Are We Trading One Bias for Another?

Consideration of Computer and Human Evaluation of Resumes

Hannah Hardin

Radford University

A Thesis

Submitted in partial fulfillment of the requirements for the degree of

Master of Arts in Industrial/ Organizational Psychology

Defended November 30, 2016

Jay Caughron (Adviser)

Nicole Petersen (Committee Member)

Nicole Petersen

Tom Pierce(Committee Member)

Thomas W. Pierce

Abstract

Research on resumes has largely focused on biases concerning applicant characteristics, ignoring the influence of decision-making styles on how resumes are analyzed by humans. Additionally, technological advances in resume screening including the use of computer aided text analysis presents a gap in research, which the current study addresses through the comparison of human evaluation and computer aided text analysis of resumes. Researchers predicted that human raters would be more accurate than computer systems when hiring applicants from resume ratings, that computer systems using synonyms would be less accurate than those that used single words when assessing resumes with more ambiguity (i.e. of average quality), and that computer systems would be less accurate than human raters when assessing resumes with ambiguity. Using signal detection theory, results demonstrated that computer systems were more accurate than human raters when ambiguity is introduced, but equally as accurate as human raters for high quality resumes (containing low ambiguity), regardless of using synonym or single word systems. Additionally, research found that human intuition-based hiring was the least accurate method, as well as the most liberal. Human hiring decisions made from ratings and subsequent rankings (logic based) including ambiguous resumes were more accurate than intuition-based methods but less accurate than hiring decisions for non-ambiguous resumes. Human and computer ratings were equally as accurate when hiring high quality (not ambiguous) resumes. The current study provides initial evidence that computer systems used in resume screening provide a valid, reliable alternative to human-based manual scoring of resumes.

Table of Contents

Page

List of Tables

Table of Figures

Introduction

The current study compares human manual scoring of resumes and subsequent hiring decisions to computer automated resume screening and hiring decisions. In the past, particularly in the mid-1970s to early 1990s, research in this area included the effect of applicant characteristics on hiring outcomes and managerial preferences (Duriau, Reger, & Pfarrer, 2007). This research largely focused on biases in human raters such as applicant impression management, attractiveness, applicant age, sex, and academic achievement (Dipboye, Arvey, & Terpstra, 1977; Dipboye, Fromkin, & Wiback, 1975; Duriau et al., 2007; Hutchinson, 1984; Knouse, 1989; Knouse, 1994; Oliphant & Alexander, 1982). However, hiring for race or other protected classes of information is prohibited by law and is not likely to be known or even considered in true selection situations (Equal Employment Opportunity Commission, n.d.).

Additionally, research rarely has considered the processes of scoring resumes related to decision-making when applicant characteristics were not known. Some decision-making literature has addressed the contrast between intuitive and logical decisions, noting that managers prefer the use of intuition compared to logical methods even when logical methods are available (Agor, 1986; Highhouse, 2008; Isenberg, 1984; Lodato, Highhouse, & Brooks, 2011; Simon, 1987). However, the majority of this literature is theoretical in nature; few studies empirically compare intuitive methods and logical methods in selection procedures (Hogarth, 2002). Interestingly, this absence of research highlights a gap that is applicable to what is truly occurring in the selection field. Further, a gap exists in addressing how selection occurs today; as technology becomes increasingly available, computer automated resume screening and content analyses are more commonly used to process resumes than ever before.

In many instances, computer-automated resume screening, or computer-assisted text analysis (CATA) is commonly used to process resumes quickly and efficiently, often in order to bypass human decision making in personnel selection, though the name implies that humans and computers are working side by side. Computer-assisted text analysis is defined as "any technique involving the use of computer software for systematically and objectively identifying specified characteristics within text in order to draw inferences from text" (Kabanoff, 1996, p.1). When used in resume screening, these programs recommend the top candidates from among hundreds, maybe thousands, of resumes and minimize the amount of work that professionals involved in the selection process need to engage in. Often, human resource (HR) or industrial/organizational psychology (I/O) professionals using these systems will only see the top resumes that the CATA program has chosen. What is likely occurring when CATA systems are used is that the CATA systems eliminate the least qualified applicants. Then, human scoring is used to determine the best applicants from the remaining pool. However, it is possible that CATA systems are making the final decisions on which applicants are hired.

Although there has been an increase in popularity and widespread use of CATA systems in selection, empirical studies on such topics are minimal. In the past five years, even well-known news sources have taken interest in computer-assisted text analysis for selecting personnel. A CNN article and an NPR article described the values and uses of automated resume screening (Boulden, 2013; Bradford, 2012), a *Time* article discussed how to "beat the machines" (Cappelli, 2012), and a *Business Insider* article described formatting secrets to help get your resume through the automated systems (Giang, 2013). Research has focused on CATA used with text formats such as trade magazines, scholarly journals, and notes from interviews focusing on strategic management, managerial cognition, and business policy and strategy (Duriau et al., 2007). Thus, the following questions remain unanswered: (1) how are computer-assisted text analyses being used on resumes? and (2) how reliable and valid are these methods?

Though studies on CATA commonly focus on the use of CATA in understanding managerial cognition (Duriau et al., 2007), it is more likely that HR professionals are using CATA for selection processes. In fact, companies such as Monster.com, Indeed.com, and LinkedIn use automated resume scanning and offer these services for both applicants seeking jobs and companies seeking new hires. It has become much easier and cheaper to apply for jobs online, warranting the use of CATA in the hiring process, as it is likely that HR professionals are overwhelmed by the volume of incoming resumes. Yet, concern arises when computer-assisted text analysis is used without human expertise or input as a way to bypass involvement in the prescreening process. "Computer-assisted text analysis" implies that humans and computers are working together, but it is possible that computers are working in place of humans. In some cases, this may be acceptable; however, a situation where a qualified applicant is overlooked because a CATA program failed to find any key words in their resume and an HR professional is not checking that program for validity and reliability is not adequate. Where CATA is used in combination with humans, the CATA system simply screens out the least qualified applicants; however, here the risk of missing qualified applicants is reduced but remains as a valid concern.

Remarkably, a recent study has investigated computer scoring of candidate essays for personnel selection (Campion, Campion, Campion, & Reider, 2016), giving evidence that computer scoring may be a reliable way to select personnel. Researchers utilized CATA on essays where human and computer raters were compared for accuracy when selecting appropriate candidates. Overall, the findings of this study indicate that computer scoring is as reliable as human scoring and that computer scoring may be a valuable alternative to human ratings in specific situations, even demonstrating a low risk of adverse impact (Campion et al., 2016).

Even in the wake of recent studies that investigated personnel selection and the use of CATA, these studies lack external validity; the text samples used by Campion et al. (2016) are not likely what HR professionals use to prescreen applicants. Campion et al. (2016) likely used essays rather than resumes because CATA programs struggle to analyze small amounts of text (McKenny, Aguinis, Short, & Anglin, 2016). However, the reality is that selection decisions are rarely, if ever, based on applicant essays. In fixing the issue of small amounts of text, Campion et al. (2016) create a problem where their findings lack generalizability to real-life selection procedures. Therefore, the current study remains highly relevant in a world of growing technology; we suggest that the use of resumes is more likely than the use of essays in job applications and in-field personnel selection scenarios. In the specific context of resumes, we propose that computer raters may not be as successful as manual scoring by human raters when available text and complete or clear syntax, lexicon, and context are limited.

Human Raters: Intuitive and Logical Decisions

The current study explores both intuitive and logical methods of decision-making, accepting the dual process theory that decision-making styles fit into intuitive and logical categories (Kahneman, 2011). Humans often make decisions based on emotions--or what we describe as "intuition". Typically, intuition is defined as "the preconscious recognition of the pattern and/or possibilities inherent in a personal stream of experience," and is described as a largely subconscious process (Crossan, Lane, & White, 1999, p. 525). Other studies have focused on intuition functioning as a personality trait (Agor, 1986), as a set of actions, an unconscious process(where analysis is at the conscious level, intuition is at the unconscious level), distilled experience, or as a residual category (when decision-making is labeled as logical, then what is left over is considered intuition; Behling & Eckel, 1991). Others have also described intuition as an emotional response, or affect-initiated (Burke & Miller, 1999). In the current study, we refer to an intuitive response as such. In other words, intuition is an emotional response to choosing a resume for hire or not. Additionally, research that has focused on intuition as a function of expertise describes that this expertise arises from experience (Behling & Eckel, 1991). As humans, we are presumably experts at dealing with text and the complexities of language and therefore intuition can also occur as a function of distilled experience.

However biased, humans have the ability to make intuitive decisions and computers do not. Humans are able to give an intuitive reaction (a "gut" reaction) with no "logical" consideration, whereas computers require instructions and logical analyses. Logical processes are typically defined as "thinking which can be expressed in words or by other symbols, that is, reasoning" (Simon, 1987, p. 1). In the dual-process theory, Epstein (2008) describes the logical processes as analytic, affect free, less resistant to change than intuitive methods, and process-oriented. This rational model is described as a deliberate system for information searching (Glockner & Betsch, 2008) or as formal analytical strategies (Mitchell & Beach, 1990). Research has also referred to the dichotomy between intuitive and logical methods of decision making as "system one" and "system two" (Alter, Oppenheimer, Epley, & Eyre, 2007; Glockner & Betsch, 2008; Kahneman, 2011). System one is intuition, and system two is the logical system. System one is automatic, results in a quick response, and is likely to occur when cognitive load is high or when under time pressure. Epstein (2008) describes the intuition system one as resistant to change, automatic, intimately associated with affect, and behavior mediated by "vibes" from past experience. System two is slow, analytical, and deliberate, and is only used when decision-makers have the capacity and motivation to use it (Alter et al., 2007; Glockner & Betsch, 2008; Kahneman, 2011). Where intuition is qualitative, logic is quantitative. In the current study, we define logical decisions as those deduced from a scoring protocol, i.e., a decision that was guided by a deliberate system. Though there has been a lot of debate on which is better, logical or intuitive decision-making, there are few studies that explicitly test the relative validity of these processes (Hogarth, 2002).

Theorists and researchers have debated on the effectiveness of intuition in decision-making. Those that argue for the use of intuition in making decisions cite managerial experiences with using intuition (Agor, 1986). This research showed that executives think using intuition is most useful when facts are limited, there is no precedent on how to act, or when time is limited and there is pressure to make a decision (Agor, 1986). According to Hammond, Hamm, Grassia, and Pearson (1987), intuitive methods may be more accurate than analytical methods. Intuitive methods allow for the consideration of larger amounts of information than analytical methods, and can integrate a lot of information without considerable effort (Betsch & Glockner, 2010). Furthermore, intuition takes advantage of the way our brains are designed, first thinking about things in a subconscious manner and then accessing this information when needed (Burke & Miller, 1999).

A meta-analytic study demonstrated that intuition and logical decision styles are opposites, rather than part of the same process (Wang, Highhouse, Lake, Petersen, & Rada, 2015). However, others state that intuition is part of any decision; even a seemingly purely analytical and logical decision includes some intuitive properties (Salas, Rosen, & DiazGranados, 2009). Additionally, the ability to use intuition is typically built from experience, adding to the validity of intuitive decision-making. When intuition is informed by experience and expertise, Hodgkinson, Sadler-Smith, Burke, Claxton, and Sparrow (2009) suggest that organizations should use intuition. Sadler-Smith and Shefy (2004) note that focusing solely on logical decision-making paints a too simplistic picture of how decision-making actually occurs. Intuition should be considered because intuition occurs involuntarily, and when under time-pressure the use of intuition is necessary. Matzler, Uzelac, and Bauer (2014) identify intuition as the “missing ingredient” for good managerial decision-making, purporting that intuition allows managers to make decisions quickly for large amounts of data. In an unstable work environment, intuitive decision-making was positively related to organizational performance (Khatri & Ng, 2000). Furthermore, we know that intuition is used by executives, so the use of intuition cannot remain ignored, with some noting that intuition had a favorable impact on the quality of decisions made (Burke & Miller, 1999; Sadler-Smith & Shefy, 2004). Research by Gigerenzer and Gaissmaier (2011) shows that when people rely on intuitive decisions (heuristics) these judgments are typically adaptive. Additionally, when these judgments are made by ignoring part of the information provided, judgments were more accurate than weighing all information possible (Gigerenzer & Gaissmaier, 2011).

Supporters of a logical approach in decision-making focus on the errors in intuition and the benefits of a logical system. Hogarth (2002) explored the advantages and disadvantages of logical and intuitive thought, stating that intuition is subject to biases, particularly when people are aware of a rule that can be used to make a decision but are not successful in executing this rule. However, Hogarth (2002) discusses that in logical decision-making the logical process can be complex, leading to errors in the logical processes as well.

Some research suggests that whether a method of decision-making is successful is dependent on the type of task that is being done. For instance, intuition is most successful in a task that is more intuitive and requires approximations in complex situations, where logical methods are most successful when there is a specific formula to come to a decision (Hogarth, 2002; Phillips, Fletcher, Marks, & Hine, 2015; Salas, Rosen, DiazGranados, 2009). When complexity of a particular decision increases, accuracy of the decision is likely to decrease. Adding further ambiguity to whether intuition decision-making is accurate, when people have intuitive preferences, they typically cannot explain why (Hogarth, 2002). Sadler-Smith and Shefy (2004) identify the most common errors in intuitive decision-making to occur as a result of (1) ease of recall (making decisions from easily recalled information), (2) presumed associations (overestimating how related two events are), (3) over-confidence (a feeling of infallibility when making an intuitive decision), (4) confirmation bias (only seeking out confirmatory evidence after making a decision), and (5) hindsight bias (overestimating the degree to which an outcome was predicted).

