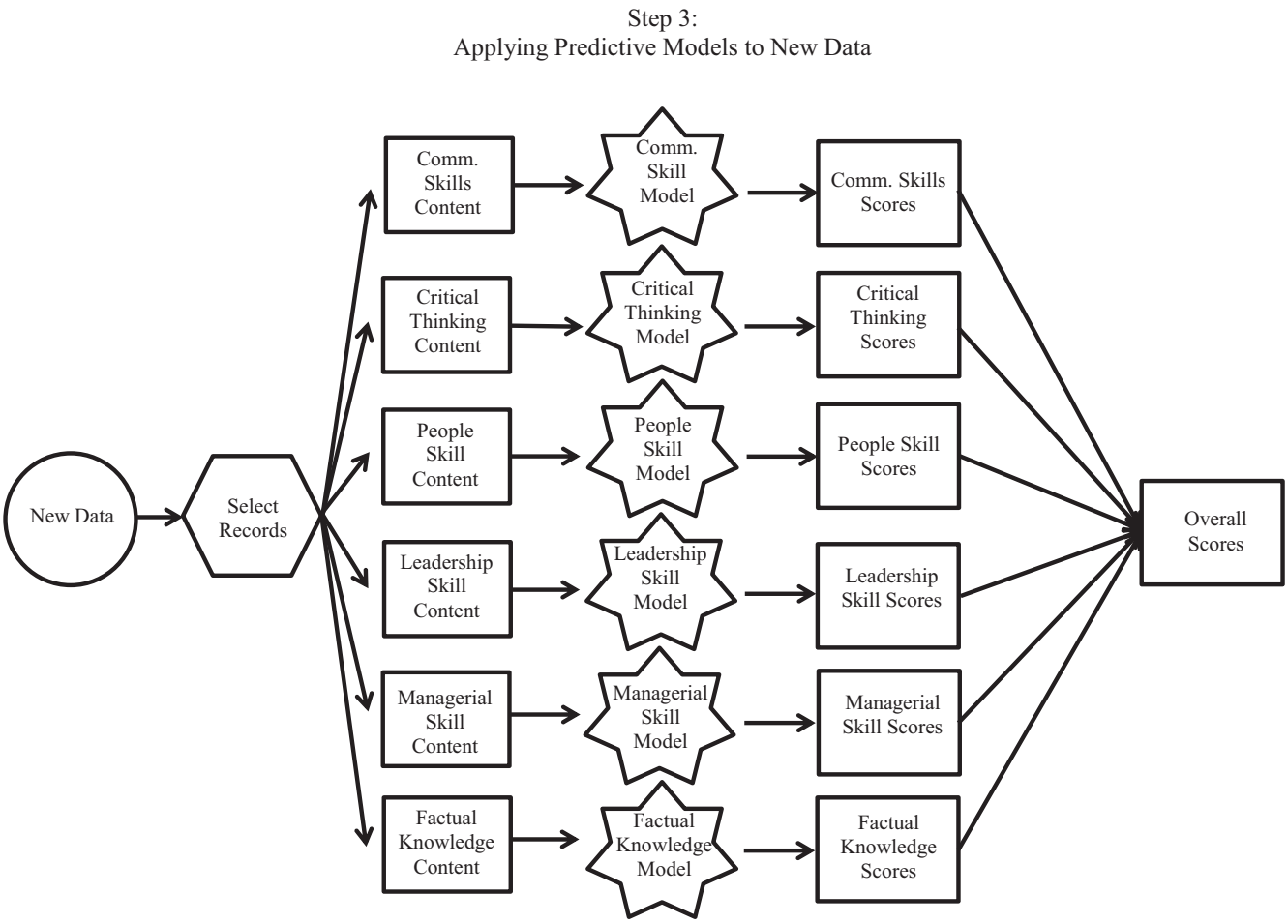

Figure 1. (*continued*).

In Step 2, the program again used regression to create an overall model for each competency (called a "predictive model") to predict the raters' scores from the text mining variables for that competency, the text mining variables for the five other text fields for that competency, and all 154 quantitative scores. This is all the information available to the raters when making the rating for each competency, thus modeling the decision-making process. Step 2 of Figure 1 shows for the process for the Leadership Skill competency. The relevant records from the data used to create the model (called the "training" data) are brought into the analysis. Then the text mining models are applied to the data (five for the five text variables and one for the relevant AR), which outputs scores on all text mined variables. The quantitative variables are also brought into the analysis at this point. Finally, all these variables are merged and a regression model is generated based on the variables that best predict the rater scores. The output is another model (titled "Leadership Skill Model"). Note there are six of these models, one for each competency.

This model is then analyzed for predictive power. The computer's estimates were correlated with the raters' scores to determine how well they were reproduced. The correlation calculated within the dataset used to create the text variables and regression models (i.e., the "training" sample) may overestimate the relationship due to capitalization on chance, so the correlation is also calculated in a separate subsample that was held out of the total sample (i.e., the "testing" sample) to cross-validate the correlation. The testing sample was 15% of our total sample, which was based on the recommendation in the software manual.

In Step 3, the process uses the predictive model to generate computer scores on new datasets. Here, records are selected from the new dataset and inputted into their respective content fields. The program then applies the model previously generated for each competency to score ARs and other text fields using the text mined categories (as well as quantitative variables), thus generating a predicted score for each competency. The total score is the sum of the six competencies, which is the same way raters' scores are combined in the actual hiring process.

In sum, in the present study the researcher played a substantial role in training the computer; however, machine learning is involved to the extent that the program counts frequencies of terms within responses (as illustrated in Table 1) and it considers probability matrices in the form of linkages among concepts (as illustrated in Figure 2). First, the researcher combined terms into concepts and concepts into categories based on whether they mean the same thing. Second, the researcher decided how many categories to keep based on their regression weights with the criterion.

Table 1
Illustration of the List of Concepts Initially Extracted for the Leadership Skill Accomplishment Records

Concept	Number of times	Number of documents
Work	17,037	12,287
Team	23,762	11,932
Leadership	15,317	11,779
Students	21,897	9,540
Group	15,373	8,921
Lead	10,800	8,809
Time	9,992	8,416
Leader	9,280	6,894
Help	8,142	6,660
Project	11,027	6,565
People	8,448	6,385
Members	8,433	5,823
Experience	6,523	5,690
Goals	6,656	5,366
Responsibility	6,095	5,280
Working	5,876	5,229
Organization	7,899	5,222
Tasks	6,353	5,101
Example	5,615	4,926
Position	5,835	4,865
Meeting	6,400	4,864
Efforts	5,541	4,791
Opportunity	5,461	4,790
Program	7,406	4,626

Third, the researcher repeated the first two steps in an iterative fashion to continue to improve the model. In the present study, this process continued until the computer model predicted the raters' scores on each competency as well as a single rater (reflected by a correlation of .60 between the computer and the rater scores), which took more than 200 hours of researcher time. Continued refinements of the computer model beyond the goal of obtaining a .60 correlation with the rater scores may be possible, but an asymptote was observed as the .60 level was approached. Gains in the correlation were increasingly more difficult to make. The initial extraction of the categories by the computer without any training produced correlations in the low .40s. Initial gains to the low .50s were relatively easy to make. However, approximately 80% of the 200 hours of training time was spent on improving the correlations from the low .50s to .60. The number of iterations depended on several factors: the complexity of the models, with

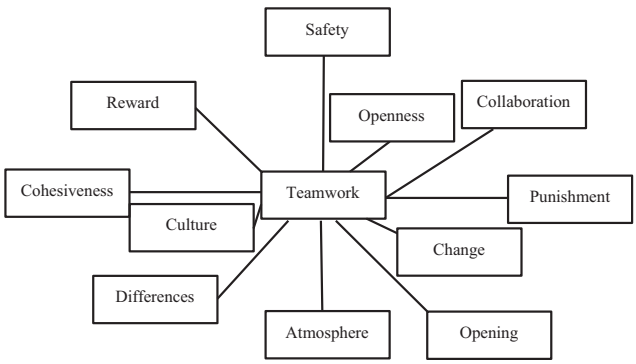

Figure 2. Illustration of a concept map for teamwork.

