AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Applied Psychology

© 2023 American Psychological Association
ISSN: 0021-9010

https://doi.org/10.1037/apl0001144

Using Natural Language Processing to Increase Prediction and Reduce Subgroup Differences in Personnel Selection Decisions

Emily D. Campion[1], Michael A. Campion[2], James Johnson[3], Thomas R. Carretta[3], Sophie Romay[3], Bobbie Dirr[3], Andrew Deregla[3], and Amanda Mouton[3]
[1] Department of Management and Entrepreneurship, University of Iowa
[2] Department of Organizational Behavior and Human Resources, Purdue University
[3] United States Air Force

The purpose of this research is to demonstrate how using natural language processing (NLP) on narrative application data can improve prediction and reduce racial subgroup differences in scores used for selection decisions compared to mental ability test scores and numeric application data. We posit there is uncaptured and job-related constructs that can be gleaned from applicant text data using NLP. We test our hypotheses in an operational context across four samples (total $N = 1{,}828$) to predict selection into Officer Training School in the U.S. Air Force. Boards of three senior officers make selection decisions using a highly structured rating process based on mental ability tests, numeric application information (e.g., number of past jobs, college grades), and narrative application information (e.g., past job duties, achievements, interests, statements of objectives). Results showed that NLP scores of the narrative application generally (a) predict Board scores when combined with test scores and numeric application information at a level of correlation equivalent to the correlation between human raters (.60), (b) add incremental prediction of Board scores beyond mental ability tests and numeric application information, and (c) reduce subgroup differences between racial minorities and nonracial minorities in Board scores compared to mental ability tests and numeric application information. Moreover, NLP scores predict (a) job (training) performance, (b) job (training) performance beyond mental ability tests and numeric application information, and (c) even job (training) performance beyond Board scores. Scoring of narrative application data using NLP shows promise in addressing the validity-adverse impact dilemma in selection.

Keywords: natural language processing, text analysis, artificial intelligence, personnel selection, subgroup differences

Supplemental materials: https://doi.org/10.1037/apl0001144.supp

An enduring challenge to scholars and practitioners is not only to increase prediction, but to do so in a manner that reduces adverse impact in selection (Outtz, 2010). Adverse impact in hiring where racial minorities have lower passing rates compared to nonracial minorities presents initial evidence (prima facie case) of discrimination in the United States based on Title VII of the Civil Rights Act of 1964 (Uniform Guidelines on Employee Selection Procedures, 1978, Section 3A). The main response is typically to demonstrate the procedure is job-related. Yet, even if job-relatedness can be exhibited, there is the additional burden of showing that alternative procedures with equal validity and less adverse impact are not available. Historically, the most valid predictors—those that are most strongly related to job performance—tend to yield the largest subgroup differences, thus creating the validity-adverse impact dilemma (Ployhart & Holtz, 2008). There have been significant advancements in personnel selection research aimed at resolving this trade-off. While researchers have achieved incremental success, the dilemma remains, and we continue to seek solutions to address this challenge facing personnel selection in societies as diverse as the United States.

Progress is limited by at least two hurdles: one conceptual and one methodological. First, there is a notable amount of narrative job-related candidate information available in the selection context that is typically only evaluated qualitatively by hiring officials and not explicitly and systematically scored to inform selection decisions. Examples include work and education history, letters of reference, statements of interest, accomplishments, narrative responses, awards, and participation in clubs (Brown & Campion, 1994). Second, the strategies that could potentially demonstrate higher validity and smaller subgroup differences are generally time-consuming and require considerable

Emily D. Campion https://orcid.org/0000-0003-1555-2089

The views expressed are those of James Johnson, Thomas R. Carretta, Sophie Romay, Bobbie Dirr, Andrew Deregla, and Amanda Mouton and do not necessarily reflect the views of the U.S. Government, the U.S. Department of Defense, or the U.S. Air Force.

The authors have no known conflict of interest to disclose. Previous versions of this research were presented at the 36th and 37th annual conferences of the Society for Industrial and Organizational Psychology. A portion of this research was conducted while Emily D. Campion was at Old Dominion University.

Correspondence concerning this article should be addressed to Emily D. Campion, Department of Management and Entrepreneurship, University of Iowa, W224 Pappajohn Business Building, Tippie College of Business, Iowa City, IA 52242, United States. Email: emily-campion@uiowa.edu

cost for human assessors (e.g., structured interviews, assessment centers; Ployhart & Holtz, 2008). What is needed, then, is a solution that offers an incremental improvement in the validity-adverse impact trade-off and also reduces costs. One emerging approach that has offered promising evidence is the use of natural language processing (NLP) to score narrative application information (e.g., M. C. Campion et al., 2016), although significant questions remain regarding its utility in addressing this tradeoff. Therefore, we ask: Can using NLP on narrative application data reduce sole reliance on testing and traditional employment assessments by increasing validity while reducing subgroup differences?

Arthur and Villado (2008) identified the distinction between construct and method when comparing alternative predictors in selection to reduce subgroup differences in scores. They argue failure to consider this difference makes the results of such comparisons uninterpretable. We maintain this distinction by focusing our theorizing on comparing constructs measured via NLP to general mental ability tests—the primary and historically most valid predictor of performance—and argue that these constructs offer incremental validity. Tesluk and Jacobs (1998) refer to such narrative information as “qualitative aspects of work experience” (p. 322). Rooting our research in the work and life experiences literature, we contend that expanding the content of what is measured by including this information might improve prediction and possibly reduce subgroup differences if the constructs are job-related and noncognitive in nature. However, the scoring method by which we measure these constructs is also a significant part of our contribution because NLP scores narrative data more efficiently than hiring officials, and text are data not currently systematically scored and likely to be rich with job-related information. Therefore, while our key comparison is among constructs, our contributions are conceptual and methodological. We position NLP as an emerging scoring method that offers the distinct advantages of (a) measuring constructs from text application data more efficiently than humans, and (b) including text data likely affords a broader sampling of noncognitive constructs that are currently uncaptured, theoretically relevant to job performance, and may demonstrate smaller subgroup differences than typical constructs measured via traditional employment tests with high validity (e.g., cognitive ability). The purpose of this work is not to examine any potential bias or discrimination in NLP as a method, but rather advance NLP as a method that enables researchers to measure additional data that may increase prediction and reduce subgroup differences due to a broadening of the predictor space. Further, we examine subgroup differences because they are the basis of a prima facie case and not as an indicator of bias due to measurement contamination.

Our intended contribution is threefold. First, we show that through NLP, we can measure job-related content that has incremental validity above and beyond traditional selection constructs (e.g., mental ability). Research on work and life experiences suggests information from a candidate’s entire application (e.g., resume, affiliations and achievements, statements of objectives) may include behaviors and accomplishments in and outside of work that are related to job performance (Hough, 1984; Mumford & Stokes, 1992; Quińones et al., 1995; Tesluk & Jacobs, 1998). Moreover, we demonstrate that the inclusion of this information reduces subgroup differences in a composite score based on the idea that greater content coverage in predictors reduces differences among subgroups, depending on the relative subgroup differences and intercorrelations (Sackett & Ellingson, 1997). We compare the variables captured through NLP to mental ability to assess this reduction in subgroup differences through the construct-change approach (Arthur & Villado, 2008; Arthur et al., 2013, 2021).

Second, we present NLP as a scoring method that can efficiently score a range of sources of job-related content not measured in most contemporary employment tests and can do so as accurately as the current common method (i.e., human ratings). Moreover, we suggest that this allows us to capture a broader range of job-related constructs. These constructs include knowledge, skills, abilities, and other characteristics such as social skills, personality, leadership, and interests, many of which may be less cognitively oriented than traditional mental ability employment tests. Candidates submit a large amount of text data including professional achievements, descriptions of their previous job duties, and their objectives (e.g., goal statements). Scoring text has historically required extensive human resources, which leaves relevant candidate information out of hiring decisions. Moreover, research on work and life experiences has lamented the difficulty in systematically scoring these qualitative elements (Tesluk & Jacobs, 1998). NLP offers a solution to these challenges by quickly scoring text data, increasing the amount of job-related information available to hiring managers, and improving decisions. Text models can be used in place of a human rater and to save resources (e.g., M. C. Campion et al., 2016). Our work potentially represents a lower bound on what is possible in this domain as new developments emerge in NLP.

Finally, we test our hypotheses in an operational context across four samples (total N = 1,828) of professional employees. We examine NLP scores independently and compare their prediction of human ratings (hiring Board scores) and initial job (training program) performance to mental ability test scores and numeric application information. We demonstrate how text scores can improve prediction and reduce subgroup differences in actual selection decisions. We also offer illustrations of how NLP scores used in concert with mental ability tests and numeric application information can reduce adverse impact ratios of hypothetical selection decisions.

A Brief Overview of Natural Language Processing

NLP broadly refers to “a set of methods for making human language accessible to computers” (Eisenstein, 2019, p. 1). It exists at the intersection of artificial intelligence and linguistics, and it relates closely to traditional text analysis methods in organizational psychology and management (e.g., content analysis, Hsieh & Shannon, 2005; grounded theory; Corbin & Strauss, 1990). NLP typically relies on closed-dictionary approaches (e.g., Linguistic Inquiry Word Count; Pennebaker et al., 2015) and open-dictionary approaches (e.g., latent Dirichlet allocation [LDA]; Blei et al., 2003).[1] NLP utilizes machine learning algorithms in a data-driven approach to summarize themes and other dimensions of text (corpus). It involves reducing words or phrases to logical categories (variables) to represent ideas (generate meaning) from the text. The scoring of text data is a relatively new practice in human resources

[1] LDA remains one of the most popular open-dictionary approaches; however, more advanced approaches have emerged including neural networks and transformer networks such as Bidirectional Encoder Representations from Transformers (Devlin et al., 2019) and RoBerta (Liu et al., 2019).

(HR), but organizational scientists and those in neighboring disciplines have utilized NLP to quicken the processing of large amounts of text data for some time. For example, education researchers were early adopters who used it to score hundreds of thousands (now millions) of essays annually, facilitate teaching English writing skills to elementary students (Dikli, 2006; Valenti et al., 2003), and now to score the writing skills of applicants domestically and globally for admission into U.S. universities (Anderson, 2018).

Research using NLP in organizational science appeared in the mid-1990s (Morris, 1994; Wolfe et al., 1993). Organizational scholars have used relatively rudimentary text analysis approaches—such as hand coding—to manage and code text data, or moderately complex versions—such as closed-dictionary approaches—to accelerate content analyses (e.g., Short et al., 2018). These efforts have demonstrated how text analysis can measure a range of constructs including communal and agentic attributions (Madera et al., 2009), leadership rhetoric (Bligh et al., 2004), and emotion (Walker et al., 2017). Now, as these advanced methods become more accessible (Eichstaedt et al., 2021; Hannigan et al., 2019), there is increasing interest in the use of NLP by HR researchers, such as for selection (e.g., M. C. Campion et al., 2016; Sajjadiani et al., 2019; see E. D. Campion & Campion, 2020, for a review of text analysis in employment research), performance management (e.g., Speer, 2018, 2021), and recruitment (Banks et al., 2019).

We propose NLP as a scoring method that is widely applicable and can be tailored to organizational needs. In the current context, we score work and life experience qualitative data that are generic (e.g., past jobs and achievements), but the scoring algorithm is adjusted to the organization. NLP might also be used to score other types of prompts, such as candidate essays in response to situations that occur on the job or job knowledge questions. Thus, NLP is a broadly applicable method that can be used to improve selection in many contexts.

Theoretical Background

Work and Life Experiences Literatures

Research on work and life experiences offers several insights pertinent to the present study. First, candidates likely have relevant work experience that current selection methods may fail to capture, and these experiences are related to work outcomes. For example, Dragoni et al. (2011) found the accumulation of leadership experiences over time related to strategic thinking competency. Second, an obstacle in the work experiences literature is measurement. Quińones et al. (1995) and Tesluk and Jacobs (1998) explain that work experience is difficult to measure due in part to its contextualization (also see Van Iddekinge et al., 2019). They called for researchers to go beyond proxy measures (e.g., seniority and tenure) to better capture relevant work experience. Quińones et al. (1995) suggested distinguishing between amount (e.g., number of jobs), time (e.g., years), and type of experience (e.g., tasks performed). Tesluk and Jacobs (1998) similarly distinguished between quantitative (amount or time) versus type and quality. NLP is particularly useful for the latter such that it measures the types and qualities of the experiences that research suggests is predictive. For example, Howard (1986) illustrated that college major and extracurricular activities predicted variance in managerial promotions. Those with backgrounds in humanities and social sciences demonstrated more effective interpersonal and other managerial skills compared to those in engineering and mathematics.

Third, research on life experiences identifies the importance of other information that might emerge in the application process such as interests, memberships groups, and personal achievements. While much of this research is based on the axiom that past behavior is a primary predictor of future behavior, Mumford and Stokes (1992) suggest a qualifier:

> [T]his statement implies that prior learning and heredity, along with the environmental circumstances in which they express themselves, make some forms of behavior more likely than others in new situations. As a result, assessment of earlier behaviors and experiences permits some accuracy in predicting future behaviors and experiences given a knowledge of environmental demands. (p. 64)

As such, patterns of successful past life experiences are likely to predict successful patterns in the future partly because they reflect adaptability to life circumstances. The solicitation of descriptions of such experiences allows for the assessment of behaviors within their original context using NLP, and correlations with job-related criteria would support their transfer to the workplace. Studies have demonstrated the value of nonwork experiences as they relate to work. For example, Ruderman et al. (2002) found that women who occupied multiple nonwork roles brought skills generated through those experiences into work, which translated into task-related and interpersonal skills. NLP can score life experiences to inform selection decisions, and it is not limited to biographical data that can easily be quantified (e.g., amounts, times) as is characteristic of research that does not use NLP. Instead, it can measure the actual words and thus leverage the rich descriptions in narrative data.

Finally, other methods of assessing candidate backgrounds bear on the meaningfulness of narrative data. For example, meta-analytic evidence shows that personality, social skills, judgment, job knowledge, and mental ability are frequently captured in employment interviews (Huffcutt et al., 2001; Salgado & Moscoso, 2002). Research on accomplishment records similarly illustrates that information on what candidates achieved in past jobs may be more relevant than numeric data such as years of experience (Hough, 1984). Moreover, qualitative components may reflect the more "elusive" aspects of work experience, such as task difficulty and job complexity (Tesluk & Jacobs, 1998, p. 323), which better assess the "density" or developmental intensity of the experiences (Tesluk & Jacobs, 1998, p. 329).

Literature on Reducing Subgroup Differences and Adverse Impact in Selection

It is important to distinguish between subgroup differences and adverse impact. *Subgroup differences* refer to, "psychological, scientific phenomena that are represented or conceptualized as standardized mean differences between groups on measures of psychological constructs" (Arthur et al., 2013, p. 475). On the other hand, *adverse impact* refers to differences in hiring or passing rates between subgroups, normally comparing the racial minority or female subgroup to the racial majority or male subgroup. While subgroup differences may occur in the psychological constructs measured, they do not necessarily result in adverse impact in selection decisions. As such, subgroup difference reduction is more likely through test design, and adverse impact reduction is through assessment administration (Arthur et al., 2013). In the present study, we focus on subgroup difference reduction, but illustrate the potential influence on adverse impact.

To appropriately theorize why using NLP to capture unmeasured constructs may reduce subgroup differences, we need to revisit the distinction between the construct-change and the method-change approach (Arthur & Villado, 2008). Researchers have examined whether the reason for reduction is due to the construct or method. There is some support for the method-change approach. For example, Chan and Schmitt (1997) found that video test administration reduced subgroup differences likely because it reduced the verbal ability requirements. Arthur and colleagues explored constructed responses such as write-ins rather than multiple-choice (e.g., Arthur et al., 2002; Edwards & Arthur, 2007) to measure mathematics and science reasoning and found smaller subgroup differences but equivalent validity in predicting grades. However, in their review of challenges associated with reducing adverse impact, Arthur et al. (2013) observed subgroup difference reduction is virtually always accompanied by a change in constructs and the constructs are typically noncognitive. Perhaps the best-known example historically is that mental ability tests have consistently shown racial subgroup differences (Roth et al., 2017), while personality and other noncognitive psychological tests have not (Hough et al., 2001) because they measure different constructs, even though both typically are assessed via structured responses (e.g., multiple-choice, rating scales, etc.).

In this study, our key argument is that adopting NLP allows us to capture constructs unmeasured in the current selection system and that measuring additional constructs likely improves validity and reduces subgroup differences. This is in line with the construct-change approach. However, it is important to recognize that advanced text analytics is an alternative method of scoring, not itself a measure of a new construct. NLP simply affords the distinct advantage of efficiently scoring information that is not typically quantified systematically. Whether that will increase prediction or reduce subgroup differences will depend on the constructs scored. We argue narrative information in applications and resumes is job-related, and thus may improve prediction (Lievens et al., 2019). The influence on subgroup differences, however, may depend on whether the constructs in the narrative data are noncognitive in nature. This is likely to be the case given the smaller subgroup differences observed for biodata questionnaires and interviews, both of which commonly measure work and life experiences, compared to mental ability tests (e.g., Bobko et al., 1999; Sackett et al., 2022). Unlike employment tests that measure verbal and math ability, problem solving, reasoning, job knowledge, and other indicators of general mental ability, narrative application information includes data on past jobs, education, activities, and accomplishments that may share smaller relationships with mental ability and therefore be less likely to show subgroup differences (Edwards & Arthur, 2007).

The decades of scholarship on alternative selection procedures reveals four challenges that NLP may help address. First, as noted, selection methods with the highest validity are employment tests that measure various mental abilities, but demonstrate the greatest subgroup differences (e.g., Roth et al., 2001; Sackett et al., 2022). NLP scores a wide range of narrative information that is not captured by mental abilities tests (e.g., information in applications), and the constructs are less cognitive and thus likely to have smaller subgroup differences. Second, some of the most popular alternatives with small subgroup differences also have the lowest validity, such as personality tests. This is in part because they are susceptible to response distortion (faking) and also the inaccuracies of self-assessments (e.g., Morgeson et al., 2007). NLP may be more resistant to these concerns because candidates do not score themselves (e.g., give high self-ratings on conscientiousness). Instead, they provide qualitative information (e.g., past job tasks or accomplishments) that is scored via NLP and therefore they do not know specifically what is scored and are less able to distort their responses or self-evaluations. Third, methods that tend to capture work experiences and education that have fairly high validity and smaller subgroup differences than mental ability—like structured interviews and accomplishment records—require more resources to score, which makes them more burdensome to deploy in high-volume hiring contests, and thus not administratively feasible. NLP provides an automated way to score cost effectively.