Research by Isenberg (1984) demonstrated that most successful managers do not follow a logical step-by-step process to make decisions. Similarly, Simon (1987) showed that in general, people are not typically rational decision-makers. Particularly when people want to hire individuals quickly and efficiently, intuition is likely to be preferred over logical methods (Miles, Sadler-Smith, 2014). Specific to selection-related procedures, people have an "inherent resistance" to using logical decision-making methods in selection situations (Highhouse, 2008). Highhouse (2008) suggests this is because people fail to consider selection processes as ones related to prediction and probability and are naturally prone to error. Highhouse (2008) fights against the use of intuition because those who use intuitive decisions lack insight on how they got to a decision, exhibit poor inter-rater agreement, and are more confident in their accuracy when irrelevant information is presented. Additionally, a meta-analysis has shown that intuitive thinking was negatively associated with performance, but that logical (reflective) thinking was positively associated with performance. Interestingly, in this study, time pressure weakened the effect of logical thinking, but not of intuition. The current study addresses the question of which method of decision-making prevails—intuitive or logical—in selection procedures involving resumes.

Human and Computer Raters in Scoring of Resumes

Knouse (1989) showed that choosing a person for hire requires the rater to make an inference about the applicant from how a resume is written and whether or not it matches the ideal job candidate. If there is a match, the person is likely to be chosen for hire. There are two points where errors can be made in these inferences: in (a) understanding the characteristics of an ideal candidate and (b) inferring information about an applicant from his or her resume. In a recent study, Cole, Feild, Giles, and Harris (2009) assess the validity of dispositional inferences from resumes. Cole et al. (2009) found that there were (a) low levels of inter-rater reliability between resume raters on inferred personality traits from resumes, and (b) these inferences that were made did not correlate with the Big Five personality characteristics of the applicants. Therefore, human raters are likely making poor hiring decisions based on incorrect inferences about applicant personalities. Additionally, these poor decisions are being made at low levels of agreement (Cole et al., 2009).

Conversely, one of the greatest limitations facing computerized algorithms is that they are the most sensitive to short phrases or single words. Computer programs will likely fail to identify occurrences in text that humans naturally detect. For instance, humans naturally identify linguistic events that rely on contextual clues and are difficult to operationalize simply into a one or two word segment (McKenny et al., 2016). In instances that require context, McKenny et al. (2016) recommend the use of manual scoring.

Computer programs have some ability to understand word pairings. However, word pairing can cause some issues for CATA systems. For instance, the situation is further complicated by the fact that pairing words with articles like ‘not’ or ‘no’ actually reverses the meaning of the word or sentence. Word pairing errors are related to context clue issues; when certain words are used in combination, it can change the meaning of an entire sentence. Weber (1990), in a review of basic content analysis, states that computer-aided text analysis addresses reliability concerns that have been inherently linked with manual coding; with CATA the coding rules are explicit and are explicitly followed. However, Weber (1990) also describes that there will always be some ambiguity between what is provided to the computer program and the characteristics of the informants, diminishing one’s ability to make accurate conclusions from this information.

To their advantage, humans are capable of detecting impression management and emotionally-laden information, where simple text analysis simply finds frequencies of word occurrences. Computers only consider word counts and quantitative analysis and therefore, cannot make intuition-based decisions; yet, human judgment is inherently influenced by this intuition (Salas, Rosen, & DiazGranados, 2009). According to Pennebaker, Mehl, and Niederhoffer (2003), a reader can be influenced simply through word choice; a specific word may incorrectly display someone’s thoughts and, typically, an emotion is attached to that word. This may be presented in the form of connotations, context dependent-interpretations, job-specific acronyms and terminology, and culturally shifting words not included in a CATA system. Though resumes are not highly emotional pieces of text, a reader will still have some sort of reaction to context clues and word choice, which is likely to impact how humans score a resume—something a computer system is not susceptible to. Though intuition is biasing in some cases, previous research demonstrates that humans consistently, across multiple studies, hire qualified applicants over unqualified applicants when controlling for other biasing variables (Agor, 1986; Betsch & Glockner, 2010; Burker & Miller, 1999; Dipboye Arvey, & Terpstra, 1977; Hammond et al., 1987; Hodgkinson et al., 2009; Knouse, 1994; Matzler et al., 2014; Khatri & Ng, 2000; Oliphant & Alexander, 1982). Therefore, we predict that in an ambiguous situation (one that requires the consideration of context clues), CATA systems will be more likely to count irrelevant information as relevant than human raters, resulting in lower accuracy in hiring applicants.

When testing the accuracy and validity of using CATA systems, lengthy organizational texts have been used. In personality research, CATA has been used to infer psychological states and personality traits. In a study comparing CATA to hand-scoring of text from free speech samples to assess psychological states and traits, it was found that computer scoring was more accurate than human hand-scoring (Rosenberg, Schnurr, & Oxman, 1990). Though computer systems were determined to be efficient, reliable, and accurate, computer-systems were criticized for their potential to exclude appropriate dictionary items, or face difficulty when the meaning of content is reliant on contextual information. Rosenberg, Schnurr and Oxman (1990) recommend that more context-sensitive texts will require programs that are more sophisticated and may paint a different picture of the reliability and validity of CATA. Therefore, we predict that CATA systems that provide scores based on a frequency of key words, including synonyms, will be less accurate than a CATA system that scores based on a single word, when choosing resumes that are ambiguous in quality.

Regardless, Duriau et al. (2007) advocates for the use of computer-aided text analysis, stating that content analysis is nonintrusive and low cost; coding schemes can be adjusted and corrected if any coding flaws are detected, methods can be easily replicated, and the use of computers is faster, easier, and cheaper. Computer-aided text analysis also holds several psychometric advantages. First, the use of CATA in practice has good external validity because of the large amount of data that can be analyzed across many units, enhancing generalizability. Additionally, McKenny et al. (2016) state that CATA has good statistical conclusion validity due to the ability to analyze large samples of tests resulting in high statistical power.

Though quick and easy, CATA systems are most likely to miss information that is being presented and to mistake information as incorrectly relevant. According to Ein-Dor and Spiegler (1995), in a natural language context, a given word can have many meanings or grammatical functions in a given database or document. The assumption is that the way in which the text analysis system organizes the information defines the word's context, and therefore, meaning. In resume screening, if the program fails to appropriately organize this information, then certain information is being counted as fitting a particular resume requirement (this is a false alarm, a false positive, or type I error), or is not counted as the requirement it should be (this is a miss, a false negative, or type II error).

Humans, in contrast, not only understand context clues, but also can give meaning to a sentence or phrase with inferred context. Meehl (1954) describes quantitative (mechanical) methods compared to qualitative (non-mechanical) methods of analysis, purporting that a situation calls for the use of qualitative methods when general rules cannot be applied to available information, and when information unique to an individual is available. Computer algorithms are only able to count and "make sense" of what is available to them and render compiled information in a quantitative fashion. Humans are likely to attribute value judgments based on personal experience and make inferences from unique and contextual information provided in resumes.

Computer-based hiring systems will only be as effective as the rating systems they are given. If these systems are flawed by lack of foresight or understanding of how that system works, then the "garbage in, garbage out" rule applies; poor quality input will produce faulty output. An algorithm that instructs the CATA program to search for only one word will only produce a count for that specific word and an algorithm that instructs that CATA program to search for synonyms and any categories those words fit under will search for just that. Single words or synonym-based algorithms may produce different quantitative answers to a selection problem; synonym-based methods may be more likely to count information that is not applicable and single word methods may be more likely to miss important information. Therefore, in the current study, we explore the use of a strict, single word algorithm and a synonym-based CATA method.

Overall we predict the following:

Hypothesis 1: CATA systems (single word and synonym) will be less accurate than human raters (intuition and logic) when choosing ambiguous resumes.

Hypothesis 2: CATA systems using synonyms will be less accurate than CATA systems using single words when choosing ambiguous resumes.

Hypothesis 3: Human raters are likely to be more accurate than CATA systems in hiring applicants from resume ratings.

Additionally, this leads us to the following exploratory research questions:

Research Question 1: Will human ratings differ based on an intuition based hiring system and a logic-based hiring system?

Research Question 2: Will human hiring systems (intuition and logic) be different from the CATA hiring systems?

Research Question 3: Will a CATA system using single words result in different decisions than a CATA system using synonyms and word categories?

Research Question 4: Are human raters, when using intuition, more conservative or more liberal in hiring applicants?

Research Question 5: Are CATA systems (a single word system and a synonym system) or humans (experts and novices) more accurate at scoring resumes?

Theoretical Overview

In order to answer these questions, we must use signal detection theory. Therefore, the current study utilizes signal detection theory (SDT) to assess the decision-making abilities of humans and computers in evaluating resumes (Swets, Tanner, & Birdsall, 1961; Stillman & Jackson, 2005). Signal detection theory has been historically used in cognitive and perceptual research and is used to assess the accuracy and bias in ratings. Signal detection theory is able to quantify the accuracy of a rater in making a particular decision (e.g. hiring or not hiring an applicant), allowing for an assessment of how conservative or liberal a particular rater is. In SDT, there are events of "signal plus noise" or just "noise". A rater must then detect if a signal is present against a background of noise (Stillman & Jackson, 2005).

In the case of this study, a signal would represent information indicating that an applicant should be hired, and noise would be irrelevant information indicating an applicant that should not be hired. All resumes include noise qualities, demonstrating all applicants have "signal plus noise”. The amount of signal decreases against a background of noise as resume quality decreases.

In SDT, it is possible to calculate how accurate rater decisions are as well as how liberal or conservative that particular rater is (Swets et al., 1961). Conservative decision makers need to be highly confident a signal is present before they say so, and a liberal decision maker does not need to be highly confident in order to be comfortable saying that a signal is present. Liberality and conservativeness, in this study, are representative of how many applicants are chosen for hire; a liberal rater would hire many applicants and one that is conservative is likely to hire few applicants. We then can learn how accurate humans and computers are at selecting appropriate applicants, as well as how liberal or conservative they are.

In the current study, a rater must determine if a signal is present against a background of noise (because signal plus noise is always the case). All resumes have noise items and signal items; some lines of a resume are meaningful and meet rating criteria for hire (signal), and others are not relevant and do not meet any criteria at all (noise). The less signal present relative to the amount of noise present in a resume, the more ambiguous a resume is. For instance, a high quality resume has more signal present than noise, but an average quality resume has an equal amount of signal and noise, containing both hirable and irrelevant qualities. Since a high quality resume contains signal with a relatively small amount of noise, it should be clear that a high quality resume meets many of the hiring criteria and the applicant should then be hired. A resume of average quality represents a signal with a moderate amount of noise, signifying an ambiguous situation. The current study explores both (a) hiring only high quality resumes and (b) hiring average and high quality resumes. The average quality resumes introduce ambiguity to a decision-making situation in the form of more noise. Finally, a resume of low quality represents a signal of almost pure noise; nearly no hiring criteria are met and the applicant should not be hired.

When a signal is present and detected this is considered a “hit”. Further, when no signal is present but the observer claims that a signal was present, this is considered a “false alarm” (Swets et al., 1961). In the current study, a hit would occur when a resume was chosen for hire and was of either high quality (condition a) or of average quality (condition b). Conversely, a false alarm would occur when a resume was of low quality but was still chosen for hire.

When the rates for hits and false alarms are known, accuracy and bias can be calculated (Swets et al., 1961). That is, accuracy is the ability for a rater to detect strong signals against a background of noise. The bias of decision-making refers to how conservative or liberal a particular rater is. Additionally, in the SDT paradigm, there are two other possible situations: a situation where a signal is not present and the rater responds that the signal is not present is a correct decision and is referred to as a “correct rejection.” Then, a situation where the signal is present and the rater responds that the signal is not present is determined to be a “miss” (see Tables 1 and 2; Swets et al., 1961). Table 1 presents the four possible outcomes of a decision as to whether a signal is present using terminology specific to signal detection theory. Table 2 displays comparable outcomes of decision making using the example of hiring or not hiring job applicants. Specifically, the decision here is whether the resume is of particularly high quality or whether the resume is of average or low quality.

Table 1: Signal Detection Outcomes

Decision	Signal Presentation	
	Signal present	Signal Absent
Respond Present	Hit	False Alarm
Respond Absent	Miss	Correct Rejection

Table 2: Signal Detection Outcomes Relevant to Hiring Decisions

Decision	Signal Presentation	
	High Quality Resume	Average or Low Quality Resume
Hired	Hit	False Alarm
Not Hired	Miss	Correct Rejection

In relation to bias, a conservative rater would have a low false alarm rate coupled with a low hit rate, setting a very high standard for hiring applicants. Furthermore, a liberal rater would have a high false alarm rate and a high hit rate, selecting high quality applicants, but also selecting many unqualified applicants. An “ideal decision maker” is one that has both a high hit rate and a low false alarm rate. In other words, an ideal decision maker in this situation would hire everyone they should and nobody that they shouldn’t. Figure 1 displays a common method for displaying visually both the accuracy and the bias in a decision maker’s performance. A common way to display this is through plotting a decision-maker’s hit rate against their false alarm rate (see Figure 1).