Table 2
Illustration of Categories Output by the Program for the Leadership Skill Accomplishment Records

Category	β	*t*
Leadership_executives	**.01***	**2.47**
Leadership_executives/class leader	**−.01***	**−2.01**
Leadership_executives/committee leader	−.00	−.35
Leadership_executives/formal leader	−.01	−1.54
Leadership_executives/good leader	−.01	−1.29
Leadership_executives/group leader	−.01	−1.67
Leadership_executives/key leaders	.01	1.95
Leadership_executives/leadership board	.00	.65
Leadership_executives/local leaders	**.01***	**2.51**
Leadership_executives/military leaders	.00	−.16
Leadership_executives/national leaders	.00	.52
Leadership_executives/natural leader	**−.01***	**−2.06**
Leadership_executives/new leaders	.01	1.15
Leadership_executives/official leader	**−.01***	**−2.42**
Leadership_executives/potential leaders	**−.01***	**−2.74**
Leadership_executives/practice leadership	.00	.10
Leadership_executives/project leader	.00	.18

Note. Categories categories significant at * $p < .05$ are bolded.

complexity determined by the number of variables included and the variation in possible responses (e.g., the AR data were much more complex than the other categories of text information mined); whether the researcher worked on one section of the model at a time, like a subset of concepts (e.g., location where the AR occurred), or the entire model; and the skill of the researcher, which improved with practice.

Results

Programming (Text Mining) Results

Table 3 shows the number of categories scored for each text mined field for each competency. The ARs yielded about 700 to 900 categories, job titles about 60 to 90, work duties about 100 to 300, employers about 50 to 130, special skills about 60 to 90, and other experiences about 10 to 20. In total, the text mining extracted about 1,200 to 1,400 categories for each competency.

Table 4 shows the categories scored for an example candidate's leadership AR. The question and candidate's answer are at the top. The categories scored by the computer are in the middle. Note that the computer scores only categories that are relevant to a given answer of the hundreds possible for that narrative, and each category is a 0/1 dummy code for the presence of the category in the answer. The bottom of the table shows the scores assigned by the three raters and the computer. The raters scored the essay a 10, and the computer scored it a 9 in this case.

In addition to the text mined variables, the analyses included all the quantitative variables available to raters as part of the hiring process. The analysis includes all the quantitative variables to predict all six competencies because they may be considered by the raters when making the ratings on all the competencies. Table 3 shows the grand total of all the variables in the final analyses with the quantitative variables included, which totals approximately 1,300 to 1,500 across competencies.

Table 3
Number of Categories Scored for Each Text Mined Field

Variable	Communication skill	Critical thinking	People skill	Leadership skill	Managerial skill	Factual knowledge
Accomplishment record	722	782	794	878	722	698
Job titles	70	87	58	76	78	93
Work duties	315	110	312	307	329	159
Employers	79	100	66	53	56	129
Special skills	60	83	69	84	75	94
Other experiences	19	16	19	9	12	23
Total text mined categories	1265	1178	1318	1407	1272	1196
Quantitative variables	153	153	153	153	153	153
Total of all variables in the final analyses	1418	1331	1471	1560	1425	1349

Research Question 1: Can a Computer Program Be Trained to Generate Scores That Demonstrate a Level of Reliability That Is Comparable With Scores of a Human Rater?

Table 5 shows a comparison of computer and human rater scores for each competency. The means are identical between the raters' scores and the computer scores for five of the competencies, and they differ by .01 for the sixth competency. The difference on the composite of all six competencies combined is .01. The means are also very similar between the training and testing samples. Most are either identical or differ by .02 points.

The main difference is that the computer scores have smaller standard deviations and usually narrower ranges than the raters' scores, which occurs in most statistical prediction situations. This is because prediction is less than perfect (which is always the case), so the predicted variation in scores will be less than the variation in the data used to create the predictions (the observed scores). For example, if the correlation between computer and human rater scores is .60, there will only be about 60% as much variation in the computer scores when compared to the human rater scores. This is approximately what is observed in the standard deviations in Table 5. Note that here we are not referring to the common variance between computer and human rater scores (coefficient of determination). We are referring to the expected variance in the predicted (computer) scores compared with the observed (human rater) scores when a criterion is predicted by one predictor and the correlation (same as the beta if only one predictor) is .60.

Table 6 shows the correlations between the computer and human rater scores for each competency. The correlations are generally in the .60s, which was the target. For comparison, Table 6 also shows the correlations between raters, which are the same as the interrater-reliabilities. The ICC(1) values reflect the reliabilities of single raters, and ICC(3) values reflect the reliabilities of the means of the three raters (based on intraclass correlations, LeBreton & Senter, 2008). The average of the ICC(1)s in Table 6 for individual competencies is .61, which is very similar to the average correlations between the computer and human rater scores of .64, demonstrating that the computer achieved reliability as high as a single rater.

Table 6 also shows the cross-validation by comparing the correlations in the training sample with the correlations in the testing sample. The results show that the correlations between the computer scores and the raters' scores cross-validated well, meaning they are close to the same size. Usually, as is seen here, there is some reduction in the size (i.e., "shrinkage") attributable to the capitalization on chance in the original sample that will not be present in the cross-validation sample. The reduction in size here is very small, averaging about .03 due partly to the stability of statistical estimates from the very large samples used.

To further cross-validate the correlations between the computer and human rater scores, we conducted cross-validations on two subsequent hiring waves. The cross-validation samples were used as a check to see whether the text mining variables and regression weights would predict in other samples, but the variables and the weights from the original sample were used going forward because they were based on the much larger sample and thus would presumably be more stable. Table 7 shows the results for wave one. As with the original sample, the means are very similar (53.12 vs. 56.86) and the standard deviations are smaller (6.23 vs. 9.47) for the computer scores than they were for the rater scores. The correlations cross-validated very well in the new sample. In fact, the sizes of the correlations are noticeably larger than they were in the cross-validation with the original sample (see Table 6). For example, the total score correlation is .75 here compared with .68 in the original sample. Table 7 also shows the results for wave two. The results were extremely similar. These findings support the validity of the computer model in predicting rater scores and suggest that the model will perform equally well in future samples when it is used to help make selection decisions.