A fourth, but less recognized, challenge is that adding procedures with smaller subgroup differences to an assessment battery may not decrease adverse impact, but can actually increase it (e.g., Potosky et al., 2005). Sackett and Ellingson (1997) showed that the influence of adding selection procedures to a composite depends on the size of the subgroup differences in the procedures and on their intercorrelation. The reduction in the composite is greater than the sum of the subgroup differences to the extent there is a higher intercorrelation. Perhaps surprisingly, the resulting subgroup difference is virtually never as low as the average subgroup difference and can be greater than the difference on the procedure with the highest difference. The latter means that adding procedures to reduce differences frequently increases differences. This has often frustrated efforts to reduce adverse impact because it seems commonsensical to laypersons (including attorneys and judges) that adding selection procedures with smaller subgroup differences should reduce impact. Whether including NLP scores to a candidate's overall score will reduce or increase subgroup differences is an empirical question at the core of this study.

Hypothesis Development

The purpose of this study is to test whether using NLP on narrative application information is a viable scoring method that adds incremental prediction and reduces subgroup differences compared to mental ability tests and numeric application information. Previous research demonstrated the feasibility of using NLP in a selection context (M. C. Campion et al., 2016), but did not address this question. The first hurdle is whether NLP, along with the test scores and numeric data, can provide a measure of the total application that adequately predicts human ratings. In the present study, we determine the adequacy based on the correlation with human ratings in the form of Board scores, which consist of ratings by a panel of three hiring officials. This is critical because the ability to replace a human rater is one potential benefit. Researchers of automated essay scoring in education have similarly adopted the goal of achieving a correlation comparable to that between human graders (e.g., Attali et al., 2013; Ramineni & Williamson, 2013).

What level of correlation to expect is uncertain due to the lack of clear benchmarks. We adopt the same goal as M. C. Campion et al. (2016) to achieve a correlation equivalent to the level of interrater reliability, which averaged .61 (M. C. Campion et al., 2016, p. 969). We use this goal for two reasons. First, there are great similarities between their study and the present study: both predicted scores of Boards comprised of three hiring officials per applicant, the Boards were highly trained, the selection process was highly structured, it occurred in governmental organizations, and it assessed the use of NLP in selection. Second, this threshold is similar to other highly structured contexts that used multiple raters to make HR decisions. For example, Conway et al. (1995) found .56 and .67 for medium and highly structured interviews. As such:

> *Hypothesis 1:* Computer scores of candidate application information will correlate as highly with human ratings (Board scores) as a level of correlation between human raters of .60.

There are conceptual and empirical reasons to expect that NLP scores will have incremental validity beyond mental ability tests. Conceptually, NLP scores new information and likely new constructs that will increase coverage of the job's content domain compared to mental ability tests, which commonly measure verbal and math skills, job knowledge, and similar attributes. Typical narrative application data include work history (e.g., job titles, duties), education (e.g., degrees, majors), past achievements, and other information (e.g., statements of objectives, activities, memberships). Past research on work and life experiences suggests that narrative application information will incrementally predict job performance. For example, application information likely reflects job-related knowledge, skills, and abilities, as well as motivation and work-related values and attitudes (e.g., Brown & Campion, 1994; Sajjadiani et al., 2019; Tesluk & Jacobs, 1998). Although research using NLP has primarily measured constructs in the domains of organizational behavior and strategic management, many are likely to be present in candidate applications, relevant to HR employment decisions, and nonoverlapping with mental ability tests (E. D. Campion & Campion, 2020).

Empirically, the psychometric formula for estimating the validity of a composite leads to the expectation that including additional valid procedures will increase validity (Ghiselli, 1964, p. 310). Past research demonstrates the estimated incremental validity of various selection procedures beyond mental ability employment tests. These include many that are highly similar to constructs likely to be measured using NLP with applicant information on work and life experiences (e.g., education, personality, interests, and reference checks; Schmidt & Hunter, 1998). Focusing first on the prediction of organizational hiring ratings, we hypothesize:

> *Hypothesis 2a:* NLP scores of narrative application information will have incremental validity in the prediction of hiring ratings (Board scores) beyond mental ability tests.

Application information includes quantitative (e.g., number of jobs, years of education, past grades) and qualitative information (e.g., past jobs, duties, achievements, interests), which are conceptually and empirically distinct (Quińones et al., 1995; Tesluk & Jacobs, 1998). Quantitative information is relatively straightforward to incorporate into selection scores as it is already quantified, yet it fails to capture critical features of previous experiences that are likely to be job-related as argued above. Because we theorize work and life experiences reflect job-related constructs that are unmeasured in current selection methods, we propose that scoring qualitative application information using NLP will add incremental prediction beyond the commonly used quantitative (numeric) application information. Therefore, we hypothesize:

> *Hypothesis 2b and c:* NLP scores of narrative application information will have incremental validity in the prediction of hiring ratings (Board scores) beyond (b) numeric application information and (c) mental ability tests and numeric application information.

Predicting hiring ratings made by humans (Board scores) is a valuable first step in validating the use of NLP for hiring, but it is also important to show that scores derived using NLP predicts job performance. Predicting human ratings is a type of construct validation evidence because the use of NLP is meant to be another measure of the human ratings, but predicting job performance is a type of criterion-related validation evidence because job performance is typically the ultimate criterion in personnel selection. In the current context, we use training performance as our measure of job performance. As such, we hypothesize:

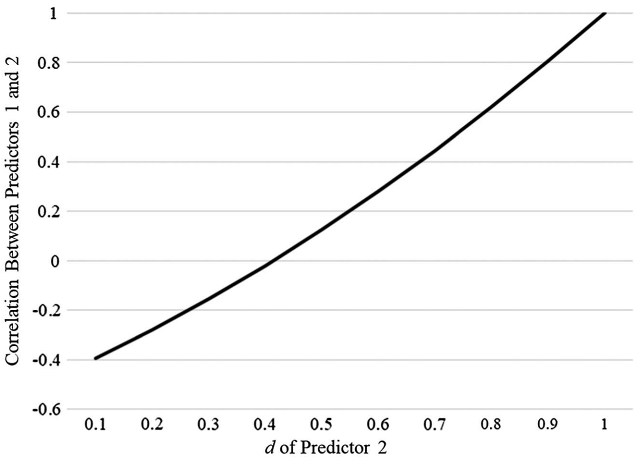
> *Hypothesis 3a–c:* NLP scores of narrative application information will have incremental validity in the prediction of training performance beyond (a) mental ability tests, (b) numeric application information, and (c) mental ability tests and numeric application information.

Although hiring ratings (Board scores) reflect the organization's overall evaluation of all candidate credentials, including mental ability test scores that have been extensively validated, numeric information, and narrative information, NLP scores may still provide incremental prediction of training performance beyond Board scores for a range of reasons, such as better measurement of work and life experiences or allowing that information to directly predict performance rather than being filtered through Board scores. Thus, we hypothesize:

> *Hypothesis 3d:* NLP scores of narrative application information will have incremental validity in the prediction of training performance beyond Board scores.

Building on these hypotheses, we argue that adding NLP scores of narrative application data to mental ability tests and other numeric data (number of jobs, years of education) will reduce subgroup differences largely due to the greater content coverage in the predictor. Yet, as noted previously, the influence of additional predictors and the resulting subgroup differences of a composite is not as straightforward as it may seem because it depends on the size of the subgroup differences of the predictors and the intercorrelations. Using Sackett and Ellingson's (1997) formula for estimating the subgroup difference (p. 713, Formula 2), we illustrate this phenomenon to support our hypothesis. Assuming an average subgroup difference of one standard deviation (*d* of 1) on mental ability tests based on meta-analytic summaries in the literature for Black–White differences (e.g., Hough et al., 2001)—which is usually the largest difference and of most concern—we estimated the value of the differences in NLP scores and the intercorrelation necessary to reduce the composite difference below 1.0. As Figure 1 shows, if NLP scores correlate .30 with the mental ability test, the subgroup difference on NLP scores must be less than about .60

Figure 1
Correlation Between Predictors 1 and 2 Versus d of Predictor 2 Necessary to Reduce d of Composite Below 1.0 When d on Predictor 1 Is 1.0



to reduce the composite difference to below 1.0. However, if the correlation between the NLP scores and test scores is zero, the subgroup difference on NLP scores must be less than .40 to reduce the composite difference to below 1.0. Put differently, if the predictors are highly correlated, the smaller subgroup differences in NLP scores pulls the subgroup differences in the mental ability scores down so that subgroup differences in the combined score is reduced. If the predictors are not highly correlated, then the subgroup differences of NLP scores need to be much lower because any new subgroup differences in NLP scores will add to the existing subgroup differences in the mental ability scores.

There are few empirical estimates of either value in the literature. M. C. Campion et al. (2016) is the only study we could find examining this issue directly. They observed correlations between an English test and NLP scores of accomplishment records of .13 to .18 and near zero with NLP scores of other applicant information (−.04 to .07). The correlations between a professional knowledge test and NLP scores of accomplishment records were .20 to .29 and the correlations with the other NLP scores were again near zero (−.02 to .10). Assuming relatively low correlations between the NLP scores and mental ability test scores, the subgroup differences of the NLP scores need to also be small to yield smaller subgroup differences in the composite, according to Sackett and Ellingson's (1997) formula. M. C. Campion et al. (2016) found that the subgroup differences on the NLP scores were essentially zero (−.03 to .06), thus the subgroup differences of the combined scores should be reduced. Based on our theorizing, these calculations, and the limited prior research, we hypothesize:

> *Hypothesis 4a:* Combining NLP scores of candidate text information with mental ability tests will have smaller subgroup differences than mental ability tests alone.

The formula can be applied similarly to predict the subgroup differences of combining NLP scores with numeric application information. While there is also limited research combining NLP with numeric information, research on biodata may be helpful. For example, in their meta-analysis of subgroup differences of biodata, which usually includes numeric application data, Bobko et al. (1999) found a *d* between Blacks and Whites of .33. The correlation between numeric information with NLP scores of the narrative information is likely to be positive because more years of experience or education will likely yield higher NLP scores because there is more information to score. Using the Sackett and Ellingson (1997) equation and assuming a *d* of .33 on the numeric application variables and a zero-subgroup difference on NLP scores, the *d* on the composite will be lower than .33 at any correlation (Figure 1). Thus:

> *Hypothesis 4b:* Combining NLP scores of candidate text information with numeric application information will have smaller subgroup differences than on numeric application information alone.

> *Hypothesis 4c:* Combining NLP scores of candidate text information with mental ability tests and the numeric application information will have smaller subgroup differences than on mental ability tests and numeric application information.

Method

Setting and Sample

The setting for this study was the operational process of selecting candidates into Air Force Officer Training School (OTS) in 2019. Each year thousands of candidates apply. Boards of officers review and score the applications in a resource-intensive and time-consuming process. The selection procedures examined herein include mental ability employment aptitude tests and a wide range of numeric and narrative application information such as past jobs, degrees, statements, and letters of reference. Thus, the setting allowed a direct comparison between mental ability tests, numeric application data, and text analyzed data in the prediction of selection decisions and job performance in an operational setting.

The setting afforded replication in four samples, consisting of two "Boards" (flying and nonflying jobs) with two panels each: (a) enlisted candidates for flying jobs with Air Force experience, (b) civilian candidates for flying jobs without Air Force experience, (c) enlisted candidates for nonflying jobs with Air Force experience, and (d) civilian candidates for nonflying jobs without Air Force experience. Table 1 shows sample sizes and race composition. The samples range from 210 to 1,057 with a total of 1,828 candidates (including 378 who applied to both flying and nonflying jobs). The total sample is 0.55% American Indian/Alaskan Native, 6.18% Asian, 7.71% Black, 14.17% Hispanic, 1.04% Native Hawaiian/Pacific Islander, and 69.09% White. About 18% are women and the average age is 28.55 ($SD = 3.87$, range 20–40).

To increase power, we combined the two panels of flying jobs and the two panels of nonflying jobs. For Hypotheses 1 and 2, statistical power was 80% to detect an incremental R^2 as low as .04 for the smallest combined sample ($N = 464$). For Hypothesis 3, statistical power was 80% to detect an incremental R^2 as low as .06 for the smallest combined sample. For Hypothesis 4, power was at least 80% to detect a change in *d* down to .57 for the smallest combined sample, and .42 for the average sample size. However, the critical values were .38 and .28 for significance, respectively, which should be small enough to detect meaningful differences. We used $p < .05$

Table 1
Descriptive Statistics of Samples by Race/Ethnicity

Samples	N[a]	Race/ethnicity						
		American Indian/ Alaskan Native	Asian	Black/African American	Hispanic	Native Hawaiian/Pacific Islander	Mixed race	White (nonracial-minority)
Flying with AF experience	287[b]	1	12	11	44	3	0	215
Flying without AF experience	210[c]	3	14	10	23	1	0	150
Flying total	497	4	26	21	67	4	0	365
Nonflying with AF experience	1,057[d]	4	56	103	162	13	1	715
Nonflying without AF experience	274[e]	2	31	17	30	2	0	183
Nonflying total	1,331	6	87	120	192	15	1	898
Total	1,828	10	113	141	259	19	1	1,263
Percentage of total		0.55%	6.18%	7.71%	14.17%	1.04%	0.05%	69.09%

Note. About 2.28% of the total sample is missing demographic data. AF = Air Force.
[a] 378 candidates applied to both flying and nonflying boards and are counted twice in this table. [b] One person missing race data in flying with AF experience sample. Total race data available = 286. [c] Nine people missing race data in flying without AF experience sample. Total race data available = 201. [d] Three people missing race data in nonflying with AF experience. One person indicated “Multiracial.” Total race data available = 1,054. [e] Nine people missing race data in nonflying without AF experience sample. Total race data available = 265.

(1-tailed) for all estimates. All hypothesis tests by board and panel are reported in the online Supplemental Tables A1–A8.

The setting was ideal for this research for many reasons. First, it was an operational hiring process with well-developed mental ability tests that have a long history of importance to staffing. Second, there were extensive narrative data that included the full range of application information. Third, there were four samples (combined into two for main analyses) allowing for replications and adequate power. Fourth, criteria included selection decisions based on a systematic and thorough evaluation process and performance based on extensive posthire training performance metrics. Finally, the goal of the Air Force was to determine whether NLP could be incorporated into their hiring process. Thus, the purpose of the study was important to the organization, which encouraged detailed, rigorous, careful, and peer-reviewed methods.

Measures

Selection Decision Process

Selection decisions into OTS are made by Boards consisting of three Colonels from the career field (flying and nonflying jobs) per 250–300 applicants, with four separate boards as described above yielding the criterion *Board scores*. Applicants applying for both flying and nonflying positions are rated independently twice, and flying positions are selected before nonflying positions. Selection is based on a “whole person” concept, meaning all aspects of the candidates’ credentials are considered. There are no specific cutoff scores, weights, or essential minimum requirements. However, the scoring process is highly structured in terms of the criteria and rating process. There are three scoring areas:

1. Experience/Leadership (2–4 points) is based on leadership potential, letters of reference, work ethic, employment history, military experience and performance, and demonstrated leadership in the form of scope of responsibility, honors and recognition, community service and activities or base involvement, and athletics, skills, and hobbies.

2. Education/Aptitude (2–3 points) is based on academic discipline (nonspecific), grade point average, academic awards and recognition, and Air Force Officer Qualifying Test (AFOQT; Drasgow et al., 2010) scores (with a reference guide as to interpretation of scores).

3. Potential/Adaptability (2–3 points) is based on an interview evaluation conducted by one interviewing officer (which can be any current officer), letters of reference indicating potential and adaptability, personal experiences, communication skills, and law violations.

In total, each Board member can assign from 6 to 10 points, with 0.10 increments, using the following scale: (10) absolutely superior, (9.5) outstanding, (9) few could be better, (8.5) strong, (8) slightly higher than average, (7.5) average, (7) slightly below average, (6.5) well below average, and (6) lowest potential. Each Board member independently rates each candidate, but they must discuss any differences of 1.5 or greater. When there are multiple sub-Boards due to more than 250–300 candidates, the scores are normalized. Only the total summed score is recorded, ranging from 18 to 30.

Selection decisions are made based on rank-order, with the number determined by class slots to fill. Because only the consensus score is recorded, interrater reliability cannot be calculated. However, it is likely to be high due to extensive training, consensus discussions, and monitoring of scores by the organization. In the highly similar context of M. C. Campion et al. (2016), the interrater reliability of a single rater was .61 and the composite of three raters was .82. The descriptive statistics on the Board scores for each of the four samples are reported in Table 2 and by flying or nonflying jobs in Table 3.

The Board scores were used as the criterion rather than whether the candidate was actually hired for two reasons. First, hiring decisions are influenced by a multitude of factors other than candidate quality that could influence the results, such as number of openings in each career field. Second, Board scores are the ideal criterion because they are the sole determinant of selection decisions, generated in a systematic and rigorous fashion, not obscured by other factors, and they are a continuous variable that avoids the statistical power loss from dichotomization.

Training Performance

Immediately following hiring, new officer recruits in the Air Force attend extensive training. Training performance is an excellent criterion for validating hiring procedures because it is the first posthire "job" assignment, it is critical to success in the Air Force because it is where new hires learn their entire job, all new hires go through the same program so the data are comparable, it is long in duration, and it has extensive performance metrics. It is a residential program that lasts between nine and 17 weeks (about 500 hr) depending on specialty. It consists of coursework and a large number of exercises. The objectives are (Allen, 2020):

> To comprehend the roles and responsibilities of an Air Force officer. Comprehend the Air Force human relations programs such as equal opportunity and treatment. Comprehend the principles and benefit of proper physical conditioning, nutrition, and lifetime wellness. Effectively apply leadership and followership skills. Comprehend the importance of adherence to Air Force core values. Effectively apply ideas verbally in a military setting. Effectively apply ideas in writing using military writing formats. Know the role of air and space power in maintaining national security. Know the role of joint operations in U.S. national security. Comprehend the principles of cross-cultural communications. (p. 1)

Officer trainees are evaluated on more than 20 metrics, all based on a 100-point scale, usually including mid-term and final evaluations clustered into four categories: (a) *academic*, consisting of tests, briefings, and papers; (b) *leadership*, consisting of scenario-based field exercises and assignments to additional duties requiring leadership; (c) *physical*, consisting of several physical fitness assessments; and (d) *presentational*, consisting of several inspections of dress, appearance, and dorms. We created equal-weighted composites for each category and a total *test performance composite* score combining all categories. We present analyses with the total score because it is the most reliable measure and results were similar for the categories.

In addition, officer trainees are evaluated at the end of OTS by the instructors (*instructor ranking*) and their peers (*peer ranking*). The rankings follow a required distribution of 10% each assigned 90–100, 80–89, 70–79, and 60–69, with the remainder assigned 0. Analyses with other values assigned to those below 60 instead of 0 (e.g., 50) yielded almost identical results. Only one instructor provided the rankings for each class, so it is not possible to estimate reliability. Only total peer rankings representing the average across peers were provided, thus reliability could not be calculated, but it should be high given it is based on 25 to 50 in a class.