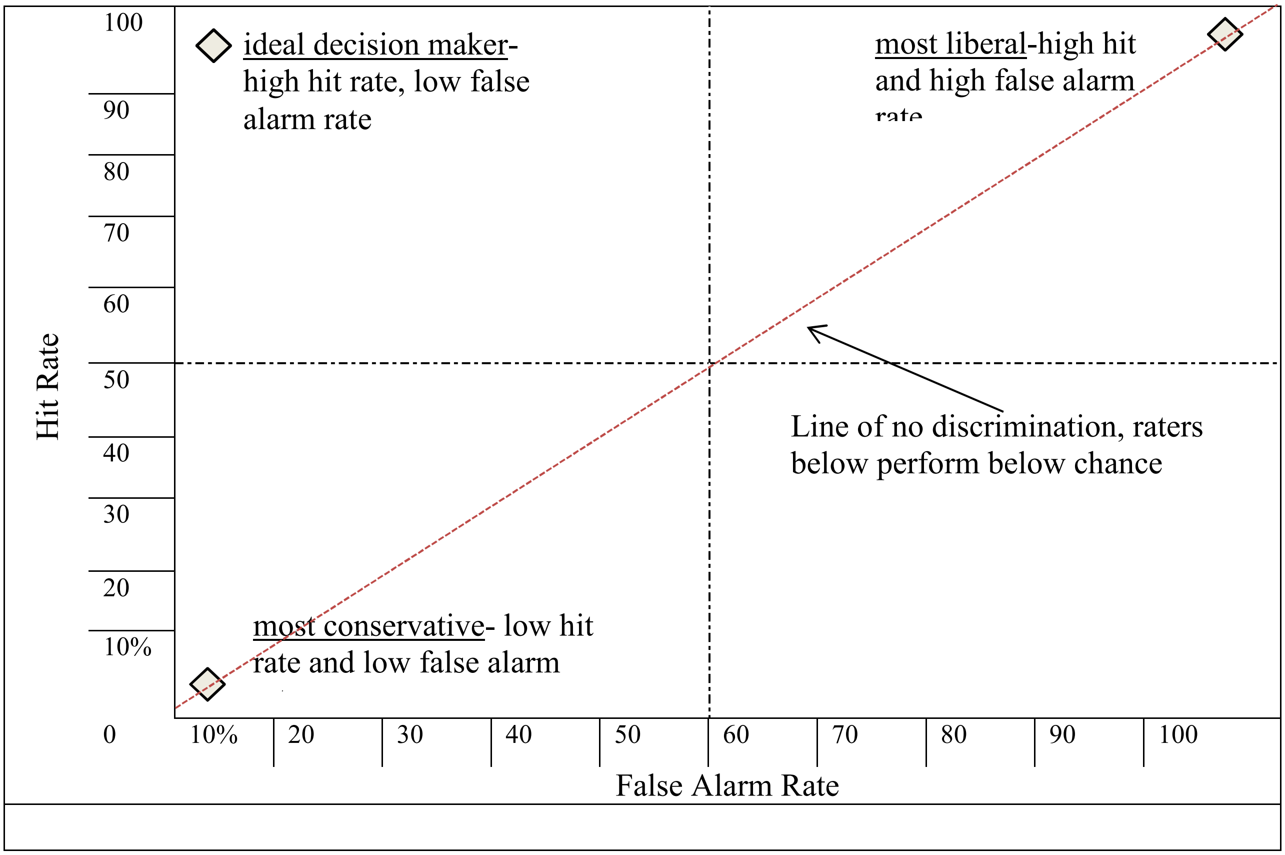

Figure 1: Scatter Plot Example of Types of Decision Makers

In Figure 1, the hit rate is displayed on the y-axis and the false alarm rate represents the x-axis. The performance of a conservative decision maker would be represented as a point located in the bottom left hand corner of the graph, corresponding to the combination of a low hit rate and a low false alarm rate. The performance of a liberal decision maker would be represented as a point located in the upper right hand corner of the graph, corresponding to the combination of a high hit rate and a high false alarm rate. The location of a point between these two corners would be observed in a decision maker that falls somewhere along the continuum of liberal to conservative decision making.

Mathematical variables. In addition to visual analysis of quadrants of decision making outcomes, there are statistics, specific to signal detection theory, that provide quantitative measures that represent accuracy and bias. A value for *d'* provides a measure of the accuracy of a rater. It represents the number of standard deviations separating the means of signal plus noise and the noise distributions (Swets et al., 1961; Macmillan & Creelman, 2004). Mathematically, *d'* is the difference between the *z* score of the hit proportion and the *z* of the false alarm proportion (Swets et al., 1961), or the standardized difference between a rater's ability to detect signal from a rater's ability to detect noise. A large *d'* indicates a rater is better at identifying a signal (should hire) from noise (should not hire).

Another way to confirm accuracy is the use of a receiver-operating characteristic curve (ROC curve). The ROC curve is a way of describing the location of a person's combination of their hit and false alarm rate as previously described in Figure 1. The shape of the ROC curve represents where a rater falls from the conservative (i.e. the bottom left corner) end of decision making spectrum to a liberal end (i.e. the top right corner; Swets et al., 1961). A perfect rater would have an ROC curve that occurs in the same place as the ideal rater (See Figure 1). However, a rater that randomly guesses would have a hit rate and a false alarm rate of .50 for each. This would fit the "line of no discrimination" (See Figure 2).

The area under the ROC curve (AUC) is an alternative approach to quantifying the accuracy of decisions. Expressed as a proportion, the maximum AUC is 1 (corresponding to a one hundred percent hit rate and a zero percent false alarm rate; Swets et al., 1961). The minimum AUC for decision making by chance alone is .5. An AUC of 1 represents a perfect rater in differentiating signal from noise (applicants who should be hired from applicants who should not be hired). When judging an ROC curve, .9-1.0 is considered excellent, .8-.9 is good, .7-.8 is fair, .6-.7 is poor, and .5-.6 is "fail", or no better than chance (TheRMUoHP Biostatistics Resource Channel, 2013).

In Figure 2, rater A's ROC curve is plotted, and the area under the curve of the plot is shaded in. Rater A is an excellent rater demonstrating high specificity and accuracy. Rater B is a good to fair rater demonstrating lower accuracy than Rater A. C is a poor rater, showing little accuracy. Additionally, the calculation for *d'* can be plotted on an ROC curve as the farthest distance from the curve to the line of no discrimination.

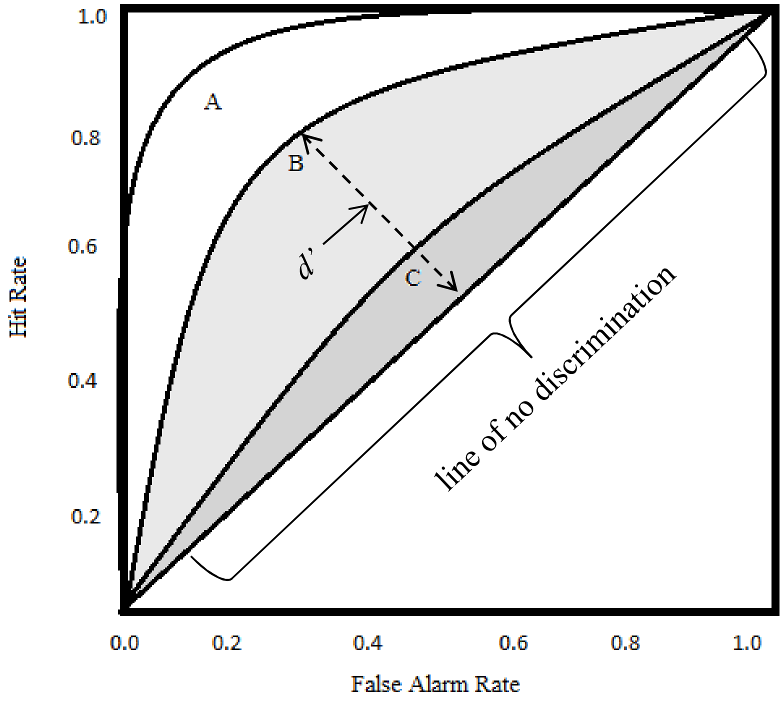

Figure 2: ROC Curve of Different Quality Raters

For determining bias, β ("beta") is used to demonstrate how liberal or conservative a particular rater is. β is a normally scaled quantity and is defined relative to the noise distribution (with a mean and standard deviation of 0 and 1, respectively; Stanislaw & Todorov,1999; Swets et al., 1961). The neutral point of β, where there is no bias, is at the value of 1. A more conservative rater will have a β below 1, demonstrating they are less likely to make a false alarm (or a bias to not hire someone over hiring someone). Further, more liberal raters will have a β above 1, demonstrating a greater likelihood to make false alarms (or a bias to hire someone over not hiring someone; Stanislaw & Todorov,1999; Swets et al., 1961).

Method

Sample

In the current study, accuracy and bias of human evaluations of resumes for hire were compared to computer evaluation of resumes for hire. A random sample of thirty-two undergraduates (“novice raters”) and six graduate students (“expert raters”) obtaining their master’s degree in I/O psychology from a mid-sized university in the southeastern United States were used as human raters. Though research by Dipboye, Fromkin, and Wiback (1975) demonstrated that student raters had inflated ratings compared to in-field professionals when evaluating resumes, their results supported the use of student populations to review resumes as these differences were not significant. Therefore, we determined that the use of students in rating resumes was an acceptable alternative to in-field professionals. The use of graduate student raters as experts functioned as a control group for undergraduate raters to ensure that student ratings were not significantly unlike those of professionals. Computer-assisted text analysis programs were simulated using the text-mining program Tropes, with an algorithm developed for synonyms and key words and an algorithm developed for the use of single key words.

Materials

Scoring protocol. A scoring guide was developed for use in human and computer ratings. This protocol included nine content domains derived from the Virginia Department of Education teaching standards and O*NET. These domains include: (1) content subject knowledge, (2) knowledge of student development, (3) instructional delivery, (4) instructional planning, (5) student assessment, (6) learning environment, (7) team work, (8) professional development, and (9) positive student and parent outcomes. Within each of these domains a set of standards to meet the domains was developed. For example, a standard under the domain of content subject knowledge was “understanding of curriculum” (see Table 3 for the full scoring protocol). Table 3 demonstrates the scoring protocol given to human raters. In the left column is the qualification domain and the right column depicts the standards that fit within that domain. When using this scoring protocol, human raters were instructed to indicate a standard and domain combination with a letter and number. For instance, instructional delivery, standard number two would be referenced as “C2”.

Table 3: Scoring Protocol

Domain	Standard
A) Content subject knowledge	1. Understanding of curriculum 2. Understanding of subject content 3. Keeping up to date with content knowledge of the job 4. Demonstrates an accurate knowledge of the subject matter (math, English, physical education) 5. Demonstrates skills relevant to the subject matter of English, physical education, statistics
B) Knowledge of student development	1. Understanding of developmental needs of student 2. Observe student social development 3. Observe student physical health
C) Instructional Delivery	1. Provides relevant learning experiences to subject 2. Plans and prepares materials for class activities 3. Coordinates extracurricular activities 4. Instructs through lectures and discussion 5. Involves students in learning process through providing hands-on activities for students and providing opportunity for dialogue 6. Adapts teaching methods to meet students’ varying needs and interests
D) Instructional Planning	1. Uses student learning to guide planning 2. Organizes and prioritizes work 3. Plans time for pacing, content mastery, and transitions 4. Develops specific goals for students 5. Plans for individualized and differentiated instruction based on student needs 6. Teaches using Virginia’s Standards Of Learning 7. Teaches using school curriculum

E) Student assessment	1. Uses data to meet the needs of students 2. Uses relevant data to measure student academic progress (e.g. test scores, behavioral data) 3. Instructional content and delivery is guided by data 4. Provides timely feedback to students and students' parents or guardians throughout the school year 5. Expectations for students are developed through assessment data 6. Observe student behavior and performance 7. Use assignments and tests to evaluate progress and document student learning 8. Student assessment is aligned with curriculum standards and benchmarks
F) Learning environment	1. Encourages students and advises students 2. Establishes and enforces rules and policies for student behavior 3. Minimizes classroom disruptions 4. Works with students individually, in small groups, in addition to lecture and instructional time 5. Makes use of routines and resources to provide a positive learning environment 6. Communicates the importance of respect to students
G) Team work	1. Collaborates with other teaching professionals (to develop educational programs, etc.)
H) Professional development	1. Attends training sessions or professional meetings to develop or maintain professional knowledge 2. Document lesson plans 3. Sets personal goals for improvement of knowledge and skills 4. Serves as a contributing member of schools professional learning community
I) Positive Student and Parent outcomes	1. Confer with parents or guardians to resolve student behavioral and academic problems 2. Confer with other teachers, counselors and administrators to resolve student behavioral and academic problems 3. Students demonstrate success through mastery of material 4. Students demonstrate success through meeting Virginia Standards or Learning and curriculum requirements 5. Positive student interactions 6. Positive parent and/or guardian interactions

Resumes. We developed eighteen mock resumes using example resumes found online for teaching positions. We developed two resumes of varying quality (low, average, and high) for three different teaching jobs (Math, English, and Physical Education teachers) totaling eighteen resumes. Because different types of teaching positions may require different context (noise items), resumes were developed for different teaching jobs to control for any potential error associated with job type.

For the eighteen resumes, there were six resumes for each teaching position. Teaching positions included Math, English, and Physical Education. Additionally, two high quality, two average quality, and two low quality resumes were developed for each teaching position. High quality resumes were developed to meet three fourths of the performance standards from the scoring protocol, average quality resumes were developed to meet half of the performance standards, and low quality resumes were developed to meet one fourth of the performance standards. Which standards were represented in each resume was determined via a random number generator for each different standard within each domain (see Table 4 for an example of a resume). Creating mock resumes directly from the scoring protocol was to control for random response error. Random response error is an inconsistent pattern in participant cognitions, which in the current study would represent as inconsistent text, appearing as if the writing of the resume were not consistent in topic or structure (McKenny et al., 2016). In resume development, keeping language and resume structure consistent between resumes was imperative to controlling for this error.