In summary, the answer to Research Question 1 is that scores generated by a computer can be as reliable as those assigned by a human rater. The means are almost identical, but the standard deviations are smaller. The computer scores also cross-validated well.

Research Question 2: What Is the Construct Validity of Computer Scores? Specifically, (a) How Will They Correlate With Human Scores, (b) What Are the Subcomponents of Computer Scores, (c) Will Computer Scores Generated by a Computer Exhibit the Same Correlations With the Other Selection Procedures, and (d) What Is the Nature of any Extreme Differences Between Computer and Human Scores?

We evaluated this research question in several ways. First, we examined intercorrelations between computer and human rater scores to examine convergent and discriminant validity (using a

Table 4
Illustration of Scored Categories of an Accomplishment Record

Accomplishment record question: Describe how you have demonstrated leadership.

Candidate answer: During my first year of law school, I noticed that there was a portion of the student population which was unrepresented within the student organizations. Together with three friends, I decided to help form a new student organization that would address the needs and interests of these students. Starting an organization, and then spending two years serving as its vice president, requires a significant amount of leadership. Along with the other founding members of the organization, I first had to organize meetings and events, as well as publicize the organization to other students on campus. I also had to ensure that there were the proper amount of funds to ensure our success. We knew that leadership would be key to the success of any new student organization. First, we had to make ourselves known as an organization. Then we had to be willing to take suggestions and to use our experiences to better the organization. Most importantly, we had to create an organization that students would want to continue after we had graduated. Through our leadership, and in only three years, the organization has grown to be one of the largest on campus. It now boasts a calendar of events which includes well known speakers, fundraisers, and roundtables with professors.

Categories scored by the computer for this answer
campus
establishment
establishment/academy
establishment/academy/students
establishment/academy/students/student organization
finance
finance/funds
meeting
members
members/founding member
occupations
occupations/professor
occupations/vice president
occupations/white collar workers
occupations/white collar workers/executives
work
work/work environment
work/work environment/company events
work/work environment/company events/organizer

Scores from human raters and computer	Score
Leadership – Reviewer 1	3
Leadership – Reviewer 2	3
Leadership – Reviewer 3	4
Leadership – Sum of 3 reviewers	10
Computer score	9

multitrait multimethod approach; Campbell & Fiske, 1959). Second, to better understand the meaning of the total computer score, we separated the text mining from the quantitative variables and examined the correlations between raters' scores and each of these subcomponents. Third, we correlated the computer scores with the other assessment scores to further explore the construct validity. Fourth, we examined the small number of extreme mis-predictions to determine why computer and computer scores differ in these instances.

Intercorrelations. Table 8 shows the intercorrelations among the computer scores and among the human rater scores. In general, the intercorrelations are fairly similar. The average intercorrelation among the human rater scores is .57, while the average among the computer scores is slightly higher at .64. Table 8 also shows the intercorrelations among the human and computer scores. For every competency, the correlation between the human and computer score for that competency is larger than the correlations with other competencies, thus providing evidence of convergent and discriminant validity.

Examining subcomponents of computer scores. Table 9 shows the correlations between the human rater scores and the text mining and quantitative variables. The previous analyses combined the variables to predict raters' scores in order to model the actual decision-making process for the overall selection procedure. However, it is instructive to consider the contribution of each separately. Table 9 shows that all the variables accounted for a portion of the raters' scores. The AR showed the highest correlations, the quantitative variables showed similar or lower correlations depending on the competency, and the other text mined fields showed the lowest correlations. For example, the text mining of the AR for leadership correlates .54, the text mining of the other narrative answers correlate from .10 to .33, and the quantitative variables as a set correlate .37, whereas all of them combined correlate .65 (as shown in Table 6 above). The results are similar for the other ARs. This suggests that each set of variables accounts for a portion of the raters' scores, which is logical given that all these variables are supposed to be considered by the raters when scoring. It is also noteworthy that the AR shows the highest correlation with the raters' scores, suggesting it has the greatest influence.

Table 9 also shows the regression weights to index the unique contribution of each set of variables to the prediction of the rater scores. The AR accounts for the greatest portion of the prediction. The other variables account for a small portion when the AR is in the equation. This is because the candidates likely discuss the other variables in their ARs. The only exception is the quantitative variables, which are probably not discussed by the candidate but may be considered by the rater. They consequently show somewhat larger regression weights.

Correlations with other selection procedures. Table 10 shows correlations among the text mining variables and the other selection procedures (the employment tests and the onsite selection procedures). These are correlations of the text mining variables separately and without the quantitative variables. As shown, the correlations between each of the other selection procedures and the text mining of the AR were stronger than the correlations between each of these other selection procedures and the other text mining variables. In addition, the quantitative variables showed larger correlations with the employment tests than the text mining variables, but this is because the employment tests include the quantitative variables. The somewhat larger correlations with the professional knowledge test, biodata, and interview suggest the AR and other text mining variables measure content (knowledge and past experience), and the somewhat lower correlations with the English test and the essay suggest the text mining variables less reflect writing skill. Finally, the correlations with the onsite procedures were higher for the computer scores than the rater scores.

As a more direct comparison, we correlated the total computer scores, including all the competencies, with a composite of the onsite scores. This correlation was .25, and the rater

Table 5
Comparison of Human Rater and Computer Scores for Each Competency

	Training		Testing		All	
Competency	Rater scores	Computer scores	Rater scores	Computer scores	Rater scores	Computer scores
Communication skill						
N	35,170	35,170	6,259	6,259	41,429	41,429
Mean	9.52	9.52	9.48	9.52	9.51	9.52
SD	2.16	1.19	2.16	1.19	2.16	1.19
Min	3.00	4.55	3.00	5.88	3.00	4.55
Max	15.00	14.91	15.00	16.12	15.00	16.12
Critical thinking						
N	35,170	35,170	6,259	6,259	41,429	4,429
Mean	9.49	9.49	9.47	9.49	9.49	9.49
SD	2.19	1.24	2.19	1.23	2.19	1.24
Min	3.00	4.90	3.00	5.53	3.00	4.90
Max	15.00	14.63	15.00	14.73	15.00	14.73
People skill						
N	35,170	35,170	6,259	6,259	41,429	41,429
Mean	9.19	9.19	9.17	9.19	9.19	9.19
SD	2.11	1.02	2.12	1.01	2.11	1.01
Min	3.00	5.73	3.00	5.38	3.00	5.38
Max	15.00	15.27	15.00	14.13	15.00	15.27
Leadership skill						
N	35,170	35,170	6,259	6,259	41,429	41,429
Mean	9.17	9.17	9.16	9.19	9.17	9.17
SD	2.14	1.05	1.01	1.01	2.14	1.05
Min	3.00	5.36	5.38	5.38	3.00	5.36
Max	15.00	14.46	14.13	14.13	15.00	14.46
Managerial skill						
N	35,170	35,170	6,259	6,259	41,429	41,429
Mean	9.44	9.44	9.44	9.45	9.44	9.44
SD	2.09	1.03	2.07	1.02	2.09	1.03
Min	3.00	5.03	3.00	5.80	3.00	5.03
Max	15.00	14.65	15.00	14.62	15.00	14.65
Factual knowledge						
N	35,170	35,170	6,259	6,259	41,429	41,429
Mean	9.47	9.47	9.48	9.47	9.47	9.47
SD	2.51	1.51	2.53	1.51	2.52	1.51
Min	3.00	3.80	3.00	4.83	3.00	3.80
Max	15.00	16.35	15.00	16.60	15.00	16.60
Total						
N	35,170	35,170	6,259	6,259	41,429	41,429
Mean	56.28	56.28	56.20	56.29	56.27	56.28
SD	10.59	5.92	10.54	5.88	10.58	5.92
Min	18.00	31.89	19.00	35.42	18.00	31.89
Max	90.00	83.81	86.00	86.01	90.00	90.00