Mental Ability Employment Tests

The AFOQT is an extensively researched employment test with a long history of use in the Air Force (Carretta et al., 2016). It has played a key role in selecting candidates into OTS since 1953. It has been revised several times (Drasgow et al., 2010), most recently in 2015 (Form T; Carretta et al., 2016). It consists of 10 subtests: (1) Verbal Analogies, (2) Arithmetic Reasoning, (3) Word Knowledge, (4) Math Knowledge, (5) Reading Comprehension, (6) Physical Science, (7) Table Reading, (8) Instrument Comprehension, (9) Block Counting, and (10) Aviation Information. Subtests are combined to create composites that have been validated for officer commissioning and specific occupations that the Boards use to make selection decisions. The AFOQT composites (subsets) are *Verbal* (1, 3, 5), *Quantitative* (2, 4), *Pilot* (4, 7, 8, 10), *Combat Systems Operator* (3, 4, 7, 9), and *Air and Battle Manager* (1, 4, 7, 8, 9, 10). In addition, the *Pilot Candidate Selection Method* score is a combination of the AFOQT Pilot composite, several scores from the Test of Basic Aviation Skills, and a measure of prior flying experience, interpreted as an indicator of aptitude for aviation jobs (Carretta, 2011). These composites are the primary predictors in the present study because the Boards consider them explicitly when making decisions. The Boards for flying jobs consider all the scores, while the Boards for nonflying jobs consider only the Verbal and Quantitative scores. Additional information on AFOQT subtests is available from the authors.

The reliability and validity of the AFOQT are well-documented. The internal consistency reliabilities of the version used in this study range from .74 to .91 across the 10 subtests, with 20 to 40 items each (Carretta et al., 2016). The factor composition compares reasonably well with the hypothesized structure, with the best fit for a model with five lower order factors reflecting verbal, math, spatial, perceptual speed, and aviation knowledge, and a hierarchical general mental ability factor (Carretta et al., 2016). The AFOQT has been validated in predicting officer training performance (Roberts & Skinner, 1996), and several specific training performance criteria such as completing training, grades, and class rank (e.g., Carretta, 2011; Olea & Ree, 1994). It also has predictive validity for nonflying officer jobs (e.g., Carretta, 2010).

Table 3 presents descriptive statistics. AFOQT subtests are t-scored, but all composites are percentile scores. Across all measures, the means tend toward the middle of the scales and the standard deviations are large, suggesting good variation and no extreme range restriction that might limit correlations. The intercorrelations are large as they normally are among mental ability tests and also because some composites may share a subtest.

Numeric Application Information

Numeric application information comprised variables that are typically collected on applications, including graduation years, grade point averages, number of jobs, years of jobs, and legal violations. Several additional variables were available for those with enlisted Air Force experience such as highest military grade attained, whether they

Table 2
Descriptive Statistics of All Predictors and Correlations With Board Scores

Variable	Flying with AF experience				Flying without AF experience				Nonflying with AF experience				Nonflying without AF experience			
	N	*M*	*SD*	Board scores	*N*	*M*	*SD*	Board scores	*N*	*M*	*SD*	Board scores	*N*	*M*	*SD*	Board scores
Board score	287	26.89	0.99	—	210	27.03	1.18	—	1,057	23.93	1.38	—	274	27.16	1.37	—
Mental ability scores																
Verbal	287	59.80	23.29	.19**	210	63.22	23.34	.29**	1,057	58.05	24.33	.17**	274	63.02	24.88	.29**
Quantitative	287	43.86	22.80	.24**	210	51.84	25.95	.37**	1,057	39.18	21.87	.07*	274	57.32	25.79	.22**
Pilot	287	67.57	18.71	.36**	210	64.12	24.07	.47**	1,057	49.51	24.45	.08**	274	55.67	25.52	.26**
Combat systems officer	287	70.88	20.41	.31**	210	70.18	21.37	.45**	1,057	61.75	24.96	.15**	274	66.85	24.83	.31**
Air and battle manager	287	65.60	19.33	.36**	210	64.71	22.89	.53**	1,057	51.33	24.22	.11**	274	60.24	25.49	.28**
Pilot candidate selection method	287	34.19	20.85	.28**	210	33.71	27.18	.39**	—	—	—	—	—	—	—	—
Text scores																
College degrees	287	1.08	0.54	.09	210	1.05	0.65	−.04	1,057	0.92	0.65	.07*	274	1.18	0.70	−.01
Career achievements	287	13.05	6.43	.26**	210	8.94	7.39	.26**	1,057	6.87	3.61	.30**	274	3.75	2.96	.16**
Personal achievements	287	10.00	6.71	.09	210	10.47	7.58	.21**	1,057	5.58	3.92	.06*	274	7.80	5.94	−.02
Professional achievements	287	3.43	2.72	.07	210	2.97	3.57	.20**	1,057	4.13	3.26	.08**	274	3.05	3.46	.22*
Personal/outside interest	287	7.06	3.31	.01	210	7.73	4.55	.16*	1,057	5.89	2.83	.02	274	7.40	4.31	.14**
Current job	287	1.25	0.73	.06	210	0.69	0.61	−.01	1,057	0.87	0.73	.04	274	0.71	0.54	.04
Current duties	287	12.23	5.70	.01	210	7.23	5.40	.24**	1,057	7.37	4.06	.06*	274	5.01	3.69	.14*
Supervisor	287	0.74	0.44	.03	210	0.30	0.46	.13*	1,057	0.73	0.44	.10*	274	0.33	0.47	.08
All jobs	287	5.50	2.87	.07	210	3.04	1.51	.16**	1,057	3.87	2.21	.13**	274	3.23	1.54	.15**
All duties	287	33.00	14.07	.09	210	16.39	10.83	.29**	1,057	31.46	13.39	.13**	274	18.39	12.39	.15**
Enlistment duty titles	287	1.09	0.72	.04	—	—	—	—	1,057	0.95	0.78	.04	—	—	—	—
Offenses	287	0.72	1.20	.04	210	0.64	1.06	−.08	1,057	0.56	0.98	−.04	274	0.49	0.89	.05
Statement of objectives	287	52.23	9.22	.08	210	35.89	13.61	.35**	1,057	45.49	9.21	.06*	274	31.39	13.09	.20**
Interviewer comments	287	27.61	10.36	.14*	210	9.45	4.41	.07	1,057	26.60	10.20	.10**	274	8.99	4.70	.17**
Letters of reference	287	34.00	7.83	.11*	210	60.52	23.66	.30**	1,057	34.11	9.31	.04	274	68.97	28.48	.16*
Numeric applicant information																
Year of most recent degree	278	2017.23	2.25	.14*	205	2016.71	1.81	−.05	1,038	2016.91	2.94	.15**	266	2016.22	2.74	−.16*
GPA of most recent degree	260	3.55	0.40	.36**	207	3.28	0.39	.39**	941	3.57	0.40	.37**	265	3.27	0.42	.34**
Year of second most recent degree	250	2015.61	2.29	.06	55	2015.60	2.77	−.05	963	2014.95	2.72	−.03	87	2014.44	3.68	−.11
GPA of second most recent degree	69	3.45	0.47	.39**	45	3.52	0.42	.30*	345	3.44	0.43	.35**	78	3.48	0.41	.23*
Year of third most recent degree	109	2013.72	3.30	−.08	2	—	—	—	512	2013.67	3.42	−.06	11	2013.09	4.57	−.03
GPA of third most recent degree	43	3.25	0.52	.33*	3	—	—	—	189	3.38	0.47	.15*	12	3.46	0.40	.56*
Current job year started	286	2015.90	1.90	−.05	206	2016.74	1.82	−.01	1,049	2016.11	1.64	−.001	269	2016.56	1.98	−.07
Number of jobs	286	4.83	1.86	.07	206	3.99	1.28	.12*	1,051	5.18	2.05	.10**	269	3.78	1.28	.17**
Year started first job	286	2010.64	3.30	.02	206	2012.56	2.71	−.10	1,051	2010.05	3.32	−.12**	269	2011.95	3.36	−.23**
Jobs per year	286	0.66	0.38	.06	206	.42	.41	−.08	1,051	.64	.34	−.02	269	3.36	.35	−.07
Number of violations of civil or military law	286	1.69	1.76	−.01	137	2.30	1.85	−.08	1,041	1.67	1.74	−.08**	159	2.20	1.69	.04
Age at most recent violation	192	22.98	4.08	−.11	135	21.87	2.95	.17*	702	23.89	4.34	−.001	158	22.37	3.60	.17*
Detained, confined, or on probation (1 = yes, 0 = no)	276	0.08	0.27	−.06	—	—	—	—	1,009	0.07	0.25	−.09**	—	—	—	—
Drugs or alcohol (1 = yes, 0 = no)	276	0.11	0.32	−.02	—	—	—	—	1,009	0.08	0.28	−.05	—	—	—	—
Military education year	284	2015.89	2.77	.12*	—	—	—	—	1,043	2016.01	2.79	.11**	—	—	—	—
Highest grade	287	5.25	.92	.16**	—	—	—	—	1,057	5.42	.92	.29**	—	—	—	—
Pilot license	286	0.04	0.19	.02	—	—	—	—	1,037	0.01	0.10	.03	—	—	—	—
Previous application for commissioned program (1 = yes, 0 = no)	285	0.26	0.44	.01	—	—	—	—	1,039	0.27	0.45	−.03	—	—	—	—

(*table continues*)

Table 2 (*continued*)

Variable	Flying with AF experience				Flying without AF experience				Nonflying with AF experience				Nonflying without AF experience			
	N	*M*	*SD*	Board scores	*N*	*M*	*SD*	Board scores	*N*	*M*	*SD*	Board scores	*N*	*M*	*SD*	Board scores
Number of applications	285	0.36	0.71	.004	—	—	—	—	1,041	0.38	0.73	−.02	—	—	—	—
Active duty service commitment (1 = yes, 0 = no)	247	0.02	0.13	−.05	—	—	—	—	929	0.02	0.13	−.06*	—	—	—	—
Eliminated from course (1 = yes, 0 = no)	284	0.01	0.08	.06	—	—	—	—	1,040	0.00	0.04	−.02	—	—	—	—
Number of previous enlistments	286	1.30	0.65	.02	—	—	—	—	1,035	1.33	0.67	.03	—	—	—	—

Note. GPA = grade point average; AF = air force. — = variable not applicable to that sample.
* $p < .05$. ** $p < .01$, one-tailed.

had a pilot license, active-duty commitments, number of previous enlistments, and previous applications. All numeric variables on the applications were analyzed for potential inclusion. Table 2 shows the descriptive statistics of the 23 numeric variables. Many have good variance and no apparent range restriction, but others do not or have small sample sizes, which limited correlations with Board scores or potential for inclusion in the final regression models due to missing data. Table 3 shows the intercorrelations among those retained in the final models (explained in the Results section). The intercorrelations are generally small, suggesting they are fairly independent.

Text Application Information

The application for OTS collected all forms of candidate information considered to be potentially relevant to the selection decisions. The text fields included numerous questions on educational history, work history, personal and professional achievements, personal and professional interests, legal offenses, a statement of objectives, at least one letter of reference, interviewer comments, past application information, and others. All text fields were examined for potential inclusion other than those that had too little commonality among responses to adequately text analyze such as names of colleges and schools, and past employers. While job titles were analyzed, we also created an indicator of whether the job was supervisory (0/1) because leadership experience is considered especially important. Table 4 provides examples of the text variable categories by text field. As explained in the next section, we scored the text fields based on counts of the number of categories extracted using NLP. Table 2 shows the descriptive statistics of the category count variables for the 15 text fields analyzed. Most show fairly large standard deviations compared to their means, suggesting good variation and no apparent range restriction. Table 3 shows the intercorrelations among those retained in the final models. The intercorrelations are generally small, suggesting they are fairly independent.

NLP Scores

We conducted NLP using SPSS Modeler Premium (Version 18.2.1; IBM, 2019). We followed a similar process as previous research using the same software (e.g., M. C. Campion et al., 2016; Helms et al., 2012). We used SPSS because it is a common and familiar software for most social scientists and makes NLP accessible to researchers who are not programmers.[2] Terms used by SPSS Modeler may not match the general literature, so we link to more familiar terms after the explanation below. SPSS extracts, categorizes, and scores text data through a seven-step process (Figure 2). The first three steps comprise the extraction process, whereby the algorithm extracts *n*-grams that are then scored. The fourth step is conducted to improve interpretability by arranging the *n*-grams in logical groupings (references to “airman of the quarter” fall under the broader category of “soldier”) so users can more readily understand what has been extracted.

Input Data Into System. Text data are inputted into SPSS Modeler using a “Stream” where the type of data inputted are identified using the “Type” node (to make sure the system reads the text data as string data) and is then connected to the “Text Mining” node. Data are then converted to a standard format, which is how

[2] SPSS Modeler is also available to academics for free as part of their “IBM Academic Initiative.” See here at https://www.ibm.com/academic/faqs/agreement.

Table 3
Descriptive Statistics and Intercorrelations of Study Variable

Variable	*M*	*SD*	*N*	1	2	3	4	5	6	7	8
1. Verbal	61.25	23.35	497	—	0.33**	a	a	a	a	0.02	−0.01
2. Quantitative	47.23	24.48	497	0.30**	—	a	a	a	a	−0.16**	−0.16**
3. Pilot	66.11	21.19	497	0.13**	0.46**	—	a	a	a	a	a
4. CSO	70.58	20.80	497	0.55**	0.45**	0.50**	—	a	a	a	a
5. ABM	65.22	20.89	497	0.27**	0.64**	0.90**	0.71**	—	a	a	a
6. PCSM	33.99	23.70	497	0.07	0.32**	0.80**	0.30**	0.67**	—	a	a
7. GPA	3.43	0.42	467	−0.03	−0.06	−0.02	0.02	−0.01	−0.03	—	0.20**
8. Number of Jobs	4.48	1.70	492	0.01	−0.01	0.07	0.11*	0.08	0.06	0.07	—
9. Career Achievement	11.29	7.15	497	−0.05	−0.07	0.05	0.05	0.03	0.04	0.22**	0.18**
10. Personal Achievement	10.20	7.09	497	−0.06	0.05	0.02	0.03	0.05	0.03	0.10*	0.02
11. Professional Achievement	3.23	3.11	497	−0.03	0.01	0.08	0.07	0.08	0.10*	0.12*	0.11*
12. Professional Affiliations	7.34	3.89	497	−0.08	−0.01	0.06	0.02	0.05	0.04	0.03	−0.01
13. Personal Interests	10.12	6.10	497	0.00	−0.07	0.02	0.01	0.02	−0.02	0.16**	0.01
14. Supervisor	0.56	0.50	497	−0.03	−0.09*	0.11*	−0.02	0.06	0.09	0.12*	0.36**
15. All Jobs	4.46	2.68	497	−0.05	−0.07	0.11*	0.08	0.08	0.06	0.19**	0.66**
16. All Duties	25.98	15.20	497	−0.03	−0.11*	0.05	0.05	0.04	0.01	0.27**	0.49**
17. Objectives	45.33	13.87	497	−0.02	−0.03	0.13**	0.11*	0.12**	0.09*	0.26**	0.19**
18. Interview Comments	19.94	12.28	497	−0.01	−0.13**	0.07	0.04	0.02	0.00	0.22**	0.23**
19. Letters of Reference	45.21	21.05	497	0.11*	0.19**	0.09*	0.09*	0.13**	0.13**	−0.15**	−0.15**
20. Board Score	0.00	1.00	497	0.23**	0.30**	0.41**	0.37**	0.44**	0.33**	0.36**	0.07
21. Test Performance Composite	90.35	2.60	300	0.15*	0.12*	0.14*	0.18**	0.17**	0.08	0.26**	0.06
22. Instructor Ranking	26.03	38.00	300	−0.01	−0.01	0.09	0.05	0.10	0.07	0.22**	0.09
23. Peer Ranking	27.29	38.21	300	−0.01	0.07	0.17**	0.08	0.16**	0.14*	0.21**	0.05
24. Asian	0.05	0.22	497	−0.09*	0.11*	−0.06	−0.03	−0.02	−0.05	−0.01	−0.05
25. Black	0.04	0.20	497	−0.04	−0.05	−0.15**	−0.09*	−0.11*	−0.15**	−0.08	−0.05
26. Hispanic	0.13	0.34	497	−0.10*	−0.01	−0.11*	−0.09*	−0.09*	−0.09*	−0.02	0.02
27. White	0.75	0.43	487	0.15**	0.00	0.22**	0.13**	0.17**	0.20**	0.08	0.03

Variable	9	10	11	12	13	14	15	16	17	18	19
1. Verbal	−0.02	−0.03	−0.10**	−0.01	−0.05	−0.08**	−0.02	−0.06*	−0.08**	−0.02	0.09**
2. Quantitative	−0.18**	0.05	−0.10**	0.06*	−0.14**	−0.18**	−0.12**	−0.24**	−0.18**	−0.19**	0.22**
3. Pilot	a	a	a	a	a	a	a	a	a	a	a
4. CSO	a	a	a	a	a	a	a	a	a	a	a
5. ABM	a	a	a	a	a	a	a	a	a	a	a
6. PCSM	a	a	a	a	a	a	a	a	a	a	a
7. GPA	0.28**	−0.01	0.13**	0.02	0.16**	0.15**	0.15**	0.28**	0.25**	0.22**	−0.14**
8. Number of Jobs	0.19**	−0.04	0.11**	−0.03	−0.05	0.35**	0.64**	0.50**	0.20**	0.23**	−0.18**
9. Career Achievement	—	0.21**	0.26**	0.12**	0.25**	0.22**	0.19**	0.35**	0.31**	0.26**	−0.16**
10. Personal Achievement	0.37**	—	0.20**	0.33**	0.05	−0.02	0.04	0.02	0.01	−0.12**	0.24**
11. Professional Achievement	0.42**	0.35**	—	0.13**	0.10**	0.10**	0.08**	0.17**	0.23**	0.08**	−0.02
12. Professional Affiliations	0.26**	0.39**	0.27**	—	0.06*	−0.03	0.04	0.07*	0.06*	−0.09**	0.24**
13. Personal Interests	0.25**	0.08	0.10*	0.11*	—	0.12**	0.03	0.57**	0.24**	0.16**	−0.11**
14. Supervisor	0.19**	−0.04	0.07	−0.01	0.21**	—	0.42**	0.34**	0.26**	0.27**	−0.20**
15. All Jobs	0.27**	−0.05	0.11*	0.01	0.21**	0.49**	—	0.42**	0.19**	0.14**	−0.03
16. All Duties	0.35**	0.03	0.16**	0.04	0.66**	0.43**	0.62**	—	0.37**	0.31**	−0.17**
17. Objectives	0.26**	0.05	0.13**	0.02	0.40**	0.33**	0.35**	0.51**	—	0.43**	−0.20**
18. Interview Comments	0.25**	0.00	0.06	−0.03	0.29**	0.33**	0.37**	0.47**	0.49**	—	−0.37**
19. Letters of Reference	−0.07	0.15**	0.04	0.16**	−0.18**	−0.18**	−0.24**	−0.27**	−0.24**	−0.41**	—
20. Board Score	0.25**	0.14**	0.13**	0.09	0.10*	0.07	0.09	0.14**	0.17**	0.08	0.17**
21. Test Performance Composite	0.17**	0.05	0.03	−0.03	0.11	0.20**	0.15*	0.18**	0.24**	0.29**	−0.09
22. Instructor Ranking	0.18**	0.00	−0.01	−0.05	0.11	0.19**	0.17**	0.23**	0.27**	0.24**	−0.08
23. Peer Ranking	0.14*	−0.10	0.01	−0.05	0.10	0.20**	0.16**	0.20**	0.25**	0.27**	−0.16**
24. Asian	−0.01	−0.07	0.00	−0.02	−0.12**	−0.05	−0.05	−0.07	−0.04	−0.05	−0.04
25. Black	0.06	0.06	0.14**	0.00	0.04	−0.03	−0.07	−0.06	−0.09*	−0.05	−0.06
26. Hispanic	−0.05	−0.05	0.00	−0.03	0.04	−0.02	−0.03	0.05	0.06	0.09*	−0.04
27. White	0.00	0.03	−0.12**	0.01	0.01	0.06	0.09	0.03	0.03	0.00	0.05