Table 4: Resume Example: Math Teacher Applicant of High Quality

Experience

- Math teacher, XXXX School District, 2011- Present
    - Teaching algebra and geometry
    - Reformed student performance data management by including goal setting for students and frequent data collection
    - Encouraged underprivileged students to succeed despite challenges
    - Taught using school curriculum standards and Virginia Standards of Learning
    - Developed field hockey team, acted as head coach
    - Motivated to provide a successful learning environment for students
- Math teacher, XXXX public schools, 2007-2011
    - Teaching algebra and geometry, using school curriculum and Virginia Standards of Learning
    - Re-instated use of smart boards for interactive learning with students, as well as lead trainings for use in the classroom
    - All students passed standardized testing requirements
    - Developed weekly teacher meetings to collaborate with team members to provide comprehensive learning experiences

Skills

- In tune with student health, emotional needs, and concerns
- Monitoring of student progress through courses
- Use of student feedback and performance (testing and assignment) data to guide teaching and lesson plans
- Use student data to manage flow through course material
- Communication with students and student parents/ guardians to promote student growth behaviorally and academically
- Classroom management through clear rule setting, communication with students about being respectful
- Vary instructional methods including one-on-one teaching methods and group activities
- Personal organization and time management

Education

- Master of Education in instructional technology, University of XXXXX, in year XXXX
- Bachelor of Arts in Secondary Education with extension in Mathematics, State University of XXXXX, in XXXX

Certification

- Skillful Teacher course
- Mastery objective training in literacy and design, 2011
- Differentiated instruction training, 2011
- SmartBoard training courses
- Currently taking courses in teaching algebra and geometry at XXXX University

Areas of interest

- Coaching field hockey
- Reading science fiction novels

Resumes were developed with consideration to empirical research and recommendations on managerial preferences for resume content. Knouse (1994) and Hutchinson (1984) demonstrated that the two most important parts of a resume were education and job experience. As such, these sections were included. All resumes followed the same format and included "noise only" items that did not meet any of the scoring protocol. Because previous research indicated that applicant information related to sex, marital status, and attractiveness led to more or less favorable ratings (Dipboye et al., 1977; Dipboye et al., 1975), the current study does not include applicant information except the content of the resume relating directly to scoring protocol.

Job descriptions. Job descriptions were developed to represent the particular job the resumes were to be applicants for. These were developed using online job descriptions and O*NET (see Table 5). Participants were instructed to utilize these to make a hiring decision.

Table 5: Job Descriptions

Subject	Description	Certification and Licensure	Education	Experience
English Teacher, Grades 9-12	The primary function of an English teacher is to teach the English language through reading, speaking, and writing to develop student skills. Additionally, an English teacher will foster student listening, speaking, reading, writing, and an appreciation of literature.. Duties include teaching grammar, reading comprehensions, writing and understanding of books, poetry and other writings. English teachers are expected to: create lesson plans and teach those plans to the entire class, create tests and assignments, grade tests and assignments, meet with colleagues to coordinate lesson plans, manage students and the classroom, hold conferences with guardians and students. An English teacher will be	Required: Hold or be eligible for Virginia teaching license in secondary English. Preferred: Professional certification through the National Board for Professional Teaching Standards.	Required: Hold a Bachelor's Degree from an accredited college or university in English and/or teaching with a concentration in English is required. Preferred: Master's degree in position-relevant subject matter is preferred. Possession of a credential authorizing	Successful prior teaching experiences in English and teaching English is preferred. Experience in some or all courses or teaching in: Shakespeare, American and British literature, adolescent literature, world literature, grammar, poetry, drama, literary criticism, and composition are required. Courses in teaching methodology are preferred.

	responsible for preparing students and providing students with the necessary skills to succeed on standardized tests.		service as a teacher of secondary level students.	Teaching experience, at minimum, under the guidance of a classroom English teacher.
Math Teacher, Grades 9-12	Secondary Math teachers work with the goal to help students develop critical-thinking abilities by gaining an understanding of mathematic concepts, mathematical skills and basis understanding of the structure of mathematics. The high school math teacher is responsible for developing competencies in mathematical skills to prepare students to meet the Virginia standards of learning. Math teachers will instruct students, create lesson plans, assign and grade tests and assignments, manage students in the classroom, and help students prepare for standardized testing. High school math teacher duties include preparing students for college-entrance exams, teach general math and specific courses in geometry and algebra, developing and using lesson plans and supplementary materials compatible with the course standards of learning.	Required: Must be eligible for or hold a valid teaching license with endorsement in Algebra and Geometry or other secondary mathematics. Preferred: Professional certification through the National Board for Professional Teaching Standards. Possession of a credential authorizing service as a teacher of secondary level students	Required: A Bachelor's degree from and accredited college or university in math and/or teaching with a concentration in math is required. Preferred: A master's degree in position-relevant subject matter is preferred.	Successful prior teaching experiences in math and teaching math is preferred. Experience in some or all courses or teaching in calculus, statistics, geometry, and algebra is required. Courses in teaching methodology are preferred. Teaching experience, at minimum, under the guidance of a classroom math teacher.
Physical Ed. Teacher, Grades 9-12	Physical Education teachers provide students with learning experiences in comprehensive health and physical education and supervision of students in a supportive and positive climate that develops in each student the skill, attitudes and knowledge to meet and exceed the Virginia's core curriculum content standards. PE teachers will teach students about good body function and exercise; motivating each student to cultivate physical fitness, hygienic habits, and good social and emotional adjustment; discovering and developing talents of students in physical achievement; and developing strength, skill, agility, poise, and coordination in individual, dual and team physical activities and	Required: Must be eligible for or hold a valid teaching license with endorsement in Physical Education. Certification for Physical Education Preferred: Professional certification through the National Board for Professional Teaching Standards. Possession of a credential	Required: A Bachelor's degree from and accredited college or university with a major or minor in physical education. Preferred: A master's degree in position-relevant subject matter (sports studies or a related instructional field is preferred.	Successful prior teaching experiences in physical education are preferred. Experience in some or all courses or teaching in physical, health, and general topics such as philosophy, kinesiology, human development and educational psychology and or teaching methodology is preferred. Teaching experience, at minimum, under the guidance of a classroom physical education teacher.

sports, in accordance with each student's ability. PE teacher duties are to: organize games and challenges that promote physical activity in students, develop student proper exercise and eating habits.	authorizing service as a teacher of secondary level students. Subject area exam in physical education and a basic skills exam.	

Computer algorithm. In accordance with the eight steps outlined by Weber (1990), we developed the computer-assisted text analysis algorithms. In order to control for specific factor error, algorithms were developed and tested on sample resumes and checked for accuracy and validity of found words. Specific factor error is when the measure itself influences the data provided by the respondents (McKenny et al., 2016). In the use of CATA, researchers provided words that demonstrated the underlying phenomena from the scoring protocol on the basis of judgment. In specific factor error, the inclusion of unfitting words or omission of essential words would produce errors in CATA ratings that were not attributable to the CATA system itself.

Single word and synonym lists were developed for each standard within each domain to enter into the computer program Tropes (for an example see Table 6). A dictionary was developed in Tropes for each job type and domain combination, in recognition that, for example, “content knowledge” may require different key words for each job type and, therefore, require a separate algorithm.

Table 6: Computer Scoring Scenario

Domain:
CONTENT SUBJECT KNOWLEDGE
ENGLISH

Standards in Domain

1. Understanding of curriculum
2. Understanding of subject content

3. Keeping up to date with content knowledge of the job
4. Demonstrates an accurate knowledge of the subject matter (math, English, physical education)
5. Demonstrates skills relevant to the subject matter of English, physical education, statistics

key word	**synonyms**
Accurate	accurate correct exact precise
Content	content composition design structure
curriculum	curriculum educational program study syllabus
English	English English language
Skill	skill accomplishment competence experience expertise
up-to-date	up-to-date advanced current up-to-the-minute

In creating the single word algorithm, single words that described each standard were chosen to include in the scenario. For the synonym algorithm, those single words were used and synonyms for those words were chosen from an online thesaurus. Researchers made informed decisions to include or exclude certain synonyms based on their relevance to the standards. For instance, for the single word "content," synonyms could include: composition, design, structure, but could not include: substance or idea.

Tropes provides the scoring approach that Pennebaker et al. (2003) describes as a word count, irrespective of the context in which it occurs, allowing for linguistic information “from a distance.” In the single word system, the computer algorithm resulted in a count of how many times each word appeared in each resume. Rosenberg et al. (1990) describe that CATA methods can use basic tagging operations which then put words into category descriptions (for instance, physical education may also include references to athletic events). In the synonym method, the computer algorithm included this tagging operation and provided frequencies of the number of times a single word, that single word’s synonyms (which researchers provided), and words within those categories were found. This count resulted in the resume’s final computer-based score.

The use of two separate algorithms helped to control for and measure algorithm error. Algorithm error is specific to the use of computer analysis of text, where algorithm error would occur if two different methods of analysis produced different scores on the same text (McKenny et al., 2016). From this, we can determine if the algorithms function differently, and therefore, if algorithm error is present in the current study.

It is important to note that in the development of materials, the algorithm was made directly from the scoring protocol and that the mock resumes were developed to meet the specific content in the scoring protocol. Both the humans and computers were “set up to win” in efforts to reduce error in all other areas.

Procedure

Human participants rated the resumes using the scoring protocol in a lab with five to ten other participants per session at a maximum of two and a half hours. Participants were told to review the job description for a particular job (i.e. English teacher, Math teacher, or Physical

Education teacher) and then were asked to evaluate applicant qualifications for those jobs using the scoring protocol provided.

When using the scoring protocol, participants were asked to indicate at each line of each resume which content domain and standard was met, for a maximum of one standard per line. Additionally, participants were instructed that for some lines of the resumes no standard would be met, and in that case, to leave that line blank. In the occurrence of two standards being met, they were instructed to choose the one that best fit. This task was comparable to the computer's task of finding words that matched each standard.

Resumes were presented in randomized order to human raters within packets of each job type. Which job type packet was received first was also randomized. Randomization of presentation was done to prevent fatigue effects, as this task was particularly long and cognitively demanding. Human rater scores indicated which content domain was addressed at each line of each resume, and were totaled for statistical analysis by frequency of standards met. This frequency was then totaled to provide a final score for each resume by each rater.

Additionally, humans were asked to make a hiring decision based on the information given from the logical scoring process. That is, human raters gave a score based on an "intuitive" hiring decision, which at the bottom of each resume asked, "Would you hire this applicant for the job? Circle yes or no". This intuitive decision functioned as a control group to the logic-based and computer rating systems to demonstrate human hiring decisions without any scoring protocol. This method, though referred to as "intuitive" is not purely an intuitive response; logical scoring information was available to participants before an intuitive decision was made. This was done to prevent confirmation bias toward raters confirming their intuitive response in their subsequent logic-based ratings.

A "logical" hiring decision could then be determined from the total score of each resume, where resume scores were ranked from the total score for each rater, and the top scores were determined to be hired within each job category. Additionally, in order to consider ambiguity, hiring decisions were calculated for both humans for the top two hires (which should match the high quality resumes) and top four hires (should match the high and average quality resumes) from each packet. A hire four condition was considered more ambiguous because it considered resumes that were of average quality. This logic-based decision was not made by the participant directly, but was determined from the total scores and subsequent ranking of resumes. That is, the top two and top four hires for each rater were determined through the rankings that were determined from total scores.

For computer scores, we entered scenarios into Tropes, which gave a frequency of the presence of words in each resume. Computer algorithms included a score for each resume including synonyms and an algorithm to score each resume by searching for only single words (i.e. no synonyms). Total scores for both humans and computers were calculated via a count of either domains addressed in the resume or key words found from that domain. Similar to the logical human hiring, computer scores were ranked within each job category and the top two resumes were determined as the hired applicants, and, in another condition, the top four resumes were determined as hired. In contrast, in the "intuition" hiring method, raters chose as many applicants for hire as they felt should be hired, with no other restrictions.

Statistical analyses. SPSS Statistics for Windows, Version 21 (2012) was used for graphing for visual analysis, aggregation, and calculations for AUC and ROC curves. An online detection theory calculator was used to calculate *d',* false alarm and hit proportions, and β (ComputerPsych LLC, 2011).

Results

For intuition raters, the maximum number of hires was eighteen, the minimum of hires was four, and the average number of people hired was *M*=11, *SD*=4.07. For the human logic and the CATA hiring systems, the number of people hired was determined by the researchers and held at two hires or four hires.

In using SDT to analyze data, hire decisions were coded as a hit (hired and should have been hired), miss (did not hire and should have been hired), false alarm (hired and should not have been hired), or correct rejection (did not hire and should not have been hired). This was done for the "hire two" applicants condition and the "hire four" applicants condition.

Bias

The first analysis available through SDT is visual. See Figures 3 and 4. Note that in Figure 3 and Figure 4 novices and experts were not separated. This is supported by Dipboye et al. (1975), who supported the use of students in place of resume ratings by expert raters.