scores including all competencies correlated .13. Correcting the rater scores for direct range restriction from selection on those scores increases the correlation to .17, and correcting the computer scores for indirect range restriction due to their correlation with the rater scores increases the correlation to .30. Although primarily providing evidence of discriminant validity, this indicates that the computer scores may identify potential passers of the onsite selection procedures slightly better than the raters, which could be viewed as another benefit of the computer scoring. It will reduce the cost of the onsite selection procedures because fewer would need to be brought onsite in order to get sufficient passers to fulfill hiring needs.

Correlations with reading level were examined. The purpose was to determine whether the reading level of the AR correlates with the scores assigned. In other words, do the raters and the computer give higher scores to ARs written at a higher reading level? To analyze this, we determined the reading level of a random sample of 100 leadership AR responses using the Flesch Reading Ease score. This index is based on the average sentence length and average number of syllables per word. The higher the rating (from 0 to 100), the easier the text is to understand. The average of the 100 narratives is 43.5 ($SD = 13.7$), indicating moderate reading difficulty and a fair amount of variation among candidates. The relatively small correlations between the computer scores and reading level ($-.25$ and $-.30$) provide evidence of discriminant validity. However, the slight negative correlations suggest that higher quality answers (i.e., higher levels of leadership) may require more complex writing to explain, or more complex writing may be interpreted by raters and the computer as higher quality. Also, small correlations may be attributable in part to the fact that the ARs are only one factor considered in the total score.

Table 6
Correlations Between the Human Rater and Computer Scores, and Interrater Reliabilities Between Human Rater Scores for Each Competency

Essay	Correlations between the human rater and computer scores		Interrater reliabilities between human rater scores		
	Training	Testing	*F*	ICC(1)	ICC(3)
Communication skill	.65	.63	5.54	.60	.82
Critical thinking	.63	.61	5.48	.60	.82
People skill	.67	.63	6.23	.64	.84
Leadership skill	.68	.65	5.43	.60	.82
Managerial skill	.62	.59	5.07	.58	.80
Factual knowledge	.66	.64	6.03	.64	.84
Total	.70	.68	10.54	.76	.91

Note. For training sample, $N = 35{,}170$. For testing sample, $N = 6{,}259$. All correlations are significant ($p < .05$). For interrater reliabilities, $N = 41{,}429$. All F values are significant ($p < .05$).

Analysis of extreme mis-predictions. Finally, we analyzed extreme mis-predictions, defined as the very small number of cases when the computer produced a very high score and the raters assigned a very low score or vice versa. We identified the most extreme mis-predictions by calculating the difference between rater scores and computer scores. We selected the leadership competency for this analysis because characteristics of leadership are fairly specific and understandable, and the question asked of candidates is straightforward ("Describe how you have demonstrated leadership."). We then identified the 100 most extreme overpredictions and 100 most extreme underpredictions to gain insight into why the mis-predictions may have occurred.

Note that reading the ARs to understand the mis-predictions does not consider the impact of the quantitative variables on the rater scores (e.g., test battery scores, level of education, major, number of jobs, years of experience, etc.). Therefore, this analysis will only shed partial light on whether there might be characteristics of the ARs that cause the mis-predictions.

We reviewed the mis-predictions by reading the text of the ARs and observing the number of text mining categories scored. Characteristics of extreme overpredictions (where the computer score was much higher than the rater score) were descriptions of experiences or listings of past jobs as opposed to descriptions of demonstrating leadership, poor examples of leadership even though the right leadership words were used (as reflect in the text mining categories), examples where the candidate was critical or harsh with team members, used authority or force, or showed a bad attitude toward them, bragging and apparent exaggeration, or rambling answers that did not answer the question (e.g., stating an opinion on leadership).

Conversely, some of the characteristics of the ARs of the sample of extreme underpredictions (where the computer score was much lower than the rater score) were answers that hit fewer of the leadership categories scored by the text mining, which might be attributable to the unusual or unique nature of the answers or to the word choices of the candidates, answers that have the opposite characteristics to the overpredictions above, and ARs were the scores were probably enhanced by

Table 7
Cross-Validation on Human Rater Scores for the First and Second Hiring Wave

Competency	Communication skill		Critical thinking		People skill		Leadership skill		Managerial skill		Factual knowledge		Total	
	Rater scores	Computer scores	Rater scores	Computer scores	Rater scores	Computer scores	Rater scores	Computer scores	Rater scores	Computer scores	Rater scores	Computer scores	Rater scores	Computer scores
N	2,300/2,198	2,300/2,198	2,300/2,198	2,300/2,198	2,300/2,198	2,300/2,198	2,300/2,198	2,300/2,198	2,300/2,198	2,300/2,198	2,300/2,198	2,300/2,198	2,300/2,198	2,300/2,198
Mean	9.52/9.14	8.14/9.38	9.56/9.33	8.62/9.58	9.18/9.05	9.72/9.06	9.46/9.22	9.46/9.33	9.66/9.48	10.48/9.56	9.66/9.30	10.48/9.53	56.86/55.52	53.12/56.44
SD	1.98/1.99	1.20/1.23	2.04/2.05	1.32/1.29	1.80/1.93	.96/.97	1.99/2.06	1.12/1.12	1.87/1.98	1.08/1.08	1.87/2.34	1.08/1.66	9.47/9.77	6.23/6.22
Min	3.00/3.00	4.18/5.53	3.00/3.00	4.55/6.16	3.00/3.00	6.03/6.03	3.00/3.00	5.90/6.41	3.00/3.00	6.98/5.84	3.00/3.00	6.98/4.87	23.00/19.00	34.28/40.57
Max	15.00/15.00	15.70/15.07	15.00/15.00	13.90/14.67	15.00/15.00	14.31/12.63	15.00/15.00	13.93	15.00/15.00	15.20/14.10	15.00/15.00	15.20/17.16	90.00/86.00	85.92/81.78
r	.64/.62		.67/.64		.65/.62		.66/.64		.63/.59		.63/.72		.75/.71	

Note. Values to the left of the slash are for the first wave, and values to the right are for the second wave. All correlations are significant ($p < .05$).