(*table continues*)

Table 3 (*continued*)

Variable	20	21	22	23	24	25	26	27	*N*	*SD*	*M*
1. Verbal	.20**	0.21**	0.07	0.09*	−0.12**	−0.16**	−0.11**	0.25**	1,331	24.52	59.07
2. Quantitative	.10**	0.00	−0.09*	−0.06	0.13**	−0.12**	−0.08**	0.09**	1,331	23.88	42.92
3. Pilot	a	a	a	a	a	a	a	a	a	a	a
4. CSO	a	a	a	a	a	a	a	a	a	a	a
5. ABM	a	a	a	a	a	a	a	a	a	a	a
6. PCSM	a	a	a	a	a	a	a	a	a	a	a
7. GPA	0.34**	.29**	0.26**	0.25**	−0.07*	−0.07*	−0.04	0.13**	1,206	0.42	3.51
8. Number of Jobs	0.10**	.10*	0.17**	0.11**	−0.05*	0.01	0.01	0.02	1,320	2.00	4.90
9. Career Achievement	0.26**	.14**	0.24**	0.21**	−0.06*	0.02	−0.02	0.04	1,331	3.71	6.23
10. Personal Achievement	0.03	−0.02	0.01	−0.05	0.00	−0.01	0.02	0.00	1,331	4.50	6.03
11. Professional Achievement	0.10**	−0.01	0.05	0.03	−0.02	0.08**	0.00	−0.05	1,331	3.33	3.91
12. Professional Affiliations	0.05	−0.04	−0.03	−0.03	0.01	−0.05	−0.03	0.04	1,331	3.25	6.20
13. Personal Interests	0.07**	0.07	0.16**	0.15**	−0.02	−0.04	0.04	0.00	1,331	4.10	6.88
14. Supervisor	0.09**	.15**	0.13**	0.14**	−0.05	−0.01	0.05	0.00	1,331	0.48	0.65
15. All Jobs	0.13**	.09*	0.11**	0.11**	−0.03	−0.01	0.00	0.01	1,331	2.10	3.73
16. All Duties	0.12**	.11**	0.21**	0.19**	−0.01	−0.03	0.04	−0.01	1,331	14.21	28.77
17. Objectives	0.08**	.17**	0.26**	0.24**	−0.04	0.02	0.06*	−0.03	1,331	11.62	42.58
18. Interview Comments	0.09**	.25**	0.25**	0.23**	−0.11**	0.05	0.07**	−0.02	1,331	11.74	22.97
19. Letters of Reference	0.06*	−.12**	−0.14**	−0.17**	0.06*	−0.06*	−0.04	0.03	1,331	20.84	41.29
20. Board Score	—	.30**	0.31**	0.28**	−0.05	−0.09**	−0.06*	0.14**	1,331	1.00	0.00
21. Test Performance Composite	0.36**	—	0.49**	0.46**	−0.08	−0.06	−0.09*	0.15**	622	2.49	90.66
22. Instructor Ranking	0.28**	.47**	—	0.79**	−0.02	0.04	−0.01	0.02	622	41.47	34.82
23. Peer Ranking	0.25**	.47**	0.78**	—	−0.05	0.02	−0.02	0.05	622	41.31	34.62
24. Asian	−0.07	−.11*	0.04	0.03	—	−0.08**	−0.11**	−0.39**	1,331	0.25	0.07
25. Black	−0.04	−0.05	0.00	−0.01	−0.05	—	−0.13**	−0.46**	1,331	0.29	0.09
26. Hispanic	−0.09*	−0.06	0.04	0.03	−0.09*	−0.08	—	−0.60**	1,331	0.35	0.14
27. White	0.13**	.15*	−0.05	−0.01	−0.41**	−0.37**	−0.69**	—	1,319	0.47	0.68

Note. Bottom triangle is flying sample ($N = 497$). Upper triangle is nonflying sample ($N = 1{,}331$). Pairwise deletion was used. CSO = Combat Systems Officer; ABM = Air Battle Manager; PCSM = Pilot Candidate Selection Method; GPA = grade point average. Racial categories are coded as 1 = that racial category, 0 = all else. a = Only the flying sample completed flying-related assessments (Pilot, CSO, ABM, and PCSM).
* $p < .05$. ** $p < .01$, two-tailed.

SPSS refers to preprocessing (e.g., to stem or lemmatize, eliminate stop words, etc.).[3]

Indicate NLP Settings. In the "Text Mining" node, the variable you intend to text mine (e.g., career achievements) and what resources will be used are indicated. We use the system's "Basic Resources (English)" template. This template is a linguistic package developed by SPSS that provides a general dictionary, information about terms such as their type (e.g., locations, organization, person), and part-of-speech tagging that informs extraction and categorization of the terms and *n*-grams.

Run NLP to Extract Features. The software extracts what SPSS calls "Concepts." These include single terms (e.g., "goals") and *n*-grams (e.g., bigrams such as "financial goals"). SPSS refers to these as "Uniterms" and "Multiterms," respectively; however, these are typically referenced as *n*-grams in the broader NLP literature. During extraction, the system counts the number of times the concept occurs in the corpus and in the number of documents. This process of extracting terms and counting occurrences is generally referred to as "bag-of-words" (BOW) in the NLP literature.[4] We text-mined the data for the flying and nonflying jobs separately because the jobs are distinct and the information assessed by the Boards differs slightly (e.g., nonflying jobs comprise a wider range of specialties). Within each job, we analyzed those with and without prior Air Force enlistments together because the jobs they are applying for are the same, and larger samples yield more stable results. The size of *n*-grams in our text extraction ranged from one (e.g., "captain") to five (e.g., "united states air force bases").

Generate Categories From Extracted Features. In this step, the purpose is to categorize the *n*-grams into coherent and meaningful groupings (Blei et al., 2003). SPSS employs traditional text analytics methods but refers to these techniques ("Grouping Techniques") as "Concept Inclusion" and "Semantic Networks." These are informed by two other techniques: "Concept Root Derivation" and "Co-occurrence Rules." These are described below (see also the IBM SPSS Modeler Text Analytics 18.2.1 User's Guide).

1. "Concept Inclusion": This approach determines whether an *n*-gram is a subset of other *n*-grams. Consider "shift lead" (an example from our data set). This would be grouped along with "team leadership" and "leads a group" under the larger category of "leadership" because these *n*-grams are subsets of leadership.

2. "Concept Root Derivation": This approach identifies synonyms by assessing whether the words are derived from each other by examining the root of the term. In linguistics, we assume words that are derived from each other having similar meaning. This algorithm works from this assumption to identify, for example, "plan," "planning," and "planned" as synonyms, or "opportunities to advance" and "opportunities for advancement" as synonyms.

[3] This is the same process one might code, for example, in Python using the NLP Toolkit ("NLTK"; https://www.nltk.org/), which is a package of linguistic resources to eliminate stop words (e.g., using the "stopwords" download) and stem (e.g., using the "PorterStemmer" download) or lemmatize (e.g., using the "WordNetLemmatizer" download). SPSS Modeler does this automatically.

[4] To confirm, we ran a BOW model using Python on one text field (career achievements) and found a correlation between the counts of .88 for flying jobs and .84 for nonflying jobs.

Table 4
Description of Text Variable Categories for Each Text Field

Text field	Description	Examples of text	Number of categories extracted by panel (flying/nonflying)	Final number of categories extracted by board and alphas (flying/nonflying); alphas are in parentheses	Examples of categories (features) extracted	Biodata and interview construct domain (from Speer et al., 2022; and Huffcutt et al., 2001, classifications)
College degrees	College degree and major	"BA Criminal Justice" "BS Business Administration—Finance"	39–50/37–56	NA	Technology/information technology Business/business management/logistics Science/environmental science Science/health sciences	Academic Achievement, Education and training, Mental capacity or capability, Knowledge and procedural skills
Career achievements[a]	List of career achievements	"Summa Cum Laude" Certifications Commendations	345–450/152–170	371 (.77)/153 (.61)	Military/air force/air force air medal/air force commendation medal Economics/statistics/graduation/distinguished graduate Military/air force/air force air medal/air force achievement medal Honors/honor society/national honor society	Academic Achievement, Education and training, Leadership, Knowledge and procedural skills, Conscientiousness
Personal achievements[a]	List of personal achievements	"Volunteer wreath bearer" "Booster club" "Team captain" "Habitat for Humanity Project Manager"	417–488/218–381	393 (.79)/198 (.67)	Military/troop/officer/captain/team captain Dean/dean's list Sports/sports events/athletic events/open road race/marathon Sports/sports events/championship	Leadership, Physical fitness, Interests/preferences, Academic Achievement, Education and training, Conscientiousness, Social skills/Sociability/Applied social skills, Values and moral standards
Professional affiliations[a]	List of professional affiliations	"Airmen Against Drunk Driving" Alumni associations "American Heart Association—cardiopulmonary resuscitation Instructor"	121–140/171–455	118 (.64)/84 (.53)	Society/honor society/national honor society American/American legion Mental processes/learning/association/alumni association Occupation/treasurer	Social skills/Sociability/Applied social skills, Academic Achievement, Education and training, Leadership
Personal/outside interest[a]	List of personal interests outside the military	"Camping in National Parks" Sports (e.g., basketball, golf) Church groups	209–224/153–279	207 (.54)/151 (.40)	Health and well-being/exercise/fitness Musical instruments/strings/guitar Sports/sports by type/watersports Outdoors/camping	Interests/preferences, Physical fitness, Openness to experience
Current Job	Title of current job	"Aerospace Maintenance Craftsman" "Lead Security Officer" "Realtor"	32–60/35–43	NA	Flight/flight chief Enlisted recruitment/enlisted accessions recruiter Occupation/manager/program manager	Knowledge and procedural skills, Social skills/Sociability/Applied social skills, Leadership

(*table continues*)

Table 4 (*continued*)

Text field	Description	Examples of text	Number of categories extracted by panel (flying/nonflying)	Final number of categories extracted by board and alphas (flying/nonflying); alphas are in parentheses	Examples of categories (features) extracted	Biodata and interview construct domain (from Speer et al., 2022; and Huffcutt et al., 2001, classifications)
Current duties[a]	Description of current job duties	"Oversees cyber training program" "Teach undergraduate mathematics classes"	266–452/245–246	343 (.73)/240 (.62)	Mathematical analysis/programming Occupation/educators/instructors Occupation/hotel personnel/guide Services/customer service	Knowledge and procedural skills, Social skills/Sociability/Applied social skills, Leadership
Supervisor[a]	Whether applicant had any previous supervisory experience (1 = yes, 0 = no)	"Flight Chief" "Special Projects Manager" "Shop Supervisor"	159/162	NA	Administrator/operations/operations supervisor Occupation/manager/case manager Occupation/manager/community manager Occupation/manager/store manager Occupation/programmer/program coordinator	Leadership, Social skills/Sociability/Applied social skills, Emotional stability and self-confidence, Knowledge and procedural skills
All jobs	List of all previous jobs	"Production Technician" "Plans and Programs Journeyman" "Computer Programmer" "Teaching Assistant"	115–272/158–161	168 (.28)/132 (.27)	Occupation/manager/deployment manager/unit deployment manager Occupation/white collar workers/analyst/language analyst/cryptologic language analyst Associate/sales associate Defense/squadron	Knowledge and procedural skills, Social skills/Sociability/Applied Social Skills, Leadership
All duties[a]	Description of duties across all previous jobs	"Provide counterintelligence support to force protection of DoD personnel." "Conduct unit and flight evaluations to ensure compliance." "After two years of working for [name redacted] High School in conjunction with my two-year commitment with Teach for America, I was asked to develop my own two classroom programs." "Responsible for the proper design of custom aerial bucket trucks for power utility companies"	444–1,025/668–747	673 (.89)/665 (.87)	Resources/human resource/career/career development/training Government/government agencies/U.S. government agencies/dod/unified combatant commands Military operations/missions Office workers/team	Knowledge and procedural skills, Social skills/Sociability/Applied social skills, Leadership

(*table continues*)

Table 4 (*continued*)

Text field	Description	Examples of text	Number of categories extracted by panel (flying/nonflying)	Final number of categories extracted by board and alphas (flying/nonflying); alphas are in parentheses	Examples of categories (features) extracted	Biodata and interview construct domain (from Speer et al., 2022; and Huffcutt et al., 2001, classifications)
Enlisted duty titles[b]	Titles from enlisted job	"Contract Specialist" "Technical Training Instructor"	52/36	NA	Technical personnel/technician Occupation/recruiter/enlisted recruiter Occupation/operator	Knowledge and procedural skills, Social skills/Sociability/Applied social skills, Leadership
Offenses	List of civil offenses	"Highway speeding" "Running a red light" "Disturbing the peace"	31–38/28–30	NA	Violation/traffic violation Crimes/public order crimes/drunkenness Real property/possession	Emotional stability and self-confidence, Conscientiousness
Statement of objectives[a]	Personal essay on reasons for applying to be a Commissioned Officer	"My desire to serve as a Commissioned Officer derives from my inspiration to further apply my leadership abilities, more influentially mentor Airmen, and to make an even greater impact within our military community." "I believe my drive, passion and pursuit of excellence stem from my families' indigent circumstances during my childhood." "I believe strongly in the principle of "Leaving our world better than we found it" and it is because of my belief in this principle that I want to serve my country by joining the Air Force as an officer."	518–785/558–586	618 (.81)/526 (.77)	Psychology/behavior and behavior mechanisms/emotion/passion Values/core values Occupation/operators/military operations/mission Behavior mechanisms/personality	Conscientiousness, Emotional stability and self-confidence, Openness to experience, Agreeableness, Extraversion, Interests/preferences
Interviewer comments[a]	Comments on applicant by interviewer	"Top performer" "Key leader and mentor of officer and enlisted corps" "1st in immigrant family to earn BA degree: trailblazing leadership/skills returned to local community by teaching new citizens English" "potential to grow/learn military discipline; shows capability/potential to be a successful officer"	258–803/313–522	466 (.87)/449 (.86)	Resourcefulness/human resources/career/skills Psychology/behavior and behavior mechanisms/social psychology/morale Human resources/internship/international relations/diplomat Human resources/management practices/certifications	Mental capacity or capability, Knowledge and procedural skills, Social skills/sociability, Leadership, Conscientiousness, Emotional stability and self-confidence, Extraversion, Openness to Experience, Agreeableness, Interests/preferences

(*table continues*)

Table 4 *(continued)*

Text field	Description	Examples of text	Number of categories extracted by panel (flying/nonflying)	Final number of categories extracted by board and alphas (flying/nonflying); alphas are in parentheses	Examples of categories (features) extracted	Biodata and interview construct domain (from Speer et al., 2022; and Huffcutt et al., 2001, classifications)
Letters of reference[a]	Letters of reference for applicants	"His skill and level of participation are unmatched by his peers, and I have been consistently impressed with his maturity, intelligence, and demonstrated capability." "She was always positive and always led by example. She made the extra effort to include the kids on the fringes and made them to feel welcomed and encouraged."	727–943/589–1,413	754 (.91)/624 (.89)	Professional/professionalism Psychology/behavior and behavior mechanisms/character Highs/highest recommendation Behavior mechanisms/motivation	Mental capacity or capability, Knowledge and procedural skills, Social skills/Sociability/Applied social skills, Leadership, Conscientiousness, Emotional stability and self-confidence, Openness to Experience, Agreeableness, Extraversion, Interests/preferences

Note. NA = not applicable.
[a] Indicates category was retained in final model. Alphas are reported only for those categories retained in the final model. [b] Enlisted Duty Title data only available for applicants in the "with Air Force Experience" groups.

3. "Semantic Networks": This technique generates categories through recognition of hyponyms—or specific terms for broad categories—in part from SPSS's "Basic Resources" dictionary. This technique assumes that hyponyms are semantically similar. Take an example from our data set (Table 4): in the Personal/Outside Interest text field, playing guitar was grouped underneath "strings" (as in string instruments) with the highest level category being "musical instruments."

4. "Co-occurrence Rules": This technique identifies the frequency with which *n*-grams co-occur and then creates a rule that these should be categorized together. Central to this algorithm is the assumption that *n*-grams that occur together (i.e., within 5 words) frequently reflect a meaningful underlying relationship, which is a key element of linguistics (see Jurafsky & Martin, 2021).

Although the terms are different, these techniques are conceptually similar to the familiar open-dictionary NLP techniques latent semantic analysis (LSA) and LDA (Eichstaedt et al., 2021). LSA is typically considered a dimensionality reduction technique and LDA is often synonymized with topic modeling (Blei et al., 2003; Hannigan et al., 2019; C. Zhang et al., 2021); however, LSA and LDA can be used for both purposes because they achieve the same end goal: reducing the data to coherent and meaningful groupings (Blei et al., 2003). Both techniques represent the text (*n*-grams) in a vector space. LSA uses singular value decomposition (Landauer & Dumais, 1997) to reduce a large term document matrix into a series of matrices that considers the co-occurrence of terms, among other text data characteristics, to arrive at a smaller number of terms and *n*-grams (Deerwester et al., 1990). This approach is similar to SPSS Modeler's "Co-occurrence Rules," as it explicitly considers *n*-gram co-occurrence.