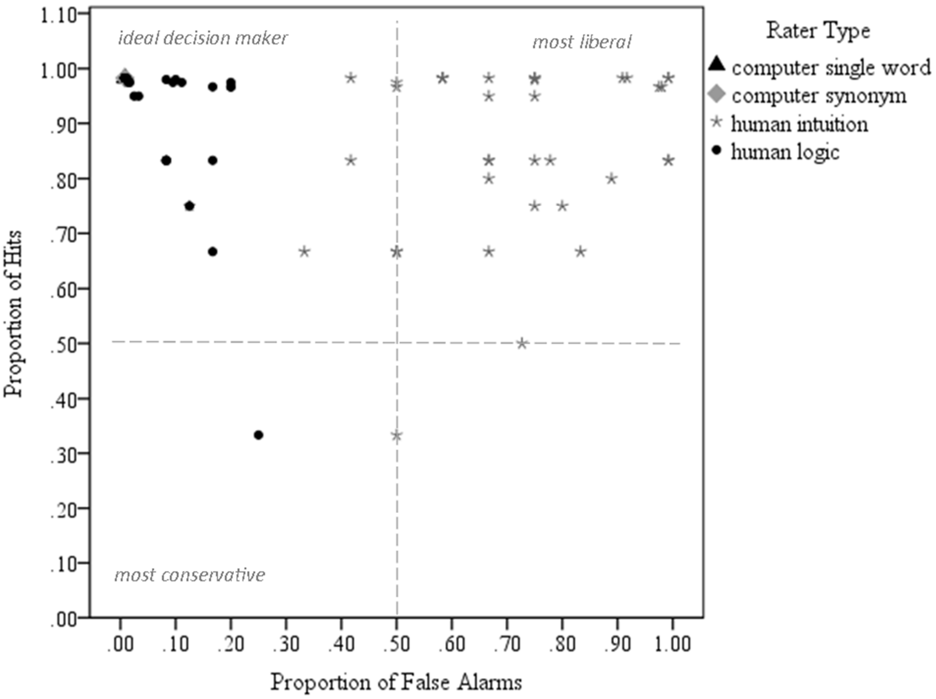

Figure 3: Signal Detection Theory: Hit and False Alarm Proportion, Hire 2

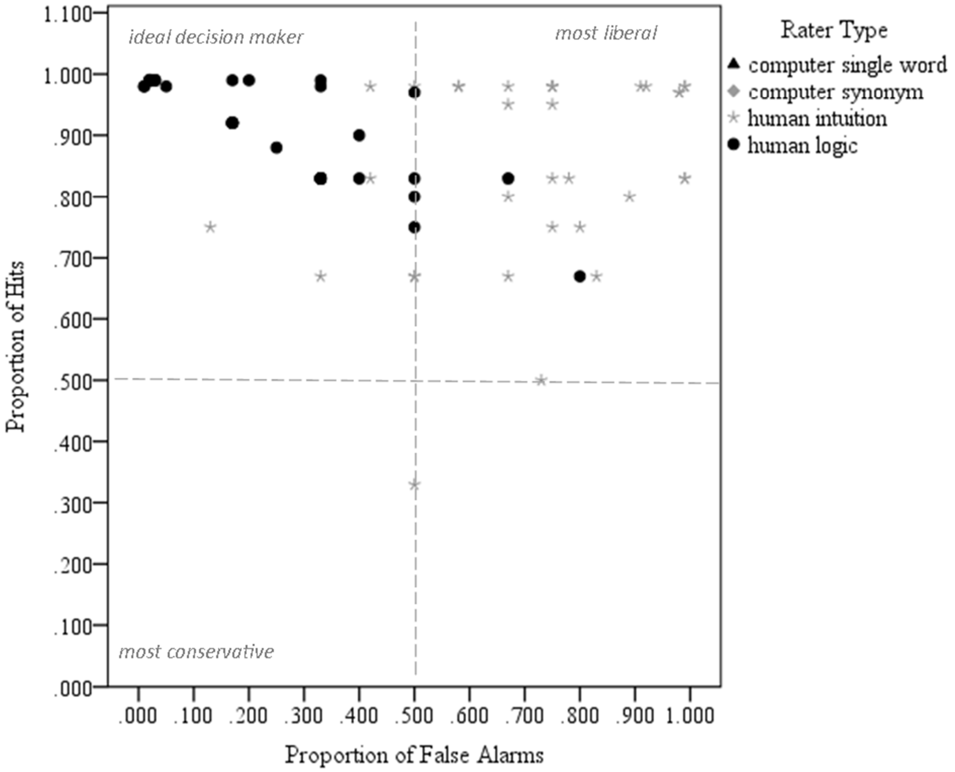

Figure 4: Signal Detection Theory: Hit and False Alarm Proportion, Hire 4

Computer single word, computer synonym, and some of the human logic based ratings are all in the “ideal decision maker” area of the graph, even overlapping one another significantly. This is represented in Figure 3, showing the hire two applicants condition. Additionally, for the hire two applicants condition, β = 1.82 for novice and expert logic decisions, and for the computer synonym and computer single word methods. The human logic and CATA systems appear to be more conservative because the number of hires was held constant for the hiring conditions at hire two applicants. This β indicates there is a bias for the raters to not hire more applicants than were hired. For the intuition condition, β = .17 for novice raters and β=.13 for expert raters, indicating a liberal bias; intuition ratings resulted in hiring more applicants than were not hired. This addresses research question 4, demonstrating that intuitive decisions are more liberal in hiring applicants than logical methods.

For Figure 4, representing the hire four condition, the human logic decisions start to appear more like the human intuition hiring decisions, moving from the ideal decision maker to the most liberal side of the graph. For human logic and computer conditions β=.55 in the hire four condition. A β of .55 indicates a more liberal bias (more liberal than the hire two condition, but less liberal than intuition-based decisions). Again, the β value is an artifact of the hire four condition necessarily controlling for the number of people that were hired.

Overall, human intuition ratings, though varied, are likely to over hire, resulting in both a high hit rate and a high false alarm rate, demonstrating a liberal bias. Further, when forced to hire only two applicants, both the human logic system and the CATA systems demonstrate a low false alarm rate and a high rate of hits, which is associated with a high degree of accuracy.

When expanding to a hire four applicants condition, the human logic raters and computer options seem to “drift” apart (in the visual analysis); human logic ratings become less accurate due to a greater proportion of false alarms, appearing more similar to human intuition ratings. From the visual analysis it can be concluded that human logic based systems are similar to CATA systems only in the hire two condition. As ambiguity increases by increasing the number of hired applicants to the hire four condition, human logic ratings become less accurate than CATA systems. Overlap in the hire two and hire four options indicates that logic based judgments from humans are not unlike the computer single word and computer synonym CATA systems.

Visual analyses suggest that human intuition-based decisions demonstrate a more liberal approach to hiring. Additionally, visual analyses show that human logic systems, when hiring four applicants (with the introduction of ambiguity), are more similar to intuition based decisions than CATA systems. These CATA systems appear to be highly accurate and the “ideal decision makers”. As such, the visual analyses provide initial evidence that Hypothesis 1 is not supported. In fact, Hypothesis 1 predicted that all CATA systems would be less accurate than human raters, but the opposite is true.

Additionally, visual analyses suggest that Hypothesis 2 may also not be supported. The visual analysis gives preliminary evidence that single word systems and synonym systems are equally accurate when choosing average quality resumes, though Hypothesis 2 predicted synonym based CATA systems would be less accurate than single word systems. Finally, visual analysis indicates that Hypothesis 3 may not be supported and, again, the opposite of what was predicted is true. That is, Hypothesis 3 predicted that human raters were likely to be more accurate than CATA systems in hiring decisions. However, human raters are not more accurate than CATA systems in hiring applicants, with intuitive decisions being the least accurate, followed by hiring four, then hiring two applicants. These visual analyses provide initial evidence that all three hypotheses are not supported; however, considering *d’* as a measurement of specificity provides more evidence in answering these hypotheses.

D-Prime Analysis

Recall that *d’* denotes the distance between the signal and the noise, which represents the accuracy of your decision makers. Mathematically, *d’* is the number of standard deviations separating the means of the signal plus noise and noise distributions (Abdi, 2007; Swets et al., 1961). A rater that obtains a larger *d’* is better at distinguishing a signal (should hire) from noise (should not hire).

Reported measures of *d’* were compared using “overall” human intuition and human logic scores. In order to determine the overall scores for all human logic decisions and all intuition ratings, a resume was determined as hired when fifty percent or greater of the raters decided to hire that applicant. This method is subsequently referred to as "aggregation" of the hiring decisions. Because CATA systems already provided a single decision per resume, no aggregation was required.

For the hire two and hire four applicants conditions, *d*' for computer single word and synonym algorithms showed *d'*=4.52. Additionally, for both novice and expert logic decisions after aggregation, *d'*=4.52. For the hire four condition, novice logic decisions remained at *d'*=4.52 and expert logic decisions decreased to *d'*= 3.36. For intuition hiring methods, novices made decisions at *d'*=1.16*,* and experts made decisions at *d'*= 1.45. For a summary of *d'* values see Table 7.

Table 7: Mathematical Interpretations of Signal Detection Theory

Choice Type	AUC	Hit Proportion	False Alarm Proportion	*d'*
Computer Single Word				
hire 2	1.0	0.98	0.01	4.52
hire 4	1.0	0.99	0.02	4.52
Computer Synonym				
hire 2	1.0	0.98	0.01	4.52
hire 4	1.0	0.99	0.02	4.52
Human Novice				
intuition overall	.634	0.98	0.83	1.16
hire 2 overall	.939	0.98	0.01	4.52
hire 4 overall	.862	0.99	0.02	4.52
Human Expert				
intuition overall	.753	0.98	0.75	1.45
hire 2 overall	.902	0.98	0.01	4.52
hire 4 overall	.901	0.99	0.17	3.36

Therefore, both single word and synonym CATA systems, expert and novice logic based decisions for the hire two condition showed a *d'* of 4.52, demonstrating that they are all good at discriminating between when a signal is present (an applicant should be hired) and noise is present (an applicant should not be hired). However, when hiring four applicants, human experts demonstrate a decrease in the ability to discriminate between qualified (signal) and unqualified (noise) applicants at *d'*= 3.36. Novice and expert intuition ratings also demonstrated poor discrimination with *d'* values of 1.16 and 1.45, respectively.

Following the consideration of *d'*, we can still say that Hypotheses 1, 2, and 3 are not supported. *D'* demonstrates that computers are more accurate than human expert raters when choosing average quality resumes, and as accurate as human novice raters when choosing average quality resumes. Additionally, Hypothesis 2 is also not supported; single word and synonym CATA systems are equally accurate and sensitive when choosing two and four resumes for hire. Finally, Hypothesis 3 is not supported; human raters seem to be less sensitive than CATA systems, particularly, human expert raters and intuition (novice and experts) ratings demonstrate less discrimination than CATA systems. However, human novice logic systems are as accurate as CATA systems in hiring applicants in both the hire two and hire four applicant conditions.

Furthermore, when taking into account *d'* values, we can tentatively answer our research questions: (1) Human ratings were different based on an intuition hiring system and a logic-based hiring system. Specifically, *d'* values indicated that the intuition methods were the least accurate and that the logic-based systems were much more accurate in hiring applicants. (2) Human intuition and logic systems were different from CATA systems. Specifically, intuition methods were the least accurate, followed by logic-based at the hire four condition, and the logic-based hire two condition and all CATA systems were equally accurate. (3) CATA systems using single words were equal to CATA systems using synonyms. (5) CATA systems were overall more accurate than humans at scoring resumes (though there was no difference for logic-based systems at the hire two condition). Research question 4 cannot be addressed through *d'* analyses.

ROC Curve and Area Under the Curve

When considering the ROC curves for each of the groups, please refer to Figures 5-14; additionally, refer to Table 7 for the reported area under the curve (AUC). The ROC curves for all CATA systems (single word and synonym) are the same, indicating an AUC of 1 for both the hire two and hire four applicant conditions. This indicates that the computers are considered the best possible raters, demonstrating high accuracy. These findings fail to support Hypothesis 2, and answer research question 3, indicating the single word and synonym systems are equally accurate, regardless of ambiguity due to choosing two or choosing four applicants. See Figures 5-8 for CATA system ROC Curves.