Table 8
Intercorrelations Among Human Rater Scores and Among Computer Scores

Variable	Communication skill score	Critical thinking score	People skill score	Leadership skill score	Managerial skill score	Factual knowledge score	Total score
Communication skill score	1	.69	.67	.54	.54	.72	.84
Critical thinking score	.59	1	.64	.62	.65	.78	.88
People skill score	.59	.57	1	.62	.59	.67	.82
Leadership skill score	.53	.55	.56	1	.66	.60	.79
Managerial skill score	.52	.56	.54	.63	1	.61	.80
Factual knowledge score	.61	.66	.57	.53	.53	1	.89
Total score	.80	.82	.79	.79	.78	.82	1

	Computer communication skill score	Computer critical thinking score	Computer people skill score	Computer leadership skill score	Computer managerial skill score	Computer factual knowledge score	Computer total score
Rater communication skill score	.65	.44	.50	.42	.39	.45	.56
Rater critical thinking score	.47	.63	.49	.48	.46	.50	.60
Rater people skill score	.40	.37	.66	.43	.38	.38	.51
Rater leadership skill score	.35	.36	.43	.67	.44	.35	.50
Rater managerial skill score	.35	.38	.42	.49	.61	.36	.51
Rater factual knowledge score	.51	.53	.53	.48	.46	.65	.64
Rater total score	.57	.57	.63	.62	.57	.57	.70

Note. The intercorrelations among human rater scores are in the lower diagonal and among the computer scores are in the upper diagonal. $N = 41{,}429$. All correlations are significant ($p < .05$).

high levels on the quantitative information on test scores and education.

In summary, the answers to Research Question 2 are as follows: the correlations between human and computer scores for the same competency are larger than with other competencies, thus providing evidence of convergent and discriminant validity; separating the text mining variables from the quantitative variables showed that all the variables accounted for a portion of the raters' scores but the ARs and quantitative variables showed the highest correlations, and the other text mined fields showed the lowest correlations; the pattern of correlations with the other selection procedures (both the test battery prescreen and the onsite assessment exercises) is similar for both the computer-generated and raters' scores, suggesting similar construct validity; relatively small correlations with reading level provide evidence of discriminant validity, but the slight negative correlations suggest that higher quality answers may require more complex writing to explain or may be interpreted that way; and some characteristics of the ARs of extreme overpredictions for leadership were descriptions of experiences or listings of past jobs, as opposed to descriptions of demonstrating leadership behavior, and poor examples of leadership.

Research Question 3: Will Scores Generated by the Computer Be Associated With the Same Lack of Subgroup Differences as Those Observed With Human Raters?

The *d* scores by race and gender were essentially zero in all cases. This is the case for the text mining variables, the quantitative variables, and the raters' scores. For example, the *d* scores for each racial group (Asian, Black, and Hispanic, dichotomously coded as 1 compared with White coded as 0) and the total computer score for leadership (with a mean of approximately 56, Table 5) ranged from −.03 to .06. The *d* score for gender (women coded 1) was .08. Because of confidentiality agreements with the

Table 9
Correlations and Regression Weights (in Parentheses) of Text Mining Variables and Quantitative Variables With the Human Rater Scores

Variable	Communication skill	Critical thinking	People skill	Leadership skill	Managerial skill	Factual knowledge
Accomplishment record	.50 (.23)	.51 (.23)	.49 (.25)	.54 (.57)	.51 (.32)	.49 (.21)
Job titles	.24 (−.02)	.29 (.02)	.24 (.02)	.26 (.04)	.29 (.08)	.32 (.01)
Special skills	.21 (.00)	.23 (.01)	.21 (.02)	.24 (.06)	.21 (.05)	.27 (.00)
Work duties	.34 (.10)	.34 (.11)	.32 (.10)	.33 (.10)	.35 (.12)	.42 (.15)
Employers	.25 (−.05)	.27 (−.06)	.27 (.01)	.25 (−.01)	.22 (−.05)	.36 (−.01)
Other experiences	.12 (.03)	.12 (.03)	.13 (.03)	.10 (.01)	.10 (.02)	.15 (.04)
Quantitative variables	.48 (.15)	.50 (.21)	.38 (.07)	.37 (−.16)	.39 (.04)	.45 (.22)

Note. $N = 41{,}432$. Regression weights are in parentheses. All correlations and all regression weights with betas whose absolute value is greater than or equal to .02 are significant ($p < .05$).

Table 10
Range of Correlations of Text Mining Variables With Scores on Other Selection Procedures Across Competencies

Variable	Total employment test battery	Professional knowledge	Biodata	English test	Essay	Total onsite	Leaderless group discussion	Management case	Interview
Accomplishment record	.36–.44	.20–.29	.12–.18	.13–.18	.01–.09	.15–.19	.10–.15	.03–.11	.18–.24
Job titles	.07–.15	.02–.10	.06–.12	−.04–.02	−.08–.00	.07–.12	.00–.10	−.05–.05	.14–.19
Work duties	.12–.20	.05–.13	.07–.13	−.01–.07	−.02–.05	.08–.14	.04–.13	.01–.09	.15–.21
Employers	.05–.13	.00–.09	.03–.10	−.03–.03	−.05–.02	.04–.12	−.01–.08	−.01–05	.07–.17
Special skills	.04–.10	−.02–.05	.09–.14	−.04–.02	−.06–.00	.02–.09	−.01–.03	−.04–.06	.06–.12
Other experiences	.01–.06	.00–.04	.00–.05	.01–.04	−.02–.04	−.04–.06	−.01–.04	−.05–.03	−.02–.06
Quantitative variables	.51.–.59	.28–.38	.18–.25	.18–.27	.04–.13	.13–.22	.07–.17	.06–.12	.17–.23
Computer scores	.50–.59	.39–.56	.14–.20	.38–.41	.09–.19	.24–.29	.20–.25	.11–.18	.23–.27
Rater score	.14–.21	.07–.17	.06–.13	.08–.14	.03–.08	.09–.16	.06–.13	.03–.06	.13–.19

Note. $N = 41{,}429$ for employment test battery ($rs > .01$, $p < .05$), and 11,892 for onsite selection procedures ($rs > .02$, $p < .05$).

organization and to conserve space, a table displaying these results is not included.

Note that text mining scores from the quantitative variables did not correlate with race, despite the fact that the quantitative variables included the test scores. This is because the quantitative variables were mined to predict the rater scores and the rater scores did not correlate with race. This is probably because the raters did not factor employment tests strongly into their ratings, which makes sense because they know that the employment tests already influenced the selection decisions at the first stage of the hiring process.

Therefore, computer scoring should not create any adverse impact beyond what is already present in the selection procedure. In this case, the ARs did not have adverse impact, so computer scoring did not create any adverse impact. However, if the ARs did have adverse impact (depending on what they measure), then the computer scores would as well, but would not add to the adverse impact.

Research Question 4: What Are the Potential Cost Savings Associated With Using Text Mining and Predictive Modeling as a Surrogate for Human Raters?

Scoring the ARs is very expensive in this organization. It takes approximately 20 minutes per candidate for each of the three raters (1 hour total). At an average cost of $70/hour for rater time, the cost of scoring the 7,000 to 9,000 candidates who move through this hurdle each year is approximately $490,000 to $630,000 annually. The computer scoring is used to replace about one third of those rater hours, so the cost savings across years range from $163,000 to $210,000 annually. Moreover, the organization saves an additional $20,000 on rater training hours because fewer raters are needed. The investment was $19,000 for the one-time purchase of the software, about $40,000 for initial programming costs, and a couple thousand dollars in programmer time to process the data each year. Thus, computer scoring would be very cost-effective for the organization.