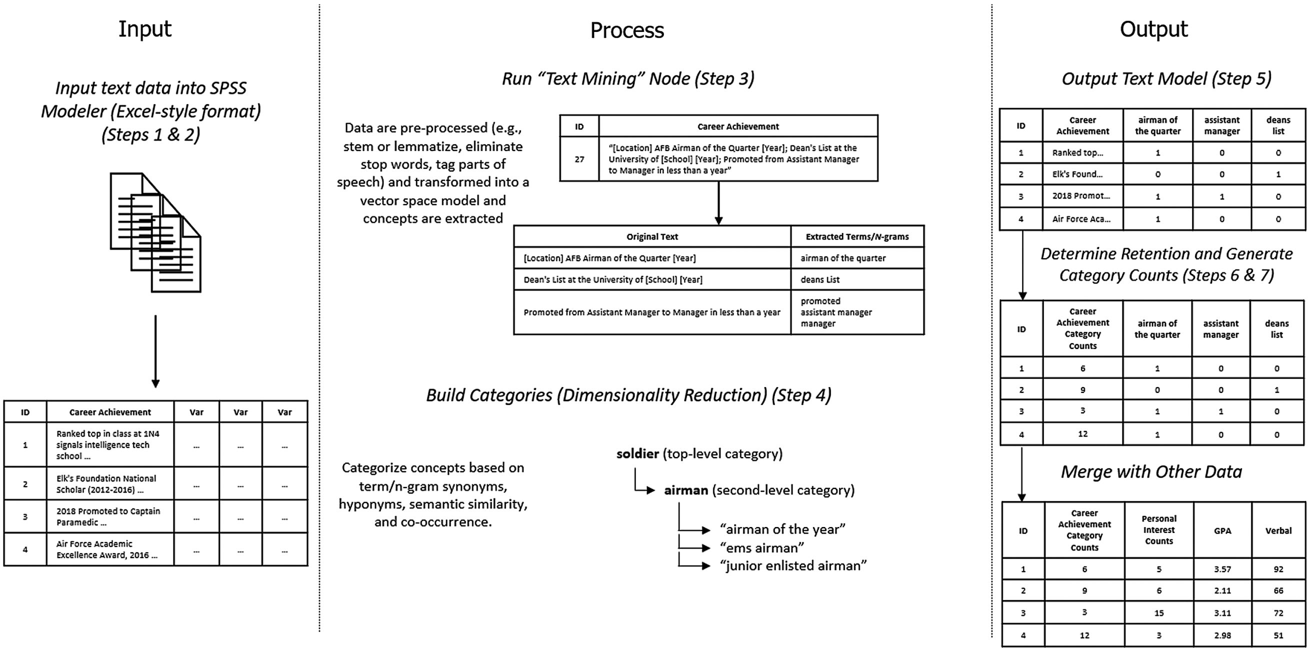
Both LSA and LDA can help identify synonyms, and they can also help identify hyponyms (Sriurai et al., 2010), which are narrower terms for broader terms, and hypernyms, which are broader terms for narrower terms (Jurafsky & Martin, 2021). For example, looking at an example of our data (Table 4) from the "Statement of Objectives" text field, "passion" is a hyponym of "emotion" and "emotion" is a hypernym of "passion." Similarly, SPSS Modeler captures synonyms using "Concept Inclusion" and "Concept Root Derivation" and homonyms and other semantic features using and "Semantic Networks." It also allows users to determine the number of top-level categories and then re-building the model. SPSS Modeler accommodates punctuation errors by normalizing (ignoring them), and it accommodates common spelling errors using "fuzzy grouping" (which allows approximate spelling matches).

In research using the same software, M. C. Campion et al. (2016) trained the text model by combining concepts that were synonyms and by combining and retaining or eliminating categories based on their similarities and meaningfulness. The same training was attempted here, which improved the apparent rational appeal of the categories, but not prediction. That finding, and to avoid the subjectivity interjected by such training, led instead to retaining all categories exceeding minimum frequencies as described below.

In Table 4, we describe the 15 text fields in the OTS applications, report the number of text categories by text field, provide illustrations of text categories extracted, and identify related construct domains based on biodata and interviewing research. We used Speer et al.'s (2022) and Huffcutt et al.'s (2001) taxonomies of construct domains in biodata and employment interview research to identify constructs in our text data

Figure 2

Input-Process-Output Model of Text Mining and Scoring


Input
Input text data into SPSS Modeler (Excel-style format) (Steps 1 & 2)
Process
Run "Text Mining" Node (Step 3)
Data are pre-processed (e.g., stem or lemmatize, eliminate stop words, tag parts of speech) and transformed into a vector space model and concepts are extracted
Build Categories (Dimensionality Reduction) (Step 4)
Categorize concepts based on term/n-gram synonyms, hyponyms, semantic similarity, and co-occurrence.
soldier (top-level category)
airman (second-level category)
"airman of the year"
"ems airman"
"junior enlisted airman"
Output
Output Text Model (Step 5)
Determine Retention and Generate Category Counts (Steps 6 & 7)
Merge with Other Data

ID	Career Achievement	Var	Var	Var
1	Ranked top in class at 1N4 signals intelligence tech school …	…	…	…
2	Elk's Foundation National Scholar (2012-2016) …	…	…	…
3	2018 Promoted to Captain Paramedic …	…	…	…
4	Air Force Academic Excellence Award, 2016 …	…	…	…

ID	Career Achievement
27	"[Location] AFB Airman of the Quarter [Year]; Dean's List at the University of [School] [Year]; Promoted from Assistant Manager to Manager in less than a year"

Original Text	Extracted Terms/*N*-grams
[Location] AFB Airman of the Quarter [Year]	airman of the quarter
Dean's List at the University of [School] [Year]	deans List
Promoted from Assistant Manager to Manager in less than a year	promoted assistant manager manager

ID	Career Achievement	airman of the quarter	assistant manager	deans list
1	Ranked top…	1	0	0
2	Elk's Found…	0	0	1
3	2018 Promot…	1	1	0
4	Air Force Aca…	1	0	0

ID	Career Achievement Category Counts	airman of the quarter	assistant manager	deans list
1	6	1	0	0
2	9	0	0	1
3	3	1	1	0
4	12	1	0	0

ID	Career Achievement Category Counts	Personal Interest Counts	GPA	Verbal
1	6	5	3.57	92
2	9	6	2.11	66
3	3	15	3.11	72
4	12	3	2.98	51

Note. GPA = grade point average.

because the narrative data in the applications are similar to biodata and interviews (which are also highly similar taxonomies to each other). The construct domains reflected by the text fields were identified independently by two authors, with minor differences discussed to consensus. In our analysis of the construct domains, we found evidence of all of Speer et al.'s (2022) and Huffcutt et al.'s (2001) construct domains: mental capacity, knowledge and procedural skills, social skills/sociability/applied social skills, agreeableness, conscientiousness, extraversion emotional stability and self-confidence, openness to experience, leadership, academic achievement, interests/preferences, and physical fitness. The most common construct domains were leadership (11 of 13 text fields), knowledge and procedural skills (10), and social skills/sociability/applied social skills (10). Online Supplemental Table A48 shows evidence of the validity of these constructs from Speer et al. (2022) and Huffcutt et al. (2001). We added evidence of subgroup differences from meta-analyses in the literature. Although this research used different methods to measure these constructs (e.g., multiple-choice or oral questions), these data illustrate the potential to increase validity and reduce differences by including measures of these constructs.

The list of categories for each text field is essentially a dictionary. Thus, we created 15 dictionaries of words reflecting the content of each text field.

Output Text Model. After generating the model, we output a document-term matrix where each candidate is a row and each column is a text category with a "1" if the candidate used that category and a "0" if not, which is generally referred to as "one-hot encoding" (Cerda et al., 2018; Deerwester et al., 1990).

Determine Feature Retention. We eliminate terms that are too rare to offer meaning. We retained text categories with at least 1% of the responses for each text field. Requiring 1% of the sample removes extremely sparse categories that will likely not be very meaningful and ensures a minimum amount of variance in the text categories so they have a chance of correlating with criteria. Previous research for the Air Force identified the 1% level as being a reasonable and useful minimum (see also C. Zhang et al., 2021).

Create Text Scores. There are a number of different ways to create text scores, but this decision is largely dependent on sample size. Sample size plays a significant role because the algorithm extracts many features and sometimes the number of features can outnumber the modest samples. Whereas data scientists may be used to sample sizes in the tens and hundreds of thousands, organizational psychologists typically have *N*s in the hundreds. We opted to use category counts as our text scores because we lacked a sample size that would allow data-derived weights such as regression. M. C. Campion et al. (2016) were able to use regression to combine the text categories because they had large sample sizes (more than 40,000). In this study, our sample size for flying jobs is 497 and for nonflying jobs is 1,331. We generated category counts by summing the number of categories present in each application blank for two reasons.[5] First, given our sample sizes, we were not able to model all

[5] Another common metric is TF-IDF (term frequency-inverse document frequency). This is a well-known metric that prioritizes rare terms because it assumes that rarer terms are more discriminatory. In addition to its use in quantifying text data to predict an outcome, TF-IDF has been a popular metric to improve information retrieval. Information retrieval is another historic and important use of NLP and constitutes the process used by search engines to help users identify the appropriate and correct websites for their query (Jurafsky & Martin, 2021; Ramos, 2003). We do not utilize TF-IDF in our research because our metric (count) already considers *n*-gram rarity. This is like a typical test situation where individuals score higher should they answer difficult questions correctly. In the current context, we are scoring the amount of attributes, so the purpose is the opposite and counts are more appropriate.

the features in a single regression because the number of text categories compared to the sample sizes was very low, and sometimes less than 1-to-1. Second, we found that category counts produced correlations with human ratings (Board scores) as high, or higher, than cross-validated multiple correlations from regressions with small samples because of the great amount of shrinkage. Category counts do not need to be cross-validated because the equal weights are not derived from the sample. With two exceptions, the internal consistency reliabilities of the features extracted from each text field exceeded .60 and averaged .72 for flying jobs and .65 for nonflying jobs (see online Supplemental Table A48). Note that categories were only counted once in this measure. Essentially, candidates were given 1 "point" if a category was in the text, regardless of how many times it was included. We learned in this and previous research that counting a category each time it was mentioned reduces prediction perhaps because it did not reflect more of the underlying construct compared to the number of distinct categories present, and it may instead reflect repetition.

An important consideration is whether to use simple word count as another measure or as a control. Conceivably, verbose candidates may benefit from writing more. We contend that candidates provide longer responses because they have more of the attribute or skill being measured. For example, if candidates have abundant leadership experience from which to speak, they will write more about leadership than candidates who have little leadership experience. Further, this is a high-stakes hiring context where candidates are motivated to provide as much information as possible. However, this raises a related concern that candidates who are more verbose simply have more experience, making age a possible confound. We found that age and word count did not share a relationship for flying jobs ($r = -0.01$, $p = .46$), although they did for nonflying jobs ($r = 0.39$, $p < .01$). Because age is not a variable that we could legally include in a selection procedure, we do not include it in our models to test our hypotheses. To check if verbosity played a role, we tested Hypotheses 2 and 3 controlling for word count and found that word count was nonsignificant in nearly every test and did not tend to influence the R. In the one model where word count was significant (predicting Board scores), the coefficient was negative. These results can be found in online Supplemental Tables A33–A42. We also controlled for average word length because it may reflect more complex writing and found it was significant in a few cases, but showed small effects with opposites signs, and did not influence the R or it increased it by .01 in a couple of cases (online Supplemental Tables A49–A58).

The means and standard deviations of the text scores in Table 2 represent how many of the extracted text categories were counted on average across sample members. For example, the text mining of the "all jobs" field of the nonflying sample identified 132 categories (Table 4), and the average candidate had a score of 3.23 (Table 2), indicating about three previous jobs. Note that for text variables missing information is considered relevant and retained. This is because the lack of an experience or credential is meaningful (e.g., lack of professional affiliation, achievements, or jobs). To ensure the correctness of this approach, we compared correlations with Board scores counting (by assigning a count of 0) and omitting missing data and found correlations were slightly larger when missing information was counted.

Looking across all of the text-mining results, we make three overall observations. First, the number of categories for each text field appears logical such that fields with more variety in responses had more categories. Thus, categories like degree, professional affiliations, current job, enlisted duty title, and offenses had relatively fewer categories compared to achievements, all duties, objectives, and letters of reference. Second, the number of categories for each text field is large and the standard deviations show variance, indicating that the jobs attract candidates from a range of backgrounds, and it is desirable for government jobs to represent the diversity of the U.S. population. Further, the large number of text categories and the standard deviations indicate that they capture many differences between candidates, which increases the likelihood of being able to statistically predict a criterion. Third, the means and standard deviations show that the average candidate is scored on many text categories, suggesting the text categories are measuring many aspects of the candidate's application. There are also some notable differences between the results for the flying and nonflying samples. For example, the flying sample has higher means on career achievements, personal achievements, and current duties.

Transparency and Openness

We describe our sampling plan, data exclusions, and measures. We adhere to the *Journal of Applied Psychology* methodological checklist. Analysis code, research material, and data are not available due to their proprietary nature. Data were analyzed using SPSS Modeler Premium (Version 18.2.1; IBM, 2019) and SPSS Statistics (Version 27). Study design, hypotheses, and analyses were not preregistered because data were collected for an applied project. This research is deemed exempt (Old Dominion University Institutional Review Board 1681843-1).

Results

The bivariate correlations with Board scores are in Table 2. We retained predictor variables—mental ability tests, numeric variables, and text scores—for the regression models if they cross-validated in at least one flying sample and one nonflying sample (with no sign reversals; $p < .05$, 1-tailed). For variables only relevant to applicants with Air Force experience, we retained those that cross-validated in both samples. In all, we retained 11 of the 15 text variables and two of the 23 numeric variables across samples (or four of the 23 in analyses by board and panel in online Supplemental Tables A1–A8). About five of the numeric variables were limited by low applicability causing restriction of range and about eight others had smaller sample sizes causing power loss. This was mostly the case for the variables only applicable to those with Air Force experience.

All tests of hypotheses are based on two samples: the combined sample for flying jobs and for nonflying jobs. Tests of hypotheses for the four individual samples are online Supplemental Tables A1–A8. Table 3 presents the intercorrelations between the text variables and the criteria using two-tailed tests for descriptive purposes. The correlations between text scores and the criteria—Board Scores and the three job performance measures—were significant for 65% of the relationships using two-tailed, bivariate tests. Because the hypotheses are directional, we also examined these relationships using one-tailed tests and found that 76% of the relationships are significant, which offers initial (bivariate) support for our hypotheses (see online Supplemental Table A59, for the one-tailed correlation matrix). We use one-tailed tests for all other tables because they test directional hypotheses.

We created six regression models to test Hypotheses 1–3, with each including a different combination of the three sets of variables (test scores, numeric application information, and NLP scores). Hypothesis 1 predicted that computer scores of all application information combined will correlate as highly with human ratings as human ratings do with each other of .60. Model VIs in Table 5 test this hypothesis. The *R*s are .62 for flying jobs and .46 for nonflying jobs. The *R*s for the four individual samples (online Supplemental Tables A1–A8) were .64, .71, .51, and .54, with an average (via Fisher's *r*-to-*z* transformation) of .60. Thus, Hypothesis 1 is supported for the flying sample but not in the nonflying sample.[6]

Adjusted ΔR^2 are interpreted for Hypotheses 2 and 3 to consider potential shrinkage. Fivefold cross-validated ΔR^2 are also reported in parentheses in Table 5. Hypothesis 2a predicted that NLP scores will have incremental validity in the prediction of Board scores beyond mental ability tests. Model IIs in Table 5 show the ΔR^2 is significant in both samples, with values of .07 and .09. The inclusion of the text variables explains an additional 7%–9% of variance in Board scores beyond mental ability tests. Thus, Hypothesis 2a is supported. Hypothesis 2b predicted incremental validity beyond numeric application information. Model IVs in Table 5 show the ΔR^2 is significant in both samples, explaining an additional 5%–6% of variance. Thus, Hypothesis 2b is supported. Hypothesis 2c predicted incremental validity beyond both mental ability tests and numeric application information. Model VIs in Table 5 show the ΔR^2 is significant in both samples, explaining an additional 3%–6% of variance. Thus, Hypothesis 2c is supported.

Hypothesis 3a predicted incremental validity in the prediction of training performance beyond mental ability tests. Model IIs in Tables 6–8 test this hypothesis for the three training performance criteria. The ΔR^2 is significant for all three criteria in both samples, with values of .09 and .07 for test performance, .08 and .09 for instructor rankings, and .10 and .08 for peer rankings, for an average of 8.5% additional variance explained. Thus, Hypothesis 3a is supported. Hypothesis 3b predicted incremental validity beyond numeric application information. Model IVs in Tables 6–8 test this hypothesis. The ΔR^2 is significant for all three criteria in both samples, with values of .06 and .03 for test performance, .06 and .03 for instructor rankings, and .08 and .04 for peer rankings, for an average of 5% additional variance explained. Thus, Hypothesis 3b is supported. Hypothesis 3c predicted incremental validity beyond both mental ability employment tests and numeric application information. Model VIs in Tables 6–8 test this hypothesis. The ΔR^2 is significant for all three criteria in both samples, with values of .06 and .03 for test performance, .06 and .03 for instructor rankings, and .08 and .06 for peer rankings, for an average of 5.3% additional variance explained. Thus, Hypothesis 3c is supported. Hypothesis 3d predicted incremental validity beyond Board scores. Model IIs in Table 9 test this hypothesis. The ΔR^2 is significant for all three criteria in both samples, with values of .07 and .04 for test performance, .08 and .06 for instructor rankings, and .09 and .05 for peer rankings, for an average of 6.5% additional variance explained. Thus, Hypothesis 3d is supported.

Hypothesis 4 examined subgroup differences. As descriptive data before testing this hypothesis, Table 10 shows the mean subgroup differences by variable. There are significant differences on most of the mental ability test scores, and some of the numeric variables, with nonracial minorities typically scoring higher. There are a few differences in text variables, but directionality is mixed. For example, Black candidates score notably higher than White candidates on professional affiliations in both samples.

Hypothesis 4a predicted that combining NLP scores with mental ability tests will yield smaller subgroup differences. To test this hypothesis, we used regression to create predicted values of the Board scores based on NLP and mental ability scores. Table 11 shows the mean differences between predicted values based on the mental ability tests (Model I) and the mental ability tests and NLP scores combined (Model II). The table also shows the *d*s between subgroups and the *t*-tests comparing the *d*s of each model. The *d* does not decrease in the Asian-White comparison, but in fact increases from 0.33 to 0.39 for flying jobs, although it decreases significantly for nonflying jobs from 0.50 to 0.38. The *d* decreases but not significantly in the Black–White comparison for flying jobs, although it decreases significantly for nonflying jobs from 0.75 to 0.44. Finally, the *d* decreases but not significantly in the Hispanic-White comparison for flying jobs, although it decreases significantly for nonflying jobs from 0.47 to 0.32. Due to the lack of support in the flying sample, we conducted supplemental analyses by those with and without Air Force experience (online Supplemental Tables A9–A32). Results show that there is a reduction of subgroup differences for those applying to flying jobs who have Air Force experience for all three comparisons, but not for those without Air Force experience. Thus, Hypothesis 4a is supported for those applying to nonflying jobs, and also those applying to flying jobs who have Air Force experience, but not for those applying to flying jobs without Air Force experience.

Hypothesis 4b predicted that combining NLP scores with numeric application information will yield smaller subgroup differences. Table 11 shows the mean differences between predicted values of the regression models based on the numeric application information (Model III) and the numeric application information and NLP scores combined (Model IV). The *d* does not change significantly in either the Asian-White or Black–White comparisons for flying or nonflying jobs. The *d* decreases significantly in the Hispanic-White comparison for flying jobs from 0.11 to 0.05, but not for nonflying jobs. Again, we analyzed these differences by subsample (online Supplemental Tables A9–A32). We found that there is a reduction in subgroup differences in all three comparisons for those applying to flying jobs who have Air Force experience, but not for those without Air Force experience. For nonflying jobs, there is a reduction in subgroup differences in all three comparisons for those with Air Force experience but is only significantly reduced in the Black–White comparison for those without Air Force experience. Thus, Hypothesis 4b is supported in two of the four subsamples (those with Air Force experience).