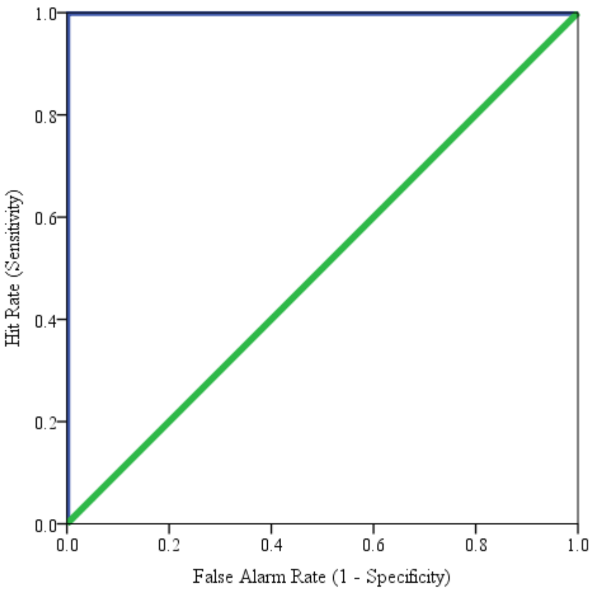

Figure 5: ROC Curve: Computer Single Word, Hire 2

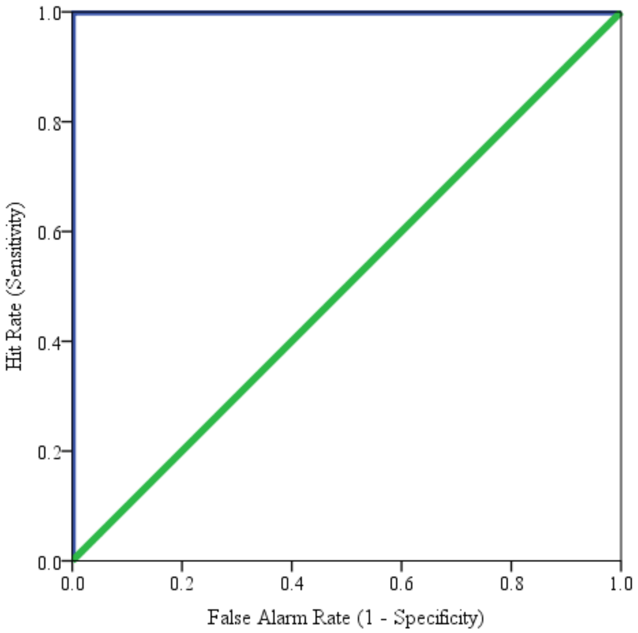

Figure 6: ROC Curve: Computer Single Word, Hire 4

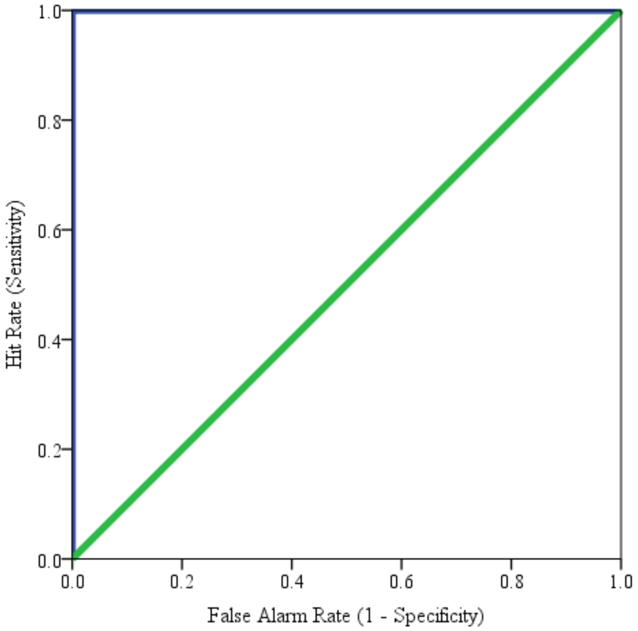

Figure 7: ROC Curve: Computer Synonym, Hire 2

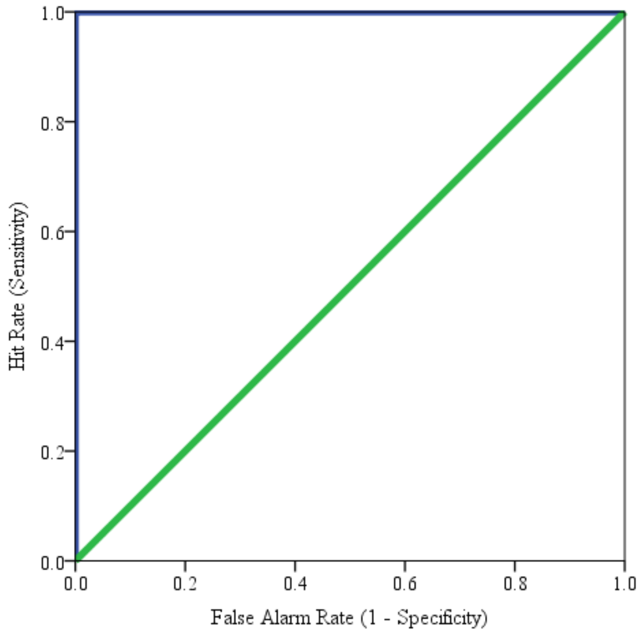

Figure 8: ROC Curve: Computer Synonym, Hire 4

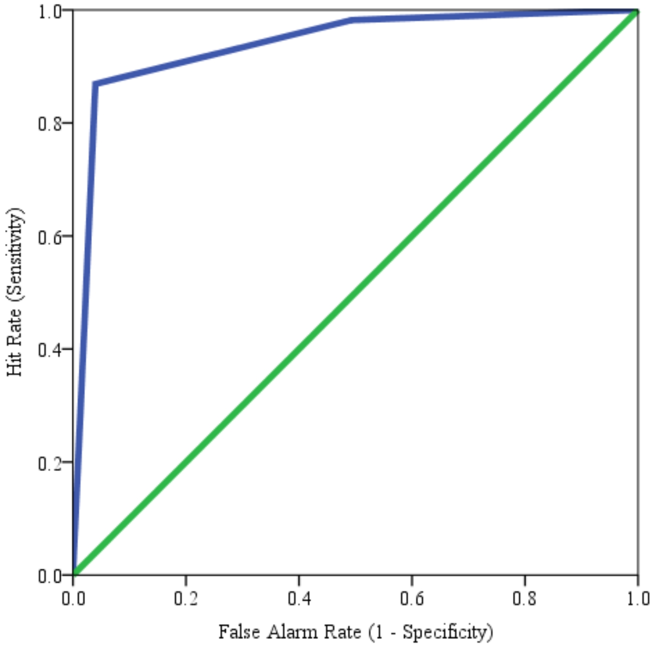

Figure 9: ROC Curve: Novice Logic, Hire 2

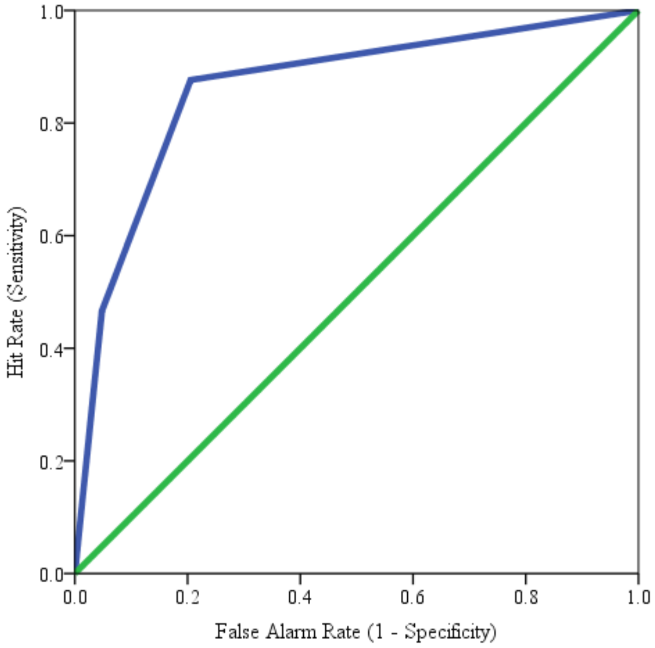

Figure 10: ROC Curve: Novice Logic, Hire 4

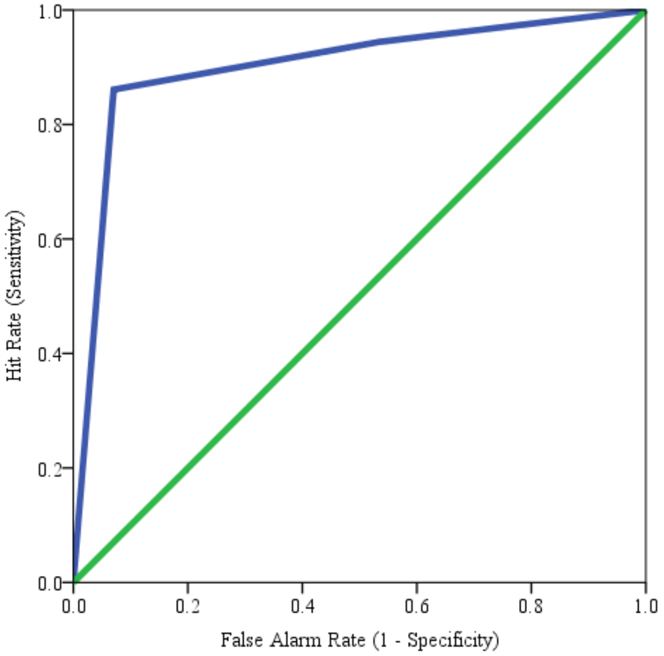

Figure 11: ROC Curve: Expert Logic, Hire 2

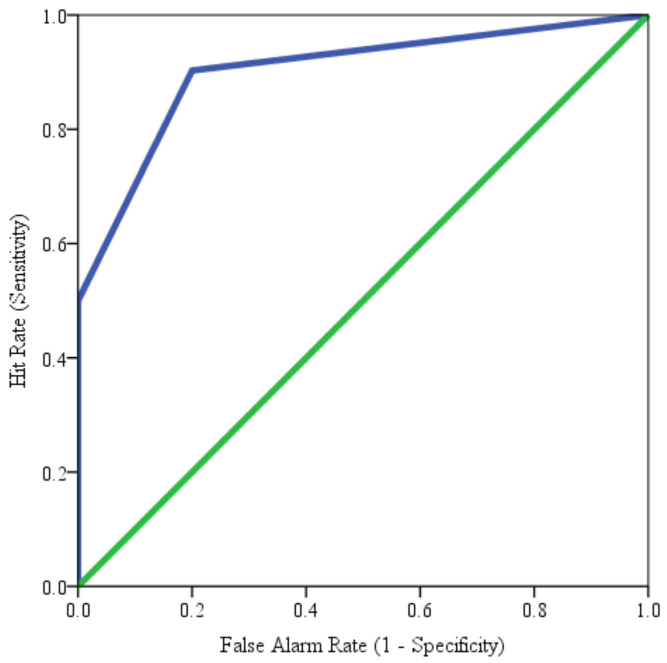

Figure 12: ROC Curve: Expert Logic, Hire 4

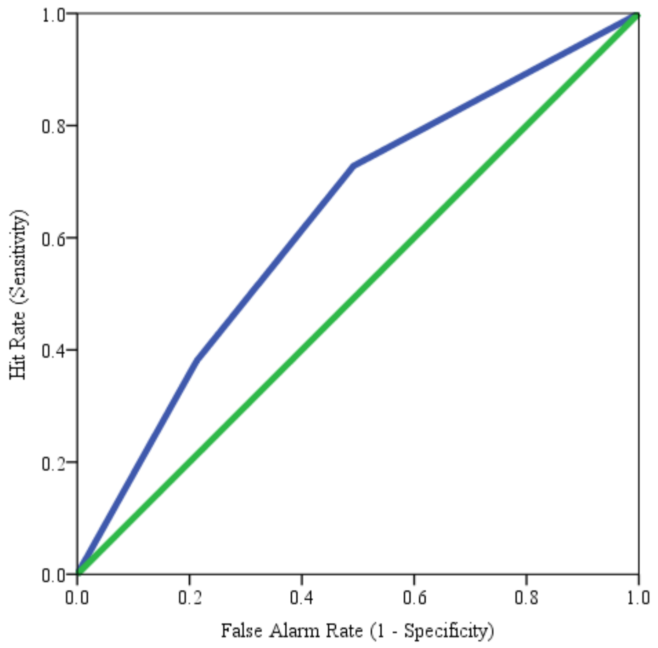

Figure 13: ROC Curve: Novice Intuition

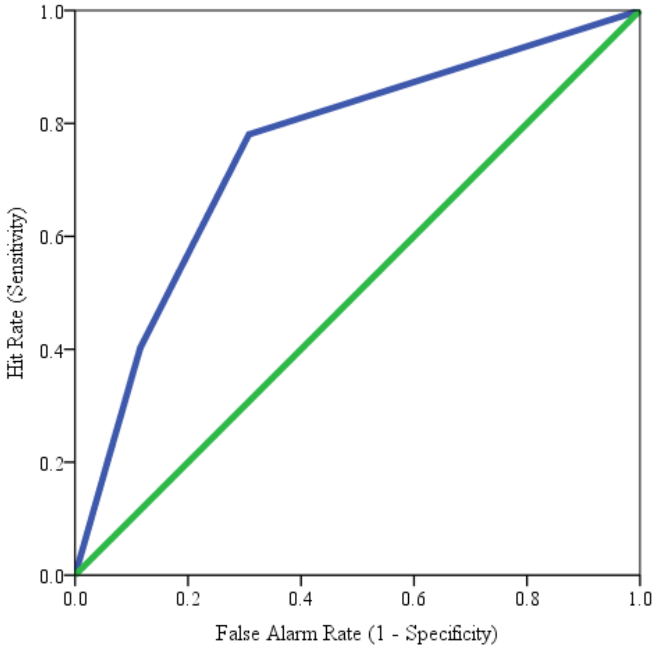

Figure 14: ROC Curve: Expert Intuition

Additionally, when comparing human novice logic and human expert logic systems, refer to Figures 9-12 for their ROC curves. For the hire two condition, the curves demonstrate that expert raters may have lower accuracy than novice raters. Conversely, the ROC curves demonstrate that expert raters have better accuracy than novice raters in the hire four applicants condition, when ambiguity is introduced. The AUC for novice raters for hiring two applicants was .94 and .86 for hiring four applicants, and expert raters at hire two applicants was .902 and .901 for hire four applicants. Furthermore, when comparing all four logic-based decisions to the four computer-based decisions, it is clear that the computers are more accurate raters.

Results showing a decrease in AUC between hire two and hire four applicants for all human raters is an indication that human raters struggle with ambiguity (i.e. choosing four applicants which includes average quality resumes with more “noise” and less “signal”). However, this may be most pronounced for novice raters. It seems that expert raters, though their

AUC decreased as well, may not struggle to a significant degree with ambiguous resumes. As such, results continue to indicate that Hypotheses 1 and 3 are not supported.

Finally, when considering the control group of the intuition choice, experts demonstrated an AUC of .753 and novices showed an AUC of .634. The expert AUC fit into the “fair” rater category, and novices fit into the “poor” rater category. The AUCs for intuition provide evidence that when human raters are left to use their intuition they hire applicants with poor accuracy. This indicates that human intuition raters are the poorest quality raters in the study and, regardless of expertise, are not likely to yield accurate decisions.

Discussion

Human Intuition and Logic Systems

Results indicated that human logic systems were much more accurate in determining signal (should hire) from noise (should not hire) than human intuition systems; this was demonstrated through *d'*, AUC, and ROC curves. Additionally, the visual interpretation of graphs of false alarm and hit proportions demonstrated that human logic systems were the "ideal decision makers" and that the human intuition methods were the "most liberal" decision makers. Humans and computers are different, but this is largely when ambiguity is introduced. When ambiguity is introduced, human rater accuracy decreases but computer accuracy remains constant. With minimal ambiguity, logical rating systems and computer systems are the ideal decision makers and have high accuracy. However, as ambiguity was introduced, the logical system raters seem to "lean" more towards an intuition-based decision, decreasing in accuracy and increasing in liberality. This provides initial evidence that as ambiguity is introduced, humans start to rely on their intuition. This supports research and theory that purports that no decision-making is without intuitive processes (Salas et al., 2009). These findings also support the notion that as the complexity of a decision increases, intuition is more likely to be utilized (Betsch & Glockner, 2010; Miles & Sadler- Smith, 2014; Matzler et al., 2014; Salas et al., 2009). Furthermore, this fits research by Hogarth (2002) that states that as analytical complexity (in the case of this study, ambiguity) increases, accuracy is predicted to decrease.