However, note that there were sunk costs that the organization had already incurred that may overestimate the financial gain to be had through adopting this scoring procedure in a new organization. That is, a new organization may have to develop the selection procedure, run it for a while to gather the data using human raters, and then develop the computer scoring model. It will only be at that point before partial costs of ongoing administration can be saved.

Nevertheless, there may be other ways to use computer scoring to save costs. First, computer scoring could be used to replace more than one rater if the reliability is adequate. Second, organizations often have archival data on preexisting selection systems, which was the case in this context. Third, consulting firms could develop products around this scoring procedure, thus amortizing costs across many clients. Finally, as noted below, a computer model may not require samples that are as large as those in this organization.

In summary, the answer to Research Question 4 is that substantial cost savings can be realized through the utilization of computer scoring, depending on how the scores are used.

Research Question 5: What Are the Potential Implementation Issues Associated With Using Computer Scoring as a Surrogate for Human Raters?

This question is addressed with three analyses. First, we examine two potential ways to implement the computer scores in this setting. Second, we examine how to respond to changes in the AR questions should they be made in the future. Third, we examined the minimum sample sizes required to use computer scoring.

How should computer scoring be implemented? We explored two practical issues. First, it is important to consider how computer scores should be used. One possible way is to replace one of the three raters with the computer scores. We conducted two analyses of this alternative. First, we correlated the total scores based on the composite of the three raters with the total scores based on using the computer scores as the third rater. The correlations for each competency and the composite are extremely high (.94–.97), suggesting that the computer scores can be substituted for the third rater. This is expected because the computer scores are as reliable as a single rater.

Second, we calculated the number of passers who would have failed had we used the computer score as the third rater. We used the top third as the passing score. A total of 1,097 of the passers would have failed. This is 7.9% of the passers (1,097 divided by 13,810), 4.0% of the failers (1,097 divided by 27,619), and 2.6%

of the total sample of candidates (1,097 divided by 41,429). Therefore, using the computer scores to replace the third rater would result in a fairly small percentage of different decisions. Note that these analyses assume the third rater is correct and the computer is wrong. In fact, it may be the reverse, and the same would occur if one rater were replaced by another.

Still, another possibility would be to use the computer score to eliminate the bottom third of candidates. We approached this analysis in several ways. First, we identified the candidates classified in the bottom third using the computer scores who would not be in the bottom third based on the rater scores. Of the 13,809 candidates classified in the bottom third by the computer, 3,473 were classified in the middle third by the raters and 879 were classified in the top third by raters. Thus, 31.4% (3,473 + 879 divided by 13,809) of the candidates classified in the bottom third by the computer were *not* classified in the bottom third by raters. In terms of the total sample of candidates, 10.5% would be different decisions in the bottom third (3,473 + 879 divided by 41,429).

Second, we examined the impact of eliminating the bottom third on the candidates who passed based on the rater scores. This analysis is relevant because the passing scores on this selection procedure are usually very high, so eliminating the candidates in the bottom third may have little impact on those actually passing. One approach to this analysis is to assume a third of people passed the selection procedure based on historical data. A total of 879 of the passers based on the rater scores would not have been passers because they would have been eliminated. This is 6.4% of the passers (879 divided by 13,810), 3.2% of the failers (879 divided by 13,809 + 13,810), and 2.1% of the total sample of candidates (879 divided by 41,429).

Another approach to this analysis is to examine the actual people who passed the selection procedure. A total of 1,347 candidates in the bottom third based on the computer scores would have scored high enough based on the rater scores to pass the selection procedure. This is 9.8% of the passers (1,347 divided by 13,687), 4.9% of the failers (1,347 divided by 27,742), or 3.3% of the total sample of candidates (1,347 divided by 41,429).

In conclusion, using the computer scores to eliminate the bottom third of candidates results in a fairly small percentage of different decisions. Depending on the analysis, it would be 2.1% or 3.3% different decisions in terms of the total number of selection decisions made, and the same result would occur if a single rater were used for this purpose.

What if accomplishment questions change? Because the computer scoring of the selection procedure in this study is based largely on historical answers to these specific AR questions and the scores raters assigned to those answers, changes to the questions will require collecting and text mining new answers. Thus, when a question needs to be revised, answers to that question cannot be scored by a computer the first time it is used. Instead, answers would be collected and scored by raters, and then text mined so that the question can be computer scored in the future. The total score will be based on the other five competencies for that administration.

To estimate the influence of omitting one of the six competencies, we calculated the reliability of the total score omitting each of the six competencies one at a time. The impact on the internal consistency reliabilities is negligible. The reliability of the composite of all six competencies is .91, and eliminating one competency reduces the reliability to .89 for every competency. Thus, using only five competencies for one administration would not be a problem.

What is the minimum sample size required to use computer scoring? This is an important statistical and practical question. It depends entirely on the range of possible answers in the ARs. If the range is wide, a larger sample will be needed. It also depends on the skill and time of the researcher in training the computer. In the context of scoring student essays to assess writing skills, Ramineni and Williamson (2013) suggest a minimum sample of 500 essays and they usually use 500 to 2000. Shermis et al. (2010) also give an illustration using 500, with 300 in the training sample and 200 in the validation sample.

To address this question empirically in the current dataset, concepts and categories were extracted using the default settings (untrained) and correlated with the human scores for samples of various sizes covering the range of the expected minimum (50, 100, 250, 500, and 1000). Then the concepts and categories were applied to 10 randomly drawn samples of the same size to estimate the stability of the estimates. Again, the leadership competency was used for illustration. The results in Table 11 show that the initial correlations drop off gradually with smaller sample sizes, but the mean cross-validated correlations drop off radically and the standard deviation of the cross-validated correlations increases rapidly with samples below 500. Based on these analyses, it appears that a minimum sample of 500 essays would be needed to obtain meaningful and reasonably stable cross-validated correlations.

Assuming that modifying the categories (training by the researcher) would have the same effect in improving the correlation as it did using the full sample of about 100% in variance explained, the .36 initial correlation in the sample of 500 would yield a correlation of about .51 with the human scores. This is not as high as the value obtained here with the full sample of .60 or higher, but it might be satisfactory if multiple competencies were scored. Therefore, a minimum sample of 500 seems reasonable from this analysis as well.

In summary, the answers to Research Question 5 are as follows: there are several possible ways to use the computer scores such as replacing one rater or eliminating the lowest scoring candidates; if an AR question is revised, the total score based on the other ARs is reliable enough to not include that AR while data on the new AR

Table 11
Influence of Sample Size on Cross-Validated Correlations

Correlation	Sample size				
	50	100	250	500	1000
Initial correlation	.24*	.29*	.32*	.36*	.35*
Mean cross-validated correlation	.03	.06	.04	.18*	.17*
Standard deviation of cross-validated correlation	.13	.09	.07	.03	.02

Note. Initial correlations are between computer scores and human rater scores in the sample used to develop computer scoring model. Cross-validated correlations are based on 10 random samples of the size indicated.
* $p < .05$, one-tailed.

is collected; and the minimum sample size to use computer scoring is not excessively high, such as about 500 in this setting.