Hypothesis 4c predicted that combining NLP scores with mental ability tests and numeric application information will yield smaller

[6] Interview comments are from any current officer, not Board members. They are not hiring interviews but more like recommendations. We tested Hypotheses 1–3 excluding interviewer comments from our model and found that the *R*s for the full models remained the same, supporting Hypothesis 1 for flying but not nonflying jobs without interviewer comments. Further, the average adjusted ΔR^2 predicting board scores for flying jobs is 5.67% and 4.75% for nonflying jobs (compared to 6.92% and 5.50%, respectively, with interviewer comments), and 6.67% for flying jobs and 4.33% for nonflying jobs predicting test performance, instructor rankings, and peer rankings beyond board scores (compared to 8.00% and 5.00%, respectively, with interview comments), replicating the results for Hypotheses 2 and 3, although with smaller effects. Results are in the online Supplemental Tables A43–A47.

Table 5

Regressions Predicting Board Scores

	Flying jobs[a]						Nonflying jobs[b]					
	Model I	Model II	Model III	Model IV	Model V	Model VI	Model I	Model II	Model III	Model IV	Model V	Model VI
DV = board score	β	β	β	β	β	β	β	β	β	β	β	β
Mental ability scores												
Verbal	0.10*	0.11*			0.07	0.07	0.19**	0.17**			0.13**	0.13**
Quantitative	0.05	0.09			0.09	0.09	0.04	0.10**			0.11**	0.12**
Pilot	0.16	0.20			0.12	0.17						
Combat systems officer	0.11	0.09			0.09	0.08						
Air and battle manager	0.11	0.04			0.11	0.07						
Pilot candidate selection method	0.08	0.05			0.13*	0.10						
Numeric applicant information												
GPA of most recent degree			0.34**	0.33**	0.36**	0.34**			0.33**	0.29**	0.34**	0.29**
Number of jobs			0.04	0.07	0.01	0.03			0.05*	0.01	0.07*	0.00
NLP scores												
Career achievements		0.21**		0.16**		0.17**		0.25**		0.21**		0.20**
Personal achievements		0.03		−0.01		0.00		−0.06*		−0.04		−0.05
Professional affiliations		−0.02		0.00		−0.03		0.06*		0.03		0.05*
Personal interests		−0.01		0.00		−0.01		0.01		0.02		0.01
Current duties		−0.02		0.02		0.00		0.03		0.05		0.05
Supervisor		0.00		0.00		0.01		0.03		0.01		0.03
All jobs		−0.05		−0.06		−0.09		0.07*		0.10**		0.10**
All duties		0.10		−0.02		0.04		0.01		−0.07		−0.05
Objectives		0.08		0.10*		0.04		−0.01		−0.03		−0.02
Interviewer comments		0.02		0.03		0.01		0.05		0.05†		0.05*
Letter of reference		0.15**		0.24**		0.15**		0.10*		0.15**		0.13**
R	0.47**	0.55**	0.35**	0.45**	0.58**	0.62**	0.20**	0.37**	0.34**	0.42**	0.39**	0.46**
R^2	0.22**	0.30**	0.12**	0.20**	0.33**	0.38**	0.04**	0.14**	0.12**	0.18**	0.16**	0.22**
Adjusted R^2	0.21**	0.28**	0.12**	0.18**	0.32**	0.35** (0.30)	0.04**	0.13**	0.12**	0.17**	0.15**	0.21** (0.19)
ΔR^2		0.09**		0.08**		0.05**		0.10**		0.06**		0.06**
Adjusted ΔR^2		0.07**		0.06**		0.03**		0.09**		0.05**		0.06**

Note. Standardized coefficients, or Betas, are presented for interpretability. Therefore, *SE*s for unstandardized regression coefficients are omitted. Numbers in parentheses in Models IVs with adjusted R^2s are fivefold cross-validation R^2s. DV = dependent varaible; GPA = grade point average; NLP = natural language processing; *SE* = standard error.

[a] $N = 464–497$. [b] $N = 1,205–1,331$.

† $p = 0.05$, one-tailed. * $p < .05$. ** $p < .01$.

Table 6

Regressions Predicting Training Test Performance

	Flying jobs[a]						Nonflying jobs[b]					
	Model I	Model II	Model III	Model IV	Model V	Model VI	Model I	Model II	Model III	Model IV	Model V	Model VI
DV = test performance	β	β	β	β	β	β	β	β	β	β	β	β
Mental ability scores												
Verbal	0.10	0.11			0.12†	0.13*	0.22**	0.21**			0.21**	0.22**
Quantitative	0.02	0.11			0.05	0.11	−0.05	0.02			0.03	0.07
Pilot	0.13	0.06			0.21	0.10						
Combat systems officer	0.04	0.05			0.05	0.05						
Air and battle manager	0.06	0.01			−0.01	0.00						
Pilot candidate selection method	−0.10	−0.05			−0.12	−0.06						
Numeric applicant information												
GPA of most recent degree			0.25**	0.16**	0.27**	0.17**			0.29**	0.24**	0.30**	0.25**
Number of jobs			0.05	−0.04	0.04	−0.08			0.04	−0.03	0.04	−0.06
Text scores												
Career achievements		0.12*		0.09		0.10		0.04		0.02		0.01
Personal achievements		0.04		0.01		0.01		0.01		0.03		0.02
Professional affiliations		−0.04		−0.04		−0.05		−0.05		−0.07†		−0.06
Personal interests		−0.02		−0.01		0.01		−0.01		−0.02		−0.01
Current duties		−0.03		−0.08		−0.06		0.00		0.00		0.00
Supervisor		0.11†		0.10		0.11†		0.08*		0.05		0.07†
All jobs		−0.02		0.01		0.03		0.04		0.05		0.07
All duties		0.04		0.02		0.04		−0.05		−0.06		−0.04
Objectives		0.09		0.13*		0.11		0.08*		0.07		0.08*
Interviewer comments		0.21**		0.21**		0.19**		0.18**		0.18**		0.17**
Letter of reference		0.02		0.09		0.05		−0.03		0.02		0.00
R	0.21*	0.41**	0.26**	0.40**	0.35**	0.46**	0.21**	0.35**	0.30**	0.36**	0.37**	0.43**
R^2	0.05*	0.17**	0.07**	0.16**	0.12**	0.21**	0.05**	0.13**	0.09**	0.13**	0.14**	0.19**
Adjusted R^2	0.03*	0.12**	0.06**	0.12**	0.10**	0.16** (0.12)	0.04**	0.11**	0.08**	0.11**	0.13**	0.16** (0.12)
ΔR^2		0.12**		0.09**		0.09**		0.08**		0.04**		0.05**
Adjusted ΔR^2		0.09**		0.06**		0.06**		0.07**		0.03**		0.03**

Note. Standardized coefficients, or Betas, are presented for interpretability. Therefore, *SE*s for unstandardized regression coefficients are omitted. Numbers in parentheses with adjusted R^2s are fivefold cross-validated R^2s. DV = dependent variable; *SE* = standard error; GPA = grade point average.

[a] N = 282–300. [b] N = 568–622.

† p = .05, one-tailed. * p < .05. ** p < .01.

Table 7
Regressions Predicting Instructor Rankings

	Flying jobs[a]						Nonflying jobs[b]					
	Model I	Model II	Model III	Model IV	Model V	Model VI	Model I	Model II	Model III	Model IV	Model V	Model VI
DV = instructor ranking	β	β	β	β	β	β	β	β	β	β	β	β
Mental ability scores												
Verbal	0.00	0.00			0.03	0.03	0.09*	0.08*			0.09*	0.08*
Quantitative	−0.13	−0.03			−0.08	−0.03	−0.11**	−0.01			−0.01	0.03
Pilot	−0.08	−0.11			0.08	−0.01						
Combat systems officer	−0.06	−0.06			−0.08	−0.09						
Air and battle manager	0.29	0.21			0.18	0.18						
Pilot candidate selection method	−0.01	0.03			−0.05	0.00						
Numeric applicant information												
GPA of most recent degree			0.21**	0.12*	0.21**	0.13*			0.23**	0.15**	0.24**	0.16**
Number of jobs			0.08	−0.05	0.06	−0.05			0.14**	0.08	0.14**	0.07
Text scores												
Career achievements		0.15*		0.14*		0.15*		0.12**		0.07		0.07
Personal achievements		−0.03		−0.05		−0.05		0.02		0.03		0.02
Professional affiliations		−0.07		−0.08		−0.08		−0.04		−0.04		−0.03
Personal interests		−0.06		−0.04		−0.04		−0.03		−0.03		−0.02
Current duties		−0.11		−0.14		−0.13		0.03		0.06		0.06
Supervisor		0.06		0.07		0.06		0.01		−0.02		−0.01
All jobs		−0.05		−0.02		−0.01		0.02		0.01		0.02
All duties		0.16		0.14		0.13		0.04		−0.02		−0.01
Objectives		0.16*		0.17*		0.16*		0.14**		0.14**		0.14**
Interviewer comments		0.12†		0.12*		0.12†		0.10*		0.11*		0.11*
Letter of reference		0.05		0.08		0.07		−0.05		0.00		−0.01
R	0.14	0.36**	0.23**	0.39**	0.27**	0.40**	0.13**	0.34**	0.30**	0.38**	0.31**	0.39**
R^2	0.02	0.13**	0.05**	0.15**	0.07**	0.16**	0.02**	0.12**	0.09**	0.14**	0.10**	0.15**
Adjusted R^2	0.00	0.08**	0.05**	0.11**	0.04**	0.10** (0.06)	0.01**	0.10**	0.09**	0.12**	0.10**	0.13** (0.10)
ΔR^2		0.11**		0.10**		0.09**		0.10**		0.06**		0.06**
Adjusted ΔR^2		0.08**		0.06**		0.06**		0.09**		0.03**		0.03**

Note. Standardized coefficients, or Betas, are presented for interpretability. Therefore, *SE*s for unstandardized regression coefficients are omitted. Numbers in parentheses with adjusted R^2s are fivefold cross-validated R^2s. DV = dependent variable; *SE* = standard error; GPA = grade point average.
[a] $N = 282–300$. [b] $N = 568–622$.
† $p = .05$, one-tailed. * $p < .05$. ** $p < .01$.

Table 8

Regressions Predicting Peer Rankings

	Flying jobs[a]						Nonflying jobs[b]					
DV = peer ranking	Model I β	Model II β	Model III β	Model IV β	Model V β	Model VI β	Model I β	Model II β	Model III β	Model IV β	Model V β	Model VI β
Mental ability scores												
Verbal	–0.02	–0.02			0.03	0.02	0.11**	0.10*			0.11*	0.11*
Quantitative	–0.07	0.03			–0.02	0.05	–0.08*	0.03			0.00	0.07
Pilot	–0.03	–0.12			0.12	–0.03						
Combat systems officer	–0.05	–0.06			–0.07	–0.08						
Air and battle manager	0.25	0.22			0.14	0.19						
Pilot candidate selection method	0.04	0.10			0.00	0.08						
Numeric applicant information												
GPA of most recent degree			0.20**	0.12*	0.20**	0.12*			0.23**	0.14**	0.23**	0.16**
Number of jobs			0.05	–0.09	0.04	–0.11			0.09*	–0.04	0.09*	–0.05
Text scores												
Career achievements		0.13*		0.12*		0.14*		0.11*		0.07		0.07
Personal achievements		–0.15*		–0.14*		–0.15*		–0.04		–0.03		–0.03
Professional affiliations		–0.02		–0.01		–0.02		–0.04		–0.04		–0.03
Personal interests		0.00		0.00		0.00		0.00		0.00		0.00
Current duties		–0.09		–0.18*		–0.16*		0.04		0.02		0.03
Supervisor		0.07		0.09		0.08		0.03		0.01		0.03
All jobs		–0.06		–0.01		0.00		0.04		0.08		0.09†
All duties		0.09		0.14		0.13		0.00		0.00		0.02
Objectives		0.12†		0.13*		0.11		0.14**		0.14**		0.15**
Interviewer comments		0.16*		0.15*		0.16*		0.07		0.09*		0.09*
Letter of reference		–0.05		0.01		–0.03		–0.09*		–0.05		–0.06
R	0.18	0.40**	0.21**	0.40**	0.30**	0.45**	0.12*	0.33**	0.26**	0.36**	0.28**	0.39**
R^2	0.03	0.16**	0.05**	0.16**	0.09**	0.20**	0.01*	0.11**	0.07**	0.13**	0.08**	0.15**
Adjusted R^2	0.01	0.11**	0.04**	0.12**	0.06**	0.14** (0.13)	0.01*	0.09**	0.07**	0.11**	0.07**	0.13** (0.09)
ΔR^2		0.12**		0.11**		0.11**		0.09**		0.06**		0.07**
Adjusted ΔR^2		0.10**		0.08**		0.08**		0.08**		0.04**		0.06**

Note. Standardized coefficients, or Betas, are presented for interpretability. Therefore, *SE*s for unstandardized regression coefficients are omitted. Numbers in parentheses with adjusted R^2s are fivefold cross-validated R^2s. DV = dependent variable; *SE* = standard error; GPA = grade point average.

[a] N = 282–300. [b] N = 568–622.

† p = 0.05, one-tailed. * $p < .05$. ** $p < .01$.

Table 9
Regressions Predicting Test Performance, Instructor Ranking, and Peer Ranking From Board and Text Scores

	Flying jobs^{a}						Nonflying jobs^{b}					
	Test performance		Instructor ranking		Peer ranking		Test performance		Instructor ranking		Peer ranking	
Variable	Model I β	Model II β	Model I β	Model II β	Model I β	Model II β	Model I β	Model II β	Model I β	Model II β	Model I β	Model II β
Board scores	0.36**	0.36**	0.28**	0.25**	0.25**	0.26**	0.30**	0.26**	0.31**	0.25**	0.28**	0.23**
Text scores												
Career achievements		0.02		0.09		0.06		−0.01		0.08†		0.06
Personal achievements		0.03		−0.03		−0.15*		0.01		0.02		−0.04
Professional affiliations		−0.03		−0.07		−0.01		−0.06		−0.05		−0.05
Personal interests		−0.03		−0.04		0.02		−0.01		−0.03		0.01
Current duties		0.00		−0.08		−0.07		0.00		0.03		0.04
Supervisor		0.09		0.06		0.08		0.05		−0.01		0.02
All jobs		0.04		−0.01		−0.02		0.03		0.02		0.03
All duties		−0.04		0.12		0.05		−0.05		0.04		−0.01
Objectives		0.08		0.15*		0.12*		0.07		0.13**		0.13**
Interviewer comments		0.20**		0.11		0.14*		0.17**		0.07		0.05
Letter of reference		−0.03		0.01		−0.08		−0.03		−0.05		−0.09*
R	0.36**	0.48**	0.28**	0.42**	0.25**	0.43**	0.30**	0.38**	0.31**	0.41**	0.28**	0.38**
R^2	0.13**	0.23**	0.08**	0.18**	0.06**	0.18**	0.09**	0.14**	0.10**	0.17**	0.08**	0.14**
Adjusted R^2	0.13**	0.20**	0.07**	0.15**	0.06**	0.15**	0.09**	0.13**	0.09**	0.15**	0.08**	0.13**
ΔR^2		0.10**		0.10**		0.12**		0.05**		0.07**		0.07**
Adjusted ΔR^2		0.07**		0.08**		0.09**		0.04**		0.06**		0.05**

Note. Standardized coefficients, or Betas, are presented for interpretability. Therefore, *SE*s for unstandardized regression coefficients are omitted. *SE* = standard error.
[a] $N = 300$. [b] $N = 622$.
† $p = 0.05$, one-tailed. * $p < .05$. ** $p < .01$.

subgroup differences. Table 11 shows the mean differences between the mental ability tests and the numeric application information combined (Model V) and the mental ability tests, numerical application information, and NLP scores combined (Model VI). The *d* is not significant in the Asian-White comparison for flying jobs but decreases significantly for nonflying jobs from 0.43 to 0.39. The *d* is not significant in the Black–White comparison for flying jobs, but it decreases significantly for nonflying jobs from 0.63 to 0.55. Finally, the *d* is not significant in the Hispanic-White comparison for flying jobs, but it decreases significantly for nonflying jobs from 0.37 to 0.33. Examined by subsample, we found significant reductions for all three comparisons for flying jobs with Air Force experience, but not for applicants without Air Force experience. Taken together, Hypothesis 4c is supported for those applying to nonflying jobs, and also those applying to flying jobs who have Air Force experience, but not for those applying to flying jobs without Air Force experience.

We conducted hypothetical adverse impact analyses to illustrate the effect sizes of these reductions in subgroup differences. Although impact reduction should be related to subgroup differences, adverse impact in practice is situationally specific, depending on the unique pool of candidates, skew in the distributions of scores, selection rates, and other factors, especially with small subsamples (Arthur & Woehr, 2013; Arthur et al., 2013). We conducted hypothetical analyses because the organization requested actual adverse impact ratios not be shared. Although the organization's actual hiring across years by minority subgroup generally tracks the labor market availability for these types of jobs, adverse impact may be a concern when the number of candidates far exceeds the number of openings and low selection ratios must be used.

Table 12 shows the hypothetical adverse impact ratios (passing rate of minorities divided by the passing rate of nonminorities) at three potential selection ratios (75%, 50%, and 25%) for the six regression models in each job. The impact is reduced (ratio increased in size) for the model with the mental ability tests and NLP scores combined (Model II) compared to the model with the mental ability tests alone (Model I) for four of the nine comparisons (across the three subgroup comparisons and three selection ratios) for flying jobs and nine of the nine for nonflying jobs. Impact is reduced for the model with the numeric information and NLP scores combined (Model IV) compared to the model with the numeric information alone (Model III) for seven of the nine comparisons for flying jobs and five of the nine for nonflying jobs. Finally, the impact is reduced for the model with the mental ability tests, numeric application information, and NLP scores combined (Model VI) compared to the model with just the mental ability tests and the numeric information (Model V) for six of the nine comparisons for flying jobs and six of nine for nonflying jobs. Thus, adverse impact was reduced in 16 of 27 comparisons for flying jobs and 20 of the 27 comparisons for nonflying jobs. The average change in the adverse impact ratio is 0.07 for both the flying jobs and nonflying jobs. The least improvement is at the high selection ratio and the most at the low selection ratio as expected, but there was wide variation due to the distribution of scores and small samples.