With aggregation, *d'* values for experts demonstrate that experts struggle more with ambiguity than novices, however, without aggregation when considering AUC, there is initial evidence that experts struggle with ambiguity less than novices. Research has indicated that experts and novices reason through situations differently (Vanlehn, & Van de Sande, 2009). A novice is said to make decisions in a situation analytically and an expert can make the same decision more intuitively; a novice will have a superficial understanding and follow a provided process where experts will display clear plans for solutions. From this research, we would expect to see our experts perform better than novices when ambiguity is introduced and in the intuition-based decisions. We did not find this. It is possible that novices are less distracted by ambiguity and the need to utilize context more than expert raters, or error occurred in the aggregation of data. On the other hand, this error may be due to the low sample size for expert raters of *n*=6 and the relatively high sample size for novice raters at *n*=32.

Due to differences in the results depending on aggregation, no conclusions can be made specifically about experts and novices. Additionally, according to Vanlehn and Van de Sande (2009), expertise arises from practice with a particular task; however, the experts in this study did not have practice in the task of rating resumes—we considered expertise as content knowledge about the selection process. The distinction between experts and novices in the current study is likely not accurate. Therefore, we suggest that future studies should explore the use of true experts.

Notably, for the intuition method, raters were instructed to use their ratings on each resume from the logical system to make a final hiring decision. Even when the logic system was available for people to see and use to inform their intuition-based systems, they still were not particularly successful at hiring the correct applicants in the intuition condition. This gave further evidence that humans are poor decision makers when biases are allowed to be a part of the decision making process.

Based on past research, it should not be surprising that intuitive methods resulted in completely different results than logical methods. This finding is supported by a meta-analysis which showed that intuition and logic decision styles were uncorrelated constructs (Wang et al., 2015). When logical decision-making resources were available, raters refused to use the logical methods to inform this "quick and easy" decision. This supports research from Highhouse (2008), Isenberg (1984), Lodato, Highhouse, and Brooks (2011), Fisher (2008), Miles and Sadler-Smith (2014), and Mitchell and Beach (1990) which indicate that decision-makers prefer to use intuition, which can arise involuntarily, particularly in selection processes. Additionally, this fits with previous research that has shown that intuitive decisions are often less accurate (Phillips et al., 2015). Salas, Rosen, and DiazGranados (2009) sated that intuition can be highly accurate if the decision task matches the decision type (an intuitive decision is paired with a judgmental task in a complex situation) and decision environment (with time pressure). In the current study, both situations were met, yet intuition remained the least effective method of decision-making.

The current study presents intuitive methods as the least efficacious, being the least accurate and most liberal methods. This is contrary to research and theory written by Agor (1986), Burke and Miller (1999), Gigerenzer and Gaissmaier (2011), Hodgkinson et al. (2009), Matzler et al., (2014), and Khatri and Ng (2000), who recommend and support the use of intuition in organizational decision making. Our findings on intuition support the idea that "the intuitive model has more risks attached to it, can be disconcerting, paradoxical and ambiguous, and in many routine situations is probably not needed" (Sadler-Smith & Shefy, 2004, p.25). Previously noted supporters of intuitive decision-making also note that it may be necessary in times of time pressure (Agor, 1986; Miles & Sadler-Smith, 2014; Phillips et al, 2015; Sadler-Smith & Shefy, 2004; Salas et al., 2009), but the current study suggests that even in time pressured situations (i.e., the current study—analyzing eighteen resumes in two and a half hours), intuitive decision-making is poor. Therefore, we do not recommend the use of intuition in selection decisions.

Computer and Human Raters

Computer methods demonstrated the highest *d'* and AUC, and after visual analysis were the ideal decision-makers. Furthermore, this fit for both computer single word and computer synonym methods. Overall, CATA systems outperformed humans, with the exception of the logic method, with CATA systems being approximately equal in the hire two condition. However, it seems that the CATA systems are the best at handling ambiguity in resume content. This is supported by the findings of Campion et al. (2016), Duriau et al. (2007), McKenny et al. (2016), and Rosenberg et al. (1990).

We predicted that when ambiguity was introduced, human raters would perform better than CATA systems. We predicted this because of the human ability to detect idiosyncrasies in text and use these context clues that CATA systems would not be able to detect. Highhouse (2008) refutes the idea that being able to "read between the lines" and the consideration of idiosyncrasies would help human raters. Highhouse (2008) purports that even if being able to detect these idiosyncrasies and reading between the lines was useful, it would provide an inconsequential difference in performance. This is separate from the theories provided by Meehl (1954). Our findings support the discussion by Highhouse (2008).

We found that when ambiguity was introduced, human logic performance decreased. Accepting the assumption that increasing ambiguity caused raters to consider more context clues about an applicant, our results show that the use of context may actually be harmful. This is aligned with Highhouse (2008) who suggests that providing detailed information for a rater to consider may be more distracting than helpful. However, when ambiguity increased, the accuracy of CATA methods did not change. This provides some evidence that intuition is present in any human decision and that intuition is linked with decision error. Additionally, when a purely logical method is provided to CATA systems, accuracy was incredibly high. This indicates that a purely logical system may be superior to any system that includes intuition.

Furthermore, since both computer algorithms performed the same, that gives evidence of no algorithm error occurring in the current study. Algorithm error occurs when two algorithms on the same text results in different outcomes (McKenny et al., 2016). Computer systems appeared to be the perfect raters, having the highest possible *d'* and AUC, and falling into the ideal decision maker part of the visual analysis. This may represent a ceiling effect for the computer algorithms. Future studies should consider when a computer algorithm will be less successful in selecting applicants.

It was noted previously that both the humans and computers were "set up to win"; both were given resumes that perfectly matched the requirements of the job. Humans were even given job descriptions that computers were not given to consider the job that applicants were hired for, further setting them up for success. However, humans still faltered in more ambiguous situations, where CATA systems remained consistent and unbiased. The further addition of extraneous information also supports the idea that Highhouse (2008) suggests, that providing detailed information to a rater may harm performance.

Implications for Practice and Future Directions

Contrary to what we predicted, computer assisted text analysis may be an ideal way to bypass human error in the hiring system. This supports the findings of Campion et al. (2016), Duriau et al. (2007), Rosenberg et al., (1990), and McKenny et al. (2016), and is contrary to the findings of Ein-Dor and Spiegler (1995). Additionally, results provide evidence that a logic system should be used for human raters. However, previous research demonstrates that intuitive methods are frequently used by managers and selection professionals (Agor, 1986; Highouse, 2008; Isenberg, 1984; Lodato et al., 2011). We propose that this is because these methods are a quick and easy approach to answering a question. This is also supported by research that has demonstrated that intuition is most likely to be used in time-pressured situations (Agor, 1986; Miles & Sadler-Smith, 2014; Phillips et al, 2015; Sadler-Smith & Shefy, 2004; Salas et al., 2009). However, another quick and easy approach that does not include the errors of intuition has been demonstrated in the current study: CATA systems. CATA systems provide a quick and easy alternative to making selection decisions that are not riddled with error.

However, it is notable that the current study provides the ideal situation for raters. Specific factor error and random response error were both controlled for. Specific factor error is error that occurs when key words given to computer assisted text analysis systems are inaccurate or misrepresent the scoring protocol (McKenny et al., 2016). Additionally, random response error occurs in the resume content, when resume content fails to match the scoring protocol or algorithm. Random response error may also occur when a resume fails to contain appropriate terms relating to the algorithm or scoring protocol. The terms that were used in the CATA algorithms were validated and well-developed, and resumes were created as perfect scenarios to match the scoring protocol.

It is unlikely that in real life selection practices both of these errors will be controlled for. Even though the “computer only” methods seem to exclude human error, human error can occur from the individual that produces the CATA algorithm as well as from the applicants. Researchers recommend that professionals in the field of selection ensure that algorithms are reliable and valid and are tested frequently for errors. Additionally, the use of CATA systems alone is ill advised. CATA systems still need to be informed of what to search for, and that information has to be accurate, well researched, job related, and likely to show up in a resume. As such, future studies should explore the effects of specific factor error.

Second, if the CATA system were to make an error, it would be extremely difficult to see without another method to compare the results to. We recommend the use of CATA systems and human systems concurrently. Human logic based systems could be used on several randomly selected resumes to check that the CATA system is accurately scoring these resumes. Thus, this must include the resumes that were chosen for hire and those that were not chosen for hire. It would be easy to see why a resume chosen by CATA would fit, but this would only help to detect any false alarms. Viewing and rating the resumes that were not chosen for hire by a CATA system would be the most important to check; if the computer is failing to find information that is present in a resume this should inform when misses are occurring.

Misses are the most concerning part of CATA systems; in most cases, human resource professionals do not even come into contact with the resumes that a CATA system did not choose. If misses are occurring for well-qualified applicants, potentially those that are the best hires for the job, then this could be occurring for reasons related to gender or race. This could lead to adverse impact, even before the professional making the hiring decisions sees the applicants. Though we do not have information on how organizations use these methods, there are two possible ways they are being used: (1) to cut out initially unqualified applicants in a hurdle method (identify low quality applicants), (2) to make final decisions about applicants when comparing (identify high quality applicants). The first situation is less likely to lead to issues of adverse impact. A miss in the “first cut” is much less concerning than a miss when making a final hiring decision. Regardless, adverse impact is a real threat when using CATA methods. For instance, Hiemstra, Derous, Serlie and Born (2013) found that ethnicity predicted differences in grammar in resumes and that layout and grammar were significant predictors of applicant success. Even more concerning is that without the extra step to validate this hiring method concurrently, any detection of adverse impact and the accuracy of the method is completely absent. Future research should explore the effects of gender (is one gender more likely to use context dependent descriptions in resumes than another?), of non-native English-speaking applicants (for those who may not have perfect English but are qualified for a job), and the effects of the development of an algorithm by a non-expert. Additionally, attention should be paid to legal issues surrounding the use of these CATA systems for resume screening. Questions should be raised about when a person becomes a true applicant. Does someone become an applicant when they put their resume into a CATA system? Or when their resume makes it through the CATA system and a hiring manager sees their resume?

Random response error, a failure for resume content to match a human or CATA rating system, cannot be controlled for by any practitioner, but should be considered. If there are articles about “cheating the system” then people are likely artificially inflating their scores using invalid techniques and, thus, misrepresenting themselves (Boulden, 2013; Bradford, 2012; Cappelli, 2012; Giang, 2013). Allowing these applicants to move further in the selection process is undesirable and a waste of time and money. Weinstein (2012) writes that “in today’s ultra-competitive job market, savvy job applicants are maximizing the resume review process by submitting ‘behaviorally-focused’ resumes for the jobs they seek….you might simply be hiring the applicant who knows how to market himself” (p.54). Weinstein (2012) describes a manual scoring method that can be used to identify behaviorally-based items in a resume: (1) collect a list of behaviors required for success in a position, (2) have subject matter experts list the job responsibilities of that job, (3) underline important action verbs in each responsibility, and (4) use these as a reference to review applicant resumes. We recommend this system also be used in order to create computer algorithms. Behaviorally-based resume content should be considered in comparison to non-behaviorally based content. Professionals should consider including both behaviorally based and non-behaviorally based key words into computer algorithms.

Considering the topic of these news articles, it is interesting that applicants find it easier to trick the computer-based system than it is to trick a person (Boulden, 2013; Bradford, 2012; Cappelli, 2012; Giang, 2013). This further brings up questions about ethics; why is it that applicants find it acceptable to misrepresent themselves to a computer system for hire rather than a human/ manual scoring system for hire?

Additionally, random response error could function as a source of both a false alarm and a miss. A miss in response to random response error would occur when an applicant fails to provide the "correct" terms, structure, or description of work experience that an algorithm is searching for but is a true match for a job. On the other hand, random response error that results in a false alarm could be a result of inflation and impression management on the part of the applicant. It is important to consider what measures are being taken in practice to prevent and catch both occurrences; a selection professional is either missing a well-qualified applicant or potentially hiring an applicant that is largely underqualified. It is likely this is exactly what happens when an applicant enters their resume into Monster.com, Linkedin, or Indeed.com; the decision to consider an applicant for hire is clouded by both random response error and specific factor error. Further, a situation where Indeed.com would hire an applicant for a job, but Monster.com would not hire an applicant for the same job implies that algorithm error could also occur, regardless of whether that applicant is qualified. Future studies should explore how the methods that are being used work, what errors they make, and how to measure these errors.

Also, it is notable that Human Resource professionals have demonstrated a preference for intuitive hiring methods (Highhouse, 2008; Isenberg ,1984; Lodato, Highhouse, & Brooks, 2011). The current study also demonstrates this preference for intuitive decisions when given the opportunity to use logic-based information. Participants, even when given a logical system to score a resume, failed to use that information to make a hiring decision. The current study also allows us to conclude that intuitive decisions are not ideal, lack accuracy and validity, and that raters using an intuitive method over-select applicants, demonstrating a liberal bias. We also know that human raters may fail to succeed in increasingly ambiguous situations (i.e. average quality resumes). The results of this study suggest that the use of manual scoring, a logical system, by an HR professional in combination with a CATA system is more accurate and reliable than using an intuitive decision alone. Future studies should explore the results when intuitive methods are explicitly included in logical decision methods.

In lieu of the inaccuracy of intuitive decision-making, we recommend providing methods to make intuitive decisions more accurate. It is clear that intuitive decision making cannot be completely ignored simply because of inaccuracies; the “stubborn reliance” on using intuition in selection processes has been found in the current study and in previous research (Highhouse, 2008; Isenberg, 1984; Lodato et al., 2011). We recommend, first, encouraging selection professionals to consider that the selection process is probability-based; regardless of how certain one is about the accuracy of a hiring decision, there is a chance that applicant may be unqualified (Highhouse, 2008). Highhouse (2008) notes that most people incorrectly believe they have near perfect accuracy in predicting employee success during selection procedures and believe that being able to predict this comes from intuition.