Discussion

The purpose of this study was to alert scholars and practitioners to the possibility of, and potential advantages and disadvantages associated with, using text mining and predictive computer modeling as an alternative to human raters in a selection context. Our study yielded five key findings. First, it appears possible to program or train a computer to emulate a human rater when scoring ARs and other narrative data. Second, we found that the computer program was capable of producing scores that were as reliable as those of a human rater. Third, scores produced by the computer appeared to demonstrate construct validity. Further, the computer scores showed no gender or race differences, similar to the human rater scores, suggesting they will create no adverse impact. In addition, close examination of mis-predictions also indicated logical reasons for these cases. Fourth, it appears possible that computer scoring can result in substantial cost savings when used to score ARs. Finally, analyses of various operational issues suggested the number of different hiring decisions that would occur using computer scoring would be very small, which was no different than what would occur by replacing one of the three raters on the panel. Revisions to AR questions would likely not impact operational effectiveness.

Implications for Scholarship and Practice and Directions for Future Research

Our study advances selection scholarship and practice in at least four important ways. First, we provide evidence suggesting that advances in text mining capabilities and the development of predictive modeling software programs have the potential to usher in a new era of selection scholarship and practice. Most directly, these techniques could enable the use of selection procedures for large-scale application that were previously too expensive. For example, these techniques could be used to inexpensively process and score the often large number of applications organizations receive. Presently, applications are generally scored using rudimentary information retrieval programs such as keyword searches. Computer scoring may not only provide a more cost-effective alternative to separating those who are qualified from those you are not, but may also identify applicants that would otherwise not have been recognized as qualified due to their use of terms other than the keywords to describe their skills. Likewise, computer scoring may also provide a superior way to score other selection procedures using constructed responses such as biodata. In addition, these techniques may allow selection researchers to score responses that combine text and numbers, examples of which include tests that score how problems are worked out as well as whether the correct answer was obtained (e.g., CPA Licensing Exam; American Institute of Certified Public Accountants, 2011). Finally, text mining could be used alone (without rater scores and a predictive model) to simply summarize applications and identify characteristics of applicants received.

Second, the present study suggests these advances may allow scholars to begin to broaden the scope of approaches used to mitigate the adverse impact-validity dilemma (Ployhart & Holtz, 2008; Pyburn et al., 2008) to include ones that focus on reducing the cost of scoring. This would render low-impact procedures (e.g., those using constructed response formats; Arthur, Edwards, & Barrett, 2002; Edwards & Arthur, 2007) more financially feasible for large-scale use. This is in contrast to the present focus of scholarship and practice on adopting alternative procedures or uses of scores.

Also, with other advances in computer technology, such as voice recognition, computers may be capable of scoring structured interviews in the future, thus making this highly valid selection procedure (e.g., Huffcutt & Arthur, 1994; McDaniel, Whetzel, Schmidt, & Maurer, 1994), that also exhibits low levels of adverse impact (e.g., McCarthy, Van Iddekinge, & Campion, 2010), cost-effective for large-scale screening. This would move well beyond current efforts at computer scoring interviews (e.g., interactive voice-response software; Bauer, Truxillo, Paronto, Weekley, & Campion, 2004). Thus, a potentially critical area for future research and practice is the use of computer scoring to reduce the cost of low adverse impact procedures.

Third, we introduce NLP as core to understanding how text mining and predictive modeling software operate as well as how they can be used effectively by organizations. This is particularly important as little prior research on the use of text-based data exists in I/O psychology and related fields. Thus, scholars and practitioners are left with few tools to pursue research requiring that information be extracted from large amounts of unstructured, text-based data. For example, research on attitudes, beliefs, values, and employee engagement often entails collection of large volumes of write-in comments from respondents. Usually, these responses are simply read for insight or content-analyzed in a labor-intensive way (Macey, Schneider, Barbera, & Young, 2009). Computer scoring may permit such data to be scored thus making it more useful for analysis, prediction, tracking, and other research uses. This is a potentially productive area for future research and practical applications. Similarly, inductive research generally also involves labor-intensive analysis of text-based data to discover underlying themes. Before developing the predictive model, our approach entailed condensing a vast quantity of text-based information into its such dimensions and themes (with the aid of the researcher). This text mining component may thus prove valuable toward reducing time spent on analyses.

Finally, our study advances the potential of text mining. Presently, applications of text mining have been limited to scoring primarily *writing skill* in educational contexts (e.g., Rudner et al., 2005). Our study expands the scope to scoring the *content* of essays in the employment context. This opens up a wide range of other potential applications where interest is in what is said, not how it is said. The refinement of computer models for scoring the content of candidate narrative answers and the application of those models to new selection contexts and techniques are prime areas for future research.

Practitioners interested in trying out this approach might start in two ways. First, use existing datasets within the organization that contain constructed response data (e.g., applications, personal statements, etc.). Here, only criterion data would be necessary to allow the creation of a computer scoring model, such as job performance, turnover, or ratings by subject matter experts. Sec-

ond, start small. This study suggests that a sample of 500 might be necessary to use computer scoring.

Limitations

The range of uses of text mining and predictive modeling programs seems almost limitless with several caveats. First, a sample of at least 500 might be burdensome to collect if the data do not currently exist. Second, there must be a criterion score against which to select the thousands of concepts and categories produced by the software, at least when this practice is being used for predictive purposes (e.g., to replace an existing scoring system). Note that a criterion may not be necessary when text mining is used alone. An example of which might include inductive research where the purpose is to condense information and identify themes. The organizational context in which this study was performed had a large corpus of previously rated ARs and text-based data, which offered an ideal context. Many organizations will not have a corpus of this magnitude available. As such, this practice should be replicated in other contexts to identify the boundaries to its feasibility as well as to ensure the generalizability of our findings. Third, a considerable amount of research time is required to train the software, particularly when the researcher is learning how to operate the program. However, this is no different from many other sophisticated analytical techniques. Fourth, in this study, a single researcher was used to train the software program. Although identifying synonyms and linking similar categories is seemingly straightforward, researchers may differ in skill and diligence, which could affect the quality of the text mining. Finally, computer scoring of essays in college entrance exams has been criticized because of its potential for gaming. Candidates may be able to write bogus essays that trick computer software programs into providing them with higher scores than they deserve. Research on the success of such strategies has been mixed (e.g., Powers, Burstein, Chodorow, Fowles, & Kukich, 2002). The scoring model in the present study is too complex to be gamed in a simplistic way. It is based on 700 to 1000 categories for each AR. Additionally, the scoring includes 154 quantitative variables that are objective and cannot be easily faked (e.g., college major, test scores). Nevertheless, this is an important concern that researchers need to consider.

Conclusion

The increasing availability of large amounts of candidate data along with the increasing sophistication of computer software may allow personnel selection techniques involving constructed responses to be implemented more broadly and cost effectively, which could bring a new era in personnel selection research.