Discussion

The purpose of this research was to demonstrate that NLP can be used on an array of narrative application data in concert with mental ability tests and numeric application information to increase validity and reduce subgroup differences by measuring additional constructs. We present NLP as an emerging scoring method that shows promise in response to the enduring validity-adverse impact dilemma. Across

Table 10
Race Differences in Model Variables

	Asian			Asian–White	Black			Black–White	Hispanic			Hispanic–White	White		
Variable	N	M	SD	d	N	M	SD	d	N	M	SD	d	N	M	SD
							Flying jobs								
Mental ability scores															
Verbal	26	51.96	24.07	0.48*	21	56.95	24.70	0.26	67	55.30	23.27	0.34*	365	63.05	23.09
Quantitative	26	58.54	29.78	−0.47*	21	41.33	22.17	0.23	67	46.84	22.37	0.01	365	47.00	24.28
Pilot	26	60.65	23.27	0.39*	21	51.24	20.07	0.86**	67	60.03	20.73	0.43**	365	68.77	20.50
Combat systems officer	26	67.85	17.67	0.21	21	61.71	22.61	0.51*	67	65.85	23.65	0.30*	365	71.97	20.19
Air and battle manager	26	63.81	22.79	0.16	21	53.95	17.63	0.65**	67	60.49	20.31	0.32*	365	67.16	20.63
Pilot candidate selection method	26	28.73	24.98	0.34†	21	16.95	15.50	0.84**	67	28.55	22.34	0.35**	365	36.74	23.84
Numeric applicant information															
GPA of most recent degree	23	3.41	0.38	0.09	20	3.27	0.39	0.44*	58	3.40	0.43	0.12	348	3.45	0.42
Number of jobs	23	4.09	1.59	0.24	21	4.05	1.80	0.27	65	4.55	1.74	−0.04	365	4.49	1.68
Text scores															
Career achievements	26	11.15	8.60	0.03	21	13.33	7.17	−0.28	67	10.34	7.27	0.15	365	11.36	6.98
Personal achievements	26	8.04	6.98	0.32	21	12.19	5.45	−0.27	67	9.33	6.96	0.14	365	10.33	7.13
Professional affiliations	26	3.27	3.27	−0.09	21	5.24	3.66	−0.77**	67	3.22	2.95	−0.08	365	3.00	2.85
Personal interests	26	7.08	3.89	0.06	21	7.29	4.61	0.01	67	7.01	3.23	0.08	365	7.33	3.87
Current duties	26	6.92	5.05	0.53**	21	11.24	4.31	−0.19	67	10.73	6.50	−0.10	365	10.11	6.05
Supervisor	26	0.46	0.51	0.24	21	0.48	0.51	0.21	67	0.54	0.50	0.09	365	0.58	0.49
All jobs	26	3.85	3.13	0.27	21	3.62	1.69	0.36*	67	4.27	2.35	0.13	365	4.61	2.77
All duties	26	21.46	18.47	0.32	21	21.67	10.77	0.31	67	28.06	16.52	−0.12	365	26.26	14.93
Objectives	26	42.92	16.88	0.22	21	39.14	18.19	0.50	67	47.30	16.04	−0.12	365	45.76	12.85
Interviewer comments	26	17.23	11.55	0.25	21	17.24	11.52	0.24	67	22.85	13.14	−0.22†	365	20.20	12.15
Letter of reference	26	41.62	22.07	0.19	21	39.48	18.21	0.30	67	43.00	18.93	0.13	365	45.75	21.36
							Nonflying jobs								
Mental ability scores															
Verbal	87	48.29	24.72	0.63**	120	46.30	21.97	0.72**	192	52.40	23.85	0.46**	898	63.23	23.75
Quantitative	87	54.44	28.19	−0.42**	120	33.64	17.98	0.46**	192	38.13	22.09	0.26**	898	44.23	23.86
Numeric applicant information															
GPA of most recent degree	79	3.40	0.43	0.34**	108	3.41	0.40	0.33**	166	3.46	0.44	0.20*	823	3.54	0.41
Number of jobs	85	4.48	1.83	0.21*	120	4.94	1.92	−0.02	189	4.95	1.91	−0.02	893	4.91	2.04
Text scores															
Career achievements	87	5.43	3.70	0.25*	120	6.42	3.76	−0.02	192	6.08	3.53	0.08	898	6.36	3.74
Personal achievements	87	5.97	4.24	0.01	120	5.91	3.78	0.03	192	6.26	4.69	−0.05	898	6.03	4.57
Professional affiliations	87	3.69	3.90	0.03	120	4.80	3.57	−0.31**	192	3.90	3.19	−0.03	898	3.80	3.21
Personal interests	87	6.36	3.05	−0.02	120	5.71	3.29	0.18*	192	5.96	2.82	0.10	898	6.29	3.34
Current duties	87	6.66	4.41	0.06	120	6.33	3.67	0.14	192	7.26	4.21	−0.09	898	6.89	4.08
Supervisor	87	0.55	0.50	0.20*	120	0.63	0.49	0.05	192	0.70	0.46	−0.11	898	0.65	0.48
All jobs	87	3.51	2.21	0.12	120	3.64	2.12	0.05	192	3.76	2.13	−0.00	898	3.75	2.07
All duties	87	28.03	15.60	0.05	120	27.62	12.29	0.08	192	30.03	15.12	−0.10	898	28.67	14.06
Objectives	87	40.79	12.69	0.15	120	43.28	11.98	−0.07	192	44.23	11.73	−0.16*	898	42.47	11.26
Interviewer comments	87	18.08	12.10	0.42**	120	24.64	11.20	−0.15	192	25.05	11.76	−0.18*	898	22.94	11.60
Letter of reference	87	45.68	23.53	−0.19†	120	37.67	14.44	0.19*	192	39.50	18.41	0.10	898	41.56	21.56

$N = 368$–432. $^{*}p < .05$. $^{**}p < .01$. $^{\dagger}p = .05$, one-tailed. A positive d means a higher mean for Whites. GPA = grade point average.
$N = 902$–1,090. $^{*}p < .05$. $^{**}p < .01$. $^{\dagger}p = .05$, one-tailed.

Table 11
Reductions in Race Differences Predicting Board Scores

				Asian			White				
Model	Mental ability scores	Numeric application information	Text scores	*N*	*M*	*SD*	*N*	*M*	*SD*	*d*	*t* test
			Flying jobs								
I	✓			26	−0.10	0.47	365	0.05	0.45	0.33	1.80*
II	✓		✓	26	−0.15	0.70	365	0.05	0.52	0.39	
III		✓		21	0.05	0.28	348	0.06	0.33	0.01	0.66
IV		✓	✓	21	0.04	0.45	348	0.05	0.40	0.03	
V	✓	✓		21	0.04	0.52	348	0.10	0.53	0.11	0.05
VI	✓	✓	✓	21	0.03	0.64	348	0.09	0.56	0.11	

				Black			White				
Model	Mental ability scores	Numeric application information	Text scores	*N*	*M*	*SD*	*N*	*M*	*SD*	*d*	*t* test
			Flying jobs								
I	✓			21	−0.31	0.38	365	0.05	0.45	0.79**	−1.50
II	✓		✓	21	−0.33	0.50	365	0.05	0.52	0.74**	
III		✓		20	−0.10	0.29	348	0.06	0.33	0.47*	1.13
IV		✓	✓	20	−0.15	0.48	348	0.05	0.40	0.50*	
V	✓	✓		20	−0.38	0.44	348	0.10	0.53	0.90**	−1.20
VI	✓	✓	✓	20	−0.38	0.52	348	0.09	0.56	0.85**	

				Hispanic			White				
Model	Mental ability scores	Numeric application information	Text scores	*N*	*M*	*SD*	*N*	*M*	*SD*	*d*	*t* test
			Flying jobs								
I	✓			67	−0.14	0.48	365	0.05	0.45	0.42**	−1.20
II	✓		✓	67	−0.15	0.61	365	0.05	0.52	0.38**	
III		✓		58	0.02	0.33	348	0.06	0.33	0.11	−2.15*
IV		✓	✓	58	0.03	0.44	348	0.05	0.40	0.05	
V	✓	✓		58	−0.08	0.55	348	0.10	0.53	0.33*	−0.93
VI	✓	✓	✓	58	−0.07	0.60	348	0.09	0.56	0.30*	

				Asian			White				
Model	Mental ability scores	Numeric application information	Text scores	*N*	*M*	*SD*	*N*	*M*	*SD*	*d*	*t* test
			Nonflying jobs								
I	✓			87	−0.06	0.21	898	0.03	0.20	0.50**	−8.90**
II	✓		✓	87	−0.10	0.35	898	0.04	0.36	0.38**	
III		✓		78	−0.06	0.34	823	0.05	0.33	0.34**	−1.25
IV		✓	✓	78	−0.07	0.40	823	0.06	0.41	0.32**	
V	✓	✓		78	−0.08	0.35	823	0.08	0.38	0.43**	−2.30*
VI	✓	✓	✓	78	−0.09	0.41	823	0.08	0.45	0.39**	

				Black			White				
Model	Mental ability scores	Numeric application information	Text scores	*N*	*M*	*SD*	*N*	*M*	*SD*	*d*	*t* test
			Nonflying jobs								
I	✓			120	−0.11	0.18	898	0.03	0.20	0.75**	−23.62**
II	✓		✓	120	−0.12	0.37	898	0.04	0.36	0.44**	
III		✓		108	−0.05	0.33	823	0.05	0.33	0.31**	−0.57
IV		✓	✓	108	−0.06	0.42	823	0.06	0.41	0.30**	
V	✓	✓		108	−0.15	0.35	823	0.08	0.38	0.63**	−4.10**
VI	✓	✓	✓	108	−0.16	0.46	823	0.08	0.45	0.55**	

				Hispanic			White				
Model	Mental ability scores	Numeric application information	Text scores	*N*	*M*	*SD*	*N*	*M*	*SD*	*d*	*t* test
			Nonflying jobs								
I	✓			192	−0.06	0.20	898	0.03	0.20	0.47**	−11.77**
II	✓		✓	192	−0.07	0.37	898	0.04	0.36	0.32**	
III		✓		166	−0.01	0.35	823	0.05	0.33	0.19*	0.18
IV		✓	✓	166	−0.02	0.42	823	0.06	0.41	0.19*	
V	✓	✓		166	−0.06	0.40	823	0.08	0.38	0.37**	−2.27*
VI	✓	✓	✓	166	−0.06	0.46	823	0.08	0.45	0.33**	

N = 434–461. *$p < .05$. **$p < .01$, one-tailed.
N = 902–1,090. *$p < .05$. **$p < .01$, one-tailed.

Table 12
Adverse Impact Ratios at Hypothetical Selection Ratios

	Number passed at 75% selection ratio		Number passed at 50% selection ratio		Number passed at 25% selection ratio		Total samples		Adverse impact at selection ratios		
Model	Asian	White	Asian	White	Asian	White	Asian	White	75%	50%	25%
					Flying jobs						
I	16	291	12	201	7	101	26	365	0.77	0.84	0.97
II	16	290	14	192	7	102	26	365	0.77	1.02	0.96
III	18	266	11	183	3	96	21	348	1.12	1.00	0.52
IV	15	269	11	181	7	88	21	348	0.92	1.01	1.32
V	14	275	10	194	7	96	21	348	0.84	0.85	1.21
VI	14	275	11	187	7	93	21	348	0.84	0.97	1.25

	Number passed at 75% selection ratio		Number passed at 50% selection ratio		Number passed at 25% selection ratio		Total samples		Adverse impact at selection ratios		
Model	Black	White	Black	White	Black	White	Black	White	75%	50%	25%
					Flying jobs						
I	10	291	5	201	1	101	21	365	0.60	0.43	0.17
II	10	290	4	192	2	102	21	365	0.60	0.36	0.34
III	12	266	6	183	2	96	20	348	0.78	0.57	0.36
IV	12	269	7	181	2	88	20	348	0.78	0.67	0.40
V	10	275	4	194	1	96	20	348	0.63	0.36	0.18
VI	11	275	5	187	0	93	20	348	0.70	0.47	0.00

	Number passed at 75% selection ratio		Number passed at 50% selection ratio		Number passed at 25% selection ratio		Total samples		Adverse impact at selection ratios		
Model	Hispanic	White	Hispanic	White	Hispanic	White	Hispanic	White	75%	50%	25%
					Flying jobs						
I	43	291	20	201	12	101	67	365	0.80	0.54	0.65
II	44	290	29	192	9	102	67	365	0.83	0.82	0.48
III	40	266	27	183	14	96	58	348	0.90	0.89	0.88
IV	42	269	27	181	15	88	58	348	0.94	0.90	1.02
V	39	275	18	194	10	96	58	348	0.85	0.56	0.63
VI	39	275	24	187	12	93	58	348	0.85	0.77	0.77

	Number passed at 75% selection ratio		Number passed at 50% selection ratio		Number passed at 25% selection ratio		Total samples		Adverse impact at selection ratios		
Model	Asian	White	Asian	White	Asian	White	Asian	White	75%	50%	25%
					Nonflying jobs						
I	50	732	32	506	14	271	87	898	0.71	0.65	0.53
II	57	710	37	495	16	251	87	898	0.83	0.77	0.66
III	48	593	27	414	12	210	78	823	0.85	0.69	0.60
IV	46	600	27	407	11	206	78	823	0.81	0.70	0.56
V	51	610	20	440	10	228	78	823	0.88	0.48	0.46
VI	46	618	29	420	12	220	78	823	0.79	0.73	0.58

	Number passed at 75% selection ratio		Number passed at 50% selection ratio		Number passed at 25% selection ratio		Total samples		Adverse impact at selection ratios		
Model	Black	White	Black	White	Black	White	Black	White	75%	50%	25%
					Nonflying jobs						
I	66	732	36	506	10	271	120	898	0.67	0.53	0.28
II	74	710	38	495	17	251	120	898	0.78	0.57	0.51
III	67	593	35	414	16	210	108	823	0.86	0.64	0.58
IV	64	600	38	407	20	206	108	823	0.81	0.71	0.74
V	57	610	28	440	11	228	108	823	0.71	0.48	0.37
VI	53	618	32	420	11	220	108	823	0.65	0.58	0.38

(*table continues*)

Table 12 (*continued*)

Model	Number passed at 75% selection ratio		Number passed at 50% selection ratio		Number passed at 25% selection ratio		Total samples		Adverse impact at selection ratios		
	Hispanic	White	Hispanic	White	Hispanic	White	Hispanic	White	75%	50%	25%
					Nonflying jobs						
I	126	732	71	506	29	271	192	898	0.81	0.66	0.50
II	135	710	82	495	39	251	192	898	0.89	0.77	0.73
III	113	593	72	414	37	210	166	823	0.94	0.86	0.87
IV	109	600	74	407	37	206	166	823	0.90	0.90	0.89
V	104	610	59	440	25	228	166	823	0.85	0.66	0.54
VI	103	618	64	420	30	220	166	823	0.83	0.76	0.68

four samples of professional employees, we found that NLP scores (a) predict human judgment at the level of the correlation between human raters, (b) add incremental validity beyond mental ability tests and numeric application information when predicting human ratings (Board scores) and subsequent job (training) performance, and (c) reduce racial subgroup mean differences and potentially adverse impact.

Theoretical Contributions

Findings from this study contribute to the literature in three ways. First, building on the work and life experiences literatures, we show that NLP scoring can quantify more job-related information on candidates than typical selection procedures, such as mental ability tests. We found through our analysis of the construct domains within the text data that this method of scoring enables researchers to complement the constructs they are measuring with common procedures (e.g., mental ability) by using automated scoring. Further, we can systematically evaluate applicant material equally and more reliably for all candidates.

Our second contribution is to illustrate how NLP offers an efficient and relatively inexpensive option for scoring the large quantities of applicant text data that have historically been too resource intensive for high-volume hiring contests. NLP helps resolve the problem of insufficiently capturing the breadth and depth of work and life experiences in application data due to the limitations of human and financial resources. This research may also advance the work and life experiences scholarship by providing a scoring method for basic research on the topic.

Finally, we contribute to the literature by offering validity evidence of NLP scores of application information for selection decisions and subsequent performance. We use data from an operational context to demonstrate how NLP scores share notable validity with Board scores and subsequent training performance after hiring, and show incremental validity beyond mental ability tests, numeric application information, and even Board scores. This is valuable because mental ability tests are one of the strongest predictors of performance, and numeric application information also has a history of validity, so showing that scoring narrative application data matters beyond them is a meaningful empirical and theoretical contribution. We also illustrate how NLP scores have smaller subgroup differences by race and can reduce subgroup differences when combined with these other constructs. We then show how this might translate into reduced adverse impact based on realistic hypothetical selection rates. This research demonstrates the construct-change approach to reducing subgroup differences in selection (Arthur et al., 2021). Aside from measuring additional constructs, such reductions may in part be because text scores are not subjected to human evaluation and the potential for intentional discrimination.

Practical Implications

This research has several obvious practical implications. First, HR professionals should incorporate NLP into their selection systems. This could also save cost and increase the speed of hiring decisions. For example, in this study, the total estimated costs of selection decisions ranges from $169,334.40 to $310,382.40 annually. This is based on 12 Board members (three members for each of the four panels), spending 40 hr on four boards per year at $43.32 per hour ($83,174.40), including travel for 1 week at a $185 per diem ($62,160) and $500 in airfare four times per year ($24,000; Total = $169,334.40). The upper bound is based on 60 hr (vs. 40) if there are six boards per year. Moreover, these estimates do not include the cost of Board members being away from their primary duties. Boards comprise Colonels and Lt. Colonels (middle- to upper level management) who control critical programs. With the prediction of the model as high as the correlation between Board members, the organization could save one-third of the annual cost by replacing one Board member ($56,444.80–$103,460.80) and perhaps also improve the speed of decisions.

Second, NLP models could be used to efficiently produce a practice application, which would provide feedback to candidates in the form of an estimated score on the actual Boards. This would not require candidates to wait until the next once-a-year hiring announcement to receive feedback. Instead, they could improve the description of their credentials (e.g., past job duties, achievements, statements of objectives) to reduce deficiencies and other sources of systematic error. These benefits would be in addition to those documented in research on practice employment tests, such as encouraging qualified candidates to apply, encouraging unqualified candidates to seek skill development, and reducing subgroup differences (M. C. Campion et al., 2019). In the same vein, NLP could enhance transparency of the evaluation of application information. Even if hiring officials are trained and their evaluations are guided by standardized procedures, there is still a subjective component, especially from the perspective of unsuccessful candidates. NLP may be easier to defend if legally challenged, just like structure has done for employment interviews (Gollub-Williamson et al., 1997). Moreover, NLP would allow for continuous improvement of the application review process because it standardizes procedures and makes them more measurable. Continuous improvement requires that random variation be reduced so that improvements can be identified (e.g., Bhuiyan & Baghel, 2005).

An important caveat in developing NLP models is the avoidance of bias. Some potential biases are obvious (e.g., stereotyping), but others may be hidden. For example, the school a candidate attended might reflect privilege in addition to differences in knowledge reflected by the degree and major. Another potential issue is whether NLP scores may actually show greater subgroup differences than the current selection process if the text scores retained in the model are based on those that predict past selection decisions. It is also possible that NLP could measure attributes more reliably than human raters, thus increasing the measurement of subgroup differences that exist in the narrative information. Moreover, some proposed approaches to reducing subgroup differences will reduce validity, such as (a) eliminating text categories that show subgroups differences, as the similar process of removing test items has shown (e.g., Ployhart & Holtz, 2008), (b) developing algorithms that statistically reduce subgroup differences, which will necessarily cause prediction bias (N. Zhang et al., 2023), and could also be similar to within-group norming that is prohibited by the Civil Rights Act, 1991, and (c) reducing the weights of the most valid text categories, as might occur with some statistical (pareto optimal) weighting schemes (e.g., Potosky et al., 2008).