However, Fisher (2008) emphasizes that providing feedback does not guarantee that learning and subsequent behavior change will occur. Fisher emphasizes that when a hit occurs, hindsight bias is likely and when a false alarm occurs, professionals are likely to deflect blame (e.g., claiming the applicant lied). To improve on this, Fisher (2008) recommends (a) providing selection professionals with a model of the selection ratio found from intuitive decisions, (b) keeping score of the outcomes, and (c) seeking out confirming and disconfirming evidence. Keeping score of the outcomes is particularly important, since hiring outcomes (i.e. performance) and the hiring decision are temporally distant. Additionally, as previously mentioned, Sadler-Smith and Shefy (2004) describe that intuitive methods are susceptible to ease of recall, over-confidence, confirmation bias, and hindsight bias. To address these biases Sadler-Smith and Shefy (2004) recommend "playing devil's advocate" by testing intuitive judgments, raising objections to them, eliciting feedback, and generating counter arguments. Though others have provided ways to manage intuition errors, research should explicitly explore how useful these tools are at making intuitive decisions more accurate.

Moreover, in accordance with previous research, the current study suggests that computer-assisted text analysis is a viable alternative to human-only screening (Campion et al., 2016; Duriau et al., 2007;McKenny et al., 2016; Rosenberg et al., 1990). CATA methods are not significantly different from human logic methods when minimal ambiguity is provided and both of these systems demonstrate high accuracy. Further, CATA methods are far distinguished from the poor decisions made through human intuition.

Additionally, raters were “set up for success”, where the resumes were developed directly from the scoring protocol. It is likely that resumes developed independently of the scoring protocol will yield different results, creating random response error (McKenny et al., 2016). Specifically, when synonyms are used in the resume and the single word CATA system is used to analyze those resumes, the single word CATA system is likely to be at a disadvantage. Similarly, using resumes with negative wording (such as no, not, etc.) is likely to produce false alarms. Future studies should explore what happens when both are set up for failure. What happens when random response error occurs in resume content that does not match a scoring protocol? What happens when negative language or synonyms are used? What happens when specific factor error occurs and a person developing the algorithm fails to choose the right words that mimic a well-developed scoring protocol? What if there is not a well-developed scoring protocol for manual scoring? What happens when there is algorithm error and different algorithmic methods do not result in the same outcomes? Future directions for research should include exploration into high and low quality algorithms for computer screening of resumes, as well as high quality but more ambiguous text samples including more random response error. In this study, the highest possible quality algorithm was given to the text-mining program. What would happen in a situation where the computer screening system is flawed?

Moreover, future studies should explore when a human rater may perform better than a CATA system. This may be investigated by testing the effect of the number of resumes to rate or varying types input into CATA systems (full sentences, phrases, essays, lists, etc.). The current study focused on making the human and computer methods comparable to one another to make comparison simple and reduce confounding factors. However, it would be important for future studies to consider when human and computer systems are not comparable, but are still able to be accurate.

Finally, in the current study, a simplified version of a CATA system was used, only focusing on frequency of words. Rosenberg, Schurr, and Oxman (1990) suggest that more context-sensitive texts will require more sophisticated programs and may paint a different picture of the reliability and validity of CATA. Future studies should test the algorithm error associated with simplified and more complete CATA systems in analyzing resumes and subsequently hiring applicants.

Limitations

There were several limitations in this study. First, researchers aimed for a realistic representation of the hiring processes that utilize CATA. However, in the formation of the study there was a lack of access to proprietary information from companies that actually utilize automated resume screening. Though this study was intended to simulate what is truly happening when someone's resume is screened, simulation reality is not guaranteed. Future studies and researchers should collaborate with companies to further understand and better mimic what is truly happening in applied situations.

Additionally, a limitation of this study is the use of undergraduate students with little to no human resource or industrial/ organizational psychology experience. Though Dipboye et al. (1975) supported the use of students to review resumes, and results between expert and novice raters were not critically different in the current study, future studies should explore the use of participants that have skill and practice in screening resumes.

Finally, in the use of the intuition-based decision, as it was a control group, the researchers did not explore the effect of instruction on participants. That is, raters were not instructed to hire a particular number of resumes from an intuitive decision. This unintentionally confounded the intuition-based decision group into seeming more liberal than they may truly be. Future studies should correct this limitation by instructing the intuition-based decision group to pick two resumes from each job type, and four resumes from each job type in a separate condition to determine if the liberal bias was a result of human error or study design.

Conclusion

Overall, there are two main stories to be told here. First, when using intuition, human raters perform poorly and massively over select applicants, demonstrating a liberal bias and the occurrence of many false alarms. Second, computers simulated human conditions, and in some cases out-performed human raters. When humans were given more ambiguous situations (hire four condition), they began to lean towards a more intuitive response, resulting in more false alarms and lower accuracy, though this may not be true for expert raters, and should be explored in future studies.

Similar to previous research on other text and content types, the current study provides initial evidence that CATA used in resume screening makes a valid, reliable alternative to manual scoring of resumes (Campion et al., 2016; Duriau et al., 2007; McKenny et al., 2016; Pennebaker et al., 2003; Rosenberg et al., 1990). In conclusion, computer automated resume screening may be a valuable alternative to human screening of resumes, especially when human bias is likely with more ambiguous resumes. Overall, CATA systems provide an advantage for hiring of resumes, as they remain less biased, faster, cheaper, and easier than manual scoring methods.

References

Abdi, H. (2007). Signal detection theory (SDT). *Encyclopedia of measurement and statistics*, 886-889.

Agor, W. H. (1986). The logic of intuition: how top executives make important decisions. *Organizational Dynamics*, *14*(3), 5-18.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, *136*(4), 569.

Boulden, J. (2013). Software weeds out weak resumes. Retrieved from: http://www.cnn.com/2013/01/08/business/resume-software-scanning/

Behling, O., & Eckel, N. L. (1991). Making sense out of intuition. *The Executive*, *5*(1), 46-54.

Betsch, T., & Glöckner, A. (2010). Intuition in judgment and decision making: extensive thinking without effort. *Psychological Inquiry*, *21*(4), 279-294.

Bradford, B. (2012). Why companies use software to scan resumes. Retrieved from: http://www.npr.org/2012/10/06/162440531/why-companies-use-software-to-scan-resumes

Burke, L. A., & Miller, M. K. (1999). Taking the mystery out of intuitive decision making. *The Academy of Management Executive*, *13*(4), 91-99.

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology, 101*(7), 958-975.

Cappelli, P. (2012). How to get a job? Beat the machines. Retrieved from: http://business.time.com/2012/06/11/how-to-get-a-job-beat-the-machines/

Cole, M. S., Feild, H. S., Giles, W. F., & Harris, S. G. (2009). Recruiters' inferences of applicant personality based on resume screening: do paper people have a personality?. *Journal of Business and Psychology, 24*(1), 5-18.

ComputerPsych LLC (2011). Online detection theory calculator. Retrieved from: www.computerpsych.com/Research_Software/NormDist/Online/Detection_Theory

Crossan, M. M., Lane, H. W., & White, R. E. (1999). An organizational learning framework: From intuition to institution. *Academy of management review*, *24*(3), 522-537.

Dipboye, R. L., Arvey, R. D., & Terpstra, D. E. (1977). Sex and physical attractiveness of raters and applicants as determinants of resume credentials. *Journal of Applied Psychology*, *62*(3), 288–294.

Dipboye, R. L., Fromkin, H. L., & Wiback, K., (1975). Relative importance of applicant sex, attractiveness, and scholastic standing in evaluation of job applicant resumes. *Journal of Applied Psychology, Vol 60*(1), 39-43.

Duriau, V. J., Reger, R. K., & Pfarrer, M. D. (2007). A content analysis of the content analysis literature in organization studies: research themes, data sources, and methodological refinements. *Organizational Research Methods*, *10*(1), 5-34.

Ein-Dor, P., & Spiegler, I. (1995). Natural language access to multiple databases: a model and a prototype. *Journal of Management Information Systems, 12*(1), 171-197.

Epstein, S. (2008). Intuition from the perspective of cognitive-experiential self-theory. *Intuition in Judgment and Decision Making*, 23-37.

Equal Employment Opportunity Commission (n.d.). Pre-employment inquiries and race. Retrieved from: www.eeoc.gov/laws/practices/inquiries_race.cfm

Fisher, C. D. (2008). Why don't they learn?. *Industrial and Organizational Psychology*, *1*(3), 364-366.

Giang, V. (2013). These formatting rules will get your resume through the screening system. Retrieved from: http://www.businessinsider.com/formatting-rules-to-get-your-resume-through-the-scanning-software-2013-2

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451-482.

Hutchinson, K. L. (1984). Personnel administrators' preferences for resume content: a survey and review of empirically based conclusions. *Journal of Business Communication, 21*(4), 5-14.

Glockner, A., & Betsch, T. (2008). Modelling option and strategy choices with connectionist networks: towards an integrative model of automatic and deliberate decision making. *Judgment and Decision Making*, *3*(3), 215–228.

Hammond, K. R., Hamm, R. M., Grassia, J., & Pearson, T. (1987). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. *IEEE transactions on Systems, Man, and Cybernetics*, *17*(5), 753-770.

Hiemstra, A. M., Derous, E., Serlie, A. W., & Born, M. P. (2013). Ethnicity effects in graduates' résumé content. *Applied Psychology*, *62*(3), 427-453.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology*, *1*(3), 333-342.

Hodgkinson, G. P., Sadler-Smith, E., Burke, L. A., Claxton, G., & Sparrow, P. R. (2009). Intuition in organizations: implications for strategic management. *Long Range Planning*, *42*(3), 277-297.

Hogarth, R. M. (2002). Deciding analytically or trusting your intuition? The advantages and disadvantages of analytic and intuitive thought. *The Advantages and Disadvantages of Analytic and Intuitive Thought (October 2002). UPF Economics and Business Working Paper*, (654).

IBM Corp. Released 2012. IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.

Isenberg, D. J. (1984). How senior managers think. Harvard Business Review. November/ December, 81-90.

Kabanoff, B. (1996). Computers can read as well as count: how computer-aided text analysis can benefit organizational research. *Trends in Organizational Behavior, 3*, 1–21.

Kahneman, D. (2011). Thinking, fast and slow. New York, NY: Macmillan.

Khatri, N., & Ng, H. A. (2000). The role of intuition in strategic decision making. *Human relations*, *53*(1), 57-86.

Knouse, S. B. (1989). The role of attribution theory in personnel employment selection: a review of the recent literature*. Journal of General Psychology, 116*(2), 183–196.

Knouse, S.B. (1994). Impressions of the resume: the effects of applicant education, experience, and impression management. *Journal of Business Psychology, 9*(1), 33-45.

Lodato, M. A., Highhouse, S., & Brooks, M. E. (2011). Predicting professional preferences for intuition-based hiring. *Journal of Managerial Psychology, 26*(5), 352-365.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology press.

Matzler, K., Uzelac, B., & Bauer, F. (2014). Intuition: the missing ingredient for good managerial decision-making. *Journal of Business Strategy*, *35*(6), 31-40.

McKenny, A. F., Aguinis, H., Short, J. C., & Anglin, A. H. (2016). What doesn't get measured does exist improving the accuracy of computer-aided text analysis. *Journal of Management*, Advance Online Publication.

Meehl, P. E. (1954). Clinical versus statistical prediction: a theoretical analysis and a review of the evidence. Minneapolis, MN: University of Minnesota.

Miles, A., & Sadler-Smith, E. (2014). "With recruitment I always feel I need to listen to my gut": the role of intuition in employee selection. *Personnel Review*, *43*(4), 606-627.

Mitchell, T. R., & Beach, L. R. (1990). "… Do I love thee? Let me count…" Toward an understanding of intuitive and automatic decision making. *Organizational Behavior and Human Decision Processes*, *47*(1), 1-20.

Oliphant, V. N., & Alexander, E. R. (1982). Reactions to resumes as a function of resume determinateness, applicant characteristics, and sex of raters. *Personnel Psychology, 35*(4), 829–842.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: our words, our selves. *Annual Review of Psychology*, *54*(1), 547-577.

Phillips, W. J., Fletcher, J. M., Marks, A. D., & Hine, D. W. (2015). Thinking styles and decision making: a meta-analysis. *Psychological Bulletin, 142*(3), 260-290.

Rosenberg, S. D., Schnurr, P. P., & Oxman, T. E. (1990). Content analysis: a comparison of manual and computerized systems. *Journal of personality assessment*, *54*(1-2), 298-310.

Sadler-Smith, E., & Shefy, E. (2004). The intuitive executive: understanding and applying 'gut feel'in decision-making. *The Academy of Management Executive*, *18*(4), 76-91.

Salas, E., Rosen, M. A., & DiazGranados, D. (2009). Expertise-based intuition and decision making in organizations. *Journal of Management*, 1-31.

Simon, H. A. (1987). Making management decisions: the role of intuition and emotion. *The Academy of Management Executive 1*(1), 57-64.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137-149.

Stillman, J. A., & Jackson, D. J. (2005). A detection theory approach to the evaluation of assessors in assessment centres. *Journal of occupational and organizational psychology*, *78*(4), 581-594.

Swets, J. A., Tanner Jr, W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68*(5), 301.

TheRMUoHP Biostatistics Resource Channel. (2013, May 14). *How to use SPSS- receiver operating characteristics (ROC) curve part 1*. [Video file]. Retrieved from: https://www.youtube.com/watch?v=_2zN2a3MgmU.

Vanlehn, K., & Van de Sande, B. (2009). Acquiring conceptual expertise from modeling: the case of elementary physics. *The Development of Professional Performance: Toward Measurement of Expert Performance and Design of Optimal Learning Environments*, 356-378.

Wang, Y., Highhouse, S., Lake, C. J., Petersen, N. L., & Rada, T. B. (2015). Meta-analytic investigations of the relation between intuition and analysis. *Journal of Behavioral Decision Making*.

Weber, R. P. (1990). *Basic content analysis* (No. 49). Sage.

Weinstein, D. (2012). The psychology of behaviorally-focused résumés on applicant selection: Are your hiring managers really hiring the 'right'people for the 'right'jobs?. *Business Horizons*, *55*(1), 53-63.