References

American Institute of Certified Public Accountants. (2011). How is the CPA exam scored? Retrieved January 10, 2014, from http://www.aicpa.org/BecomeACPA/CPAExam/PsychometricsandScoring/ScoringInformation/DownloadableDocuments/How_the_CPA_Exam_is_Scored.pdf

Arthur, W., Jr., Edwards, B. D., & Barrett, G. V. (2002). Multiple-choice and constructed response tests of ability: Race-based subgroup performance differences on alternative paper-and-pencil test formats. *Personnel Psychology, 55,* 985–1008. http://dx.doi.org/10.1111/j.1744-6570.2002.tb00138.x

Attali, Y., Bridgeman, B., & Trapani, C. S. (2010). Performance of a generic approach in automated essay scoring. *The Journal of Technology, Learning, and Assessment, 10*.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with E-Rater V. 2. *The Journal of Technology, Learning, and Assessment, 4,* 3–30.

Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing, 30,* 125–141. http://dx.doi.org/10.1177/0265532212452396

Bauer, T. N., Truxillo, D. M., Paronto, M. E., Weekley, J. A., & Campion, M. A. (2004). Applicant reactions to different selection technology: Face-to-face interactive voice response and computer-assisted telephone screening interviews. *International Journal of Selection and Assessment, 12,* 135–148. http://dx.doi.org/10.1111/j.0965-075X.2004.00269.x

Ben-Simon, A., & Bennett, R. E. (2007). Toward more substantively meaningful automated essay scoring. *The Journal of Technology, Learning, and Assessment, 6*.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105. http://dx.doi.org/10.1037/h0046016

Chowdhury, G. (2003). Natural language processing. *Annual Review of Information Science & Technology, 37,* 51–89. http://dx.doi.org/10.1002/aris.1440370103

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18,* 7–24. http://dx.doi.org/10.1016/j.asw.2012.10.002

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment, 5,* 3–35.

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science & Technology, 38,* 189–230.

Edwards, B. D., & Arthur, W., Jr. (2007). An examination of factors contributing to a reduction in subgroup differences on a constructed-response paper-and-pencil test of scholastic achievement. *Journal of Applied Psychology, 92,* 794–801. http://dx.doi.org/10.1037/0021-9010.92.3.794

Foltz, P. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, & Computers, 28,* 197–202. http://dx.doi.org/10.3758/BF03204765

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. *Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications*.

Frick, T. W. (1992). Computerized adaptive mastery tests as expert systems. *Journal of Educational Computing Research, 8,* 187–213. http://dx.doi.org/10.2190/J87V-6VWP-52G7-L4XX

Gaizauskas, R., & Wilks, Y. (1998). Information extraction: Beyond document retrieval. *The Journal of Documentation, 54,* 70–105. http://dx.doi.org/10.1108/EUM0000000007162

Guzzo, R., & Carlisle, T. (2014). *Big data: Catch the wave*. Workshop presented at annual conference of the Society for Industrial and Organizational Psychology, Honolulu, HI.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning, 42,* 177–196. http://dx.doi.org/10.1023/A:1007617005950

Hough, L. M. (1984). Development and evaluation of the "accomplishment record" method of selecting and promoting professionals. *Journal of Applied Psychology, 69,* 135–146. http://dx.doi.org/10.1037/0021-9010.69.1.135

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection, and amelioration of adverse impact in personnel selection procedures: Issues, evidence, and lessons learned. *International Journal*

of Selection and Assessment, 9, 152–194. http://dx.doi.org/10.1111/1468-2389.00171

Huffcutt, A. I., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79,* 184–190. http://dx.doi.org/10.1037/0021-9010.79.2.184

IBM. (2012). *IBM SPSS Modeler Text Analytics 15 user's guide*. Armonk, NY: Author.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

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,* 211–240. http://dx.doi.org/10.1037/0033-295X.104.2.211

Landauer, T. K., Laham, D., & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice, 10,* 295–308. http://dx.doi.org/10.1080/0969594032000148154

Leacock, C., & Chodorow, M. (2003). C-Rater: Automated scoring of short-answer questions. *Computers and the Humanities, 37,* 389–405. http://dx.doi.org/10.1023/A:1025779619903

LeBreton, J. L., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11,* 815–852. http://dx.doi.org/10.1177/1094428106296642

Liu, H., Lieberman, H., & Selker, T. (2003). A model of textual affect sensing using real-world knowledge. *Proceedings of the International Conference on Intelligent User Interfaces*.

Liu, H., & Maes, P. (2004). What would they think? A computational model of attitudes. *Proceedings of the ACM International Conference on Intelligent User Interfaces*.

Liu, H., & Singh, P. (2004). ConceptNet—A practical commonsense reasoning tool-kit. *BT Technology Journal, 22,* 211–226. http://dx.doi.org/10.1023/B:BTTJ.0000047600.45421.6d

Macey, W. H., Schneider, B., Barbera, K. M., & Young, S. A. (2009). *Employee engagement: Tools for analysis, practice, and competitive advantage*. West Sussex, UK: Wiley-Blackwell. http://dx.doi.org/10.1002/9781444306538

McCarthy, J. M., Van Iddekinge, C. H., & Campion, M. A. (2010). Are highly structured job interviews resistant to demographic similarity effects? *Personnel Psychology, 63,* 325–359. http://dx.doi.org/10.1111/j.1744-6570.2010.01172.x

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). A meta-analysis of the validity of methods for rating training and experience in personnel selection. *Personnel Psychology, 41,* 283–309. http://dx.doi.org/10.1111/j.1744-6570.1988.tb02386.x

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79,* 599–616. http://dx.doi.org/10.1037/0021-9010.79.4.599

Padmaja, S., & Fatima, S. S. (2013). Opinion mining and sentiment analysis–an assessment of peoples' belief: A survey. *International Journal of Ad Hoc*. Sensor & Ubiquitous Computing, *4,* 21–33. http://dx.doi.org/10.5121/ijasuc.2013.4102

Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61,* 153–172. http://dx.doi.org/10.1111/j.1744-6570.2008.00109.x

Poeppelman, T., Blacksmith, N., & Yang, Y. (2013). "Big data" technologies: Problem or solution? *The Industrial-Organizational psychologist, 51,* 119–126.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2000). *Comparing the validity of automated and human essay scoring*. Princeton, NJ: ETS.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping E-Rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior, 18,* 103–134. http://dx.doi.org/10.1016/S0747-5632(01)00052-8

Pyburn, K. M., Ployhart, R. E., & Kravitz, D. A. (2008). The diversity-validity dilemma: Overview and legal context. *Personnel Psychology, 61,* 143–151. http://dx.doi.org/10.1111/j.1744-6570.2008.00108.x

Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing, 18,* 25–39. http://dx.doi.org/10.1016/j.asw.2012.10.004

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. http://dx.doi.org/10.3758/BRM.41.3.647

Rudner, L. M., Garcia, V., & Welch, C. (2005). *An evaluation of Intellimetric essay scoring system using responses to GMAT AWA prompts*. McLean, VA: GMAC.

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning, and Assessment, 1*(2). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1668/1512

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. *International Encyclopedia of Education, 4,* 20–26.

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education, 2,* 319–330.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education, 15,* 391–412. http://dx.doi.org/10.1207/S15324818AME1504_04

Received January 31, 2015
Revision received February 16, 2016
Accepted February 20, 2016 ■