A potential practical concern is whether candidates can "game" the system and fake their responses. We believe it is highly improbable that candidates will be more likely or able to fake responses where NLP rather than humans is used to score for several reasons. First, how applicant data are collected will look the same to candidates whether scored by an algorithm or a person. Second, it is unlikely that candidates would know the variables in the model and how the algorithm scores them. There has been some research on this topic, and it generally shows that "gaming" the system is quite difficult. For example, Powers et al. (2002) sought to answer this question by offering very detailed information on the computer models, including the "particular cue words on which it focuses" (p. 108), to testing experts, computer scoring experts, researchers, and critics of computer scoring and had them generate responses to be scored. Powers et al. found that they were only able to earn a higher score a slight majority of the time. Third, resumes and similar work experience documentation are objective information difficult to fake unless someone commits resume fraud (actually lying). In such a case, the algorithm would be as vulnerable as a human scorer. We can assume all candidates will try to provide the most positive information but lying on applications is less likely than giving positive answers to personality tests or other assessments where candidates essentially score themselves. Fourth, candidates will not know whether a human or an algorithm scores the materials. An organization would probably not advertise this fact, just like they would not reveal other technical details of their assessments. Even if it is scored by an algorithm, a human will likely still review the material such as in instances, where the algorithm replaces a single human rater on a board, or when interviewers read the application to prepare for an interview. We conclude that while there is some evidence that computer models can be "gamed" under the right conditions, we believe it is unlikely. Nevertheless, additional scholarship is needed to understand this more clearly.

Limitations and Future Research

First, this study was conducted within one organization. While this controls for many exogenous influences, future research should replicate our study across a broader array of occupations. Although Air Force Officers are likely to have many similarities to private and other public sector jobs in terms of skill and ability requirements, the generalizability of findings cannot be assumed. While we argue NLP is broadly beneficial to selection because organizations tend to collect a large amount of text data from applicants, parameters of our current model may limit its use in alternative contexts. Also, our sample was fairly typical for selection, but it was still relatively small. The sample size requirements for NLP will depend on the amount of text collected per text field from each sample member, with more text yielding more stable text scores and thus smaller sample size requirements. It will also depend on the diversity of information across sample members, with greater diversity requiring larger samples to create comprehensive models. However, the models developed from samples as small as 200 may have value, as illustrated by the present study. Further, while we showed that subgroup differences shrank with the inclusion of NLP scores, the influence of NLP scores on adverse impact will depend as much or more so on features of the hiring system and the specific candidates. For example, the influence of NLP scores may depend on how the scores are used (e.g., to augment or replace other selection procedures), the impact of those other procedures, and the cutting scores used. Finally, a value-add of NLP is that text models can be developed specifically for each selection context, but future research might develop generic models that could be used across contexts.

Second, our main comparisons in the present study involved mental ability tests and numeric application information. The advantages of NLP may differ when compared to other procedures. The employment interview is one obvious comparison due to its reliance on narrative data. NLP scores might not be able to reduce subgroup differences because the interview usually does not exhibit large subgroup differences, but it may have the advantage of scoring interview information well, thus saving costs. Moreover, this approach to scoring application data leaves some uncertainty as to the constructs being measured. While we imposed an organizing framework from Speer et al.'s (2022) and Huffcutt et al.'s (2001) meta-analyses on our text variables, future research may seek to uncover the exact constructs assessed. Further, to address concerns as to the precise operation of proprietary software such as SPSS versus open-source languages such as Python, researchers using proprietary software should compare their results to open-source methods to assess convergence. To examine this here, we ran a bag-of-words model in Python on one text field (career achievements) and found that the relationship between the scores was .88 for flying jobs and .84 for nonflying jobs, suggesting equivalence. Finally, future researchers might evaluate whether NLP can improve on well-developed empirically keyed biodata. Although NLP is likely to add value by improving the scoring of text information, biodata has a history of validity (Speer et al., 2022).

Third, NLP scores used here were operationalized as counts based on all the categories identified with at least a minimal frequency across respondents without weighting because sample sizes did not allow cross-validation. Although the positive results are encouraging and show that NLP can be used even in small samples, future research should examine the potential of improved prediction that might be possible with differential weighting. Another related issue is negative weighting. We found that while some text variables show a negative bivariate correlation with criteria, they tend to be few in number and small in magnitude and giving them a negative weight does not improve prediction and can fail to cross-validate. We think this is because candidates likely do not write about negative indicators of success, so these relationships are mostly due to chance. However, it is possible that in an alternative context where individuals are not motivated to impress hiring officials, negative predictors of the criterion may be meaningful. Nevertheless, negative weighting might also reduce response bias, similar to reverse-scored Likert items, but psychometric properties should be examined (e.g., Schriesheim & Eisenbach, 1995).

Fourth, while we propose that we captured knowledge, skills, and abilities in narrative application data, we believe additional research can be done particularly as it relates to "O's." For example, NLP may be a superior way to measure personality from narrative data given the concerns with faking on personality tests (Morgeson et al., 2007). Because applicants will not know the exact content of the NLP model, they would not be able to tailor their narrations. Nevertheless, candidates may still exaggerate and commit other forms of faking (Roulin & Krings, 2020), and they may be able to guess what information will sound more desirable on an application given the job requirements. As such, response distortion is a potentially fruitful area of future research.

Finally, although we focused on criterion-related validity, future research could determine how to content validate NLP models so as to provide another type of validation evidence. For example, perhaps the text categories could be evaluated by subject matter experts on job relatedness and needed-at-entry or linked to job tasks and other information identified through job analyses, as suggested by professional testing guidelines (Society for Industrial & Organizational Psychology, 2018). Moreover, as a statistical approach to the content analysis of hiring information, NLP may allow quantifiable content-related validation metrics.

References

Allen, J. L. (October 2020). *Office training school (syllabus)*. United States Air Force.

Anderson, N. (2018, October 23). SAT reclaims title of most widely used college admission test. *The Washington Post*. https://www.washingtonpost.com/education/2018/10/23/sat-reclaims-title-most-widely-used-college-admission-test/

Arthur, W., Jr., Doverspike, D., Barrett, G. V., & Miguel, R. (2013). Chasing the Title VII Holy Grail: The pitfalls of guaranteeing adverse impact elimination. *Journal of Business and Psychology*, *28*(4), 473–485. https://doi.org/10.1007/s10869-013-9289-6

Arthur, W., Jr., Edwards, B. D., & Barrett, G. V. (2002). Multiple-choice and constructed response tests of ability: Race-based subgroup performance differences on alternative paper-and-pencil test formats. *Personnel Psychology*, *55*(4), 985–1008. https://doi.org/10.1111/j.1744-6570.2002.tb00138.x

Arthur, W., Jr., Keiser, N. L., Atoba, O. A., Cho, I., & Edwards, B. D. (2021). Does the use of alternative predictor methods reduce subgroup differences? It depends on the construct. *Human Resource Management*, *60*(4), 479–498. https://doi.org/10.1002/hrm.22027

Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, *93*(2), 435–442. https://doi.org/10.1037/0021-9010.93.2.435

Arthur, W., Jr., & Woehr, D. J. (2013). No steps forward, two steps back: The fallacy of trying to "eradicate" adverse impact? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *6*(4), 438–442. https://doi.org/10.1111/iops.12081

Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, *30*(1), 125–141. https://doi.org/10.1177/0265532212452396

Banks, G. C., Woznyj, H. M., Wesslen, R. S., Frear, K. A., Berka, G., Heggestad, E. D., & Gordon, H. L. (2019). Strategic recruitment across borders: An investigation of multinational enterprises. *Journal of Management*, *45*(2), 476–509. https://doi.org/10.1177/0149206318764295

Bhuiyan, N., & Baghel, A. (2005). An overview of continuous improvement: From the past to the present. *Management Decision*, *43*(5), 761–771. https://doi.org/10.1108/00251740510597761

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://dl.acm.org/doi/10.5555/944919.944937

Bligh, M. C., Kohles, J. C., & Meindl, J. R. (2004). Charisma under crisis: Presidential leadership, rhetoric, and media responses before and after the September 11th terrorist attacks. *The Leadership Quarterly*, *15*(2), 211–239. https://doi.org/10.1016/j.leaqua.2004.02.005

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology*, *52*(3), 561–589. https://doi.org/10.1111/j.1744-6570.1999.tb00172.x

Brown, B. K., & Campion, M. A. (1994). Biodata phenomenology: Recruiters' perceptions and use of biographical information in resume screening. *Journal of Applied Psychology*, *76*(6), 897–908. https://doi.org/10.1037/0021-9010.79.6.897

Campion, E. D., & Campion, M. A. (2020). Using computer-assisted text analysis (CATA) to inform employment decisions: Approaches, software, and findings. In M. R. Buckley, A. R. Wheeler, J. E. Baur, & J. R. B. Halbesleben (Eds.), *Research in personnel and human resources management* (Vol. 38, pp. 285–325). Emerald Publishing Limited. https://www.emerald.com/insight/content/doi/10.1108/S0742-730120200000038010/full/html

Campion, M. C., Campion, E. D., & Campion, M. A. (2019). Using practice employment tests to improve recruitment and personnel selection outcomes for organizations and job seekers. *Journal of Applied Psychology*, *104*(9), 1089–1102. https://doi.org/10.1037/apl0000401

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. https://doi.org/10.1037/apl0000108

Carretta, T. R. (2010). Air Force Officer Qualifying Test validity for non-rated officer specialties. *Military Psychology*, *22*, 450–464. https://doi.org/10.1080/08995605.2010.513261

Carretta, T. R. (2011). Pilot candidate selection method: Still an effective predictor of US Air Force pilot training performance. *Aviation Psychology and Applied Human Factors*, *1*(1), 3–8. https://doi.org/10.1027/2192-0923/a00002

Carretta, T. R., Rose, M. R., & Trent, J. D. (2016). *Air Force Officer Qualifying Test form T: Initial item-, test-, factor-, and composite-level analyses* (AFRL-RH-WP-TR-2016-0093). Wright-Patterson Air Force Base, Air Force Research Laboratory, Human Performance Wing, Airman Systems Directorate.

Cerda, P., Varoquaux, G., & Kegl, B. (2018). Similarity encoding for learning with dirty categorical variables. *Machine Learning*, *107*(8–10), 1477–1494. https://doi.org/10.1007/s10994-018-5724-2

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology*, *82*(1), 143–159. https://doi.org/10.1037/0021-9010.82.1.143

Civil Rights Act, 42 USCS § 1981a (1991).

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validation of situational judgment tests. *Journal of Applied Psychology*, *86*(3), 410–417. https://doi.org/10.1037/0021-9010.86.3.410

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology*, *80*(5), 565–579. https://doi.org/10.1037/0021-9010.80.5.565

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, *13*(1), 3–21. https://doi.org/10.1007/BF00988593

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, L. K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of NAACL-HLT* (pp. 4171–4186). Association for Computational Linguistics. https://www.aclweb.org/anthology/N19-1423.pdf

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment*, *5*(1), 3–35. https://ejournals.bc.edu/index.php/jtla/article/view/1640

Dragoni, L., Oh, I. S., Vankatwyk, P., & Tesluk, P. E. (2011). Developing executive leaders: The relative contribution of cognitive ability, personality, and the accumulation of work experience in predicting strategic thinking competency. *Personnel Psychology*, *64*(4), 829–864. https://doi.org/10.1111/j.1744-6570.2011.01229.x

Drasgow, F., Nye, C. D., Carretta, T. R., & Ree, M. J. (2010). Factor structure of the Air Force Officer Qualifying Test form S: Analysis and comparison with previous forms. *Military Psychology*, *22*(1), 68–85. https://doi.org/10.1080/08995600903249255

Edwards, B. D., & Arthur, W., Jr. (2007). An examination of factors contributing to a reduction in subgroup differences on a constructed-response paper-and-pencil test of scholastic achievement. *Journal of Applied Psychology*, *92*(3), 794–801. https://doi.org/10.1037/0021-9010.92.3.794

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., Hagan, C. A., Tobolsky, V. A., Smith, L. K., Buffone, A., Iwry, J., Seligman, M. E. P., & Ungar, L. H. (2021). Closed- and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods*, *26*(4), 398–427. https://doi.org/10.1037/met0000349

Eisenstein, J. (2019). *Introduction to natural language processing*. The MIT Press.

Foldes, H. J., Duehr, E. E., & Ones, D. S. (2008). Group differences in personality: Meta-analyses comparing five US racial groups. *Personnel Psychology*, *61*(3), 579–616. https://doi.org/10.1111/j.1744-6570.2008.00123.x

Ghiselli, E. E. (1964). *Theory of psychological measurement*. McGraw-Hill.

Gollub-Williamson, L. G., Campion, J. E., Malos, S. B., Roehling, M. V., & Campion, M. A. (1997). Employment interview on trial: Linking interview structure with litigation outcomes. *Journal of Applied Psychology*, *82*(6), 900–912. https://doi.org/10.1037/0021-9010.82.6.900

Hannigan, T. R., Haans, R. F., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., Kaplan, S., & Jennings, P. D. (2019). Topic modeling in management research: Rendering new theory from textual data. *The Academy of Management Annals*, *13*(2), 586–632. https://doi.org/10.5465/annals.2017.0099

Helms, W. S., Oliver, C., & Webb, K. (2012). Antecedents of settlement on a new institutional practice: Negotiation of the ISO 26000 standard on social responsibility. *Academy of Management Journal*, *55*(5), 1120–1145. https://doi.org/10.5465/amj.2010.1045

Hough, L. M. (1984). Development and evaluation of the "accomplishment record" method of selecting and promoting professionals. *Journal of Applied Psychology*, *69*(1), 135–146. https://doi.org/10.1037/0021-9010.69.1.135

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment*, *9*(1–2), 152–194. https://doi.org/10.1111/1468-2389.00171

Howard, A. (1986). College experiences and managerial performance. *Journal of Applied Psychology*, *71*(3), 530–552. https://doi.org/10.1037/0021-9010.71.3.530

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology*, *86*(5), 897–913. https://doi.org/10.1037/0021-9010.86.5.897

IBM. (2019). *IBM SPSS modeler text analytics 18.2.1 user's guide*. https://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.2.1/en/ModelerTextAnalytics.pdf

Jurafsky, D., & Martin, J. H. (2021). *Speech and language processing* (3rd ed.). Pearson. https://web.stanford.edu/~jurafsky/slp3/6.pdf

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Lievens, F., Sackett, P. R., Dahlke, J. A., Oostrom, J. K., & De Soete, B. (2019). Constructed response formats and their effects on minority-majority differences and validity. *Journal of Applied Psychology*, *104*(5), 715–726. https://doi.org/10.1037/apl0000367

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBerta: A robustly optimized BERT pretraining approach*. arXiv. https://doi.org/10.48550/arXiv.1907.11692

Madera, J. M., Hebl, M. R., & Martin, R. C. (2009). Gender and letters of recommendation for academia: Agentic and communal differences. *Journal of Applied Psychology*, *94*(6), 1591–1599. https://doi.org/10.1037/a0016539

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, *60*(3), 683–729. https://doi.org/10.1111/j.1744-6570.2007.00089.x

Morris, R. (1994). Computerized content analysis in management research: A demonstration of advantages & limitations. *Journal of Management*, *20*(4), 903–931. https://doi.org/10.1177/014920639402000410

Mumford, M. D., & Stokes, G. S. (1992). Developmental determinants of individual action: Theory and practice in applying background measures. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 3, pp. 61–138). Consulting Psychologists Press.

Olea, M. M., & Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than g. *Journal of Applied Psychology*, *79*(6), 845–851. https://doi.org/10.1037/0021-9010.79.6.845

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology*, *89*(2), 187–207. https://doi.org/10.1037/0021-9010.89.2.187

Outtz, J. L. (Ed.). (2010). *Adverse impact: Implications for organizational staffing and high stakes selection*. Routledge/Taylor & Francis Group. https://doi.org/10.4324/9780203848418

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. University of Texas at Austin.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, *61*(1), 153–172. https://doi.org/10.1111/j.1744-6570.2008.00109.x

Potosky, D., Bobko, P., & Roth, P. L. (2005). Forming composites of cognitive ability and alternative measures to predict job performance and reduce adverse impact: Corrected estimates and realistic expectations. *International Journal of Selection and Assessment*, *13*(4), 304–315. https://doi.org/10.1111/j.1468-2389.2005.00327.x

Potosky, D., Bobko, P., & Roth, P. L. (2008). Some Comments on Pareto Thinking, Test Validity, and Adverse Impact: When "and" is optimal and "or" is a trade-off. *International Journal of Selection and Assessment*, *16*(3), 201–205. https://doi.org/10.1111/j.1468-2389.2008.00425.x

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping E-Rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior*, *18*(2), 103–134. https://doi.org/10.1016/S0747-5632(01)00052-8

Quińones, M. A., Ford, J. K., & Teachout, M. S. (1995). The relationship between work experience and job performance: A conceptual and meta-analytic review. *Personnel Psychology*, *48*(4), 887–910. https://doi.org/10.1111/j.1744-6570.1995.tb01785.x

Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, *18*(1), 25–39. https://doi.org/10.1016/j.asw.2012.10.004

Ramos, J. (2003, December). *Using tf-idf to determine word relevance in document queries*. Department of Computer Science, Rutgers University. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b3bf6373ff41a115197cb5b30e57830c16130c2c

Roberts, H. E., & Skinner, J. (1996). Gender and racial equity of the Air Force Officer Qualifying Test in officer training school selection decisions. *Military Psychology*, *8*(2), 95–113. https://doi.org/10.1207/s15327876mp0802_4

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., III, & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology*, *54*(2), 297–330. https://doi.org/10.1111/j.1744-6570.2001.tb00094.x

Roth, P. L., Van Iddekinge, C. H., DeOrtentiis, P. S., Hackney, K. J., Zhang, L., & Buster, M. A. (2017). Hispanic and Asian performance on selection procedures: A narrative and meta-analytic review of 12 common predictors. *Journal of Applied Psychology*, *102*(8), 1178–1202. https://doi.org/10.1037/apl0000195

Roulin, N., & Krings, F. (2020). Faking to fit in: Applicants' response strategies to match organizational culture. *Journal of Applied Psychology*, *105*(2), 130–145. https://doi.org/10.1037/apl0000431

Ruderman, M. N., Ohlott, P. J., Panzer, K., & King, S. N. (2002). Benefits of multiple roles for managerial women. *Academy of Management Journal*, *45*(2), 369–386. https://doi.org/10.2307/3069352

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology*, *50*(3), 707–721. https://doi.org/10.1111/j.1744-6570.1997.tb00711.x

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*, *107*(11), 2040–2068. https://doi.org/10.1037/apl0000994

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology*, *104*(10), 1207–1225. https://doi.org/10.1037/apl0000405

Salgado, J. F., & Moscoso, S. (2002). Comprehensive meta-analysis of the construct validity of the employment interview. *European Journal of Work and Organizational Psychology*, *11*(3), 299–324. https://doi.org/10.1080/13594320244000184

Schmidt, F., & Hunter, J. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Schriesheim, C. A., & Eisenbach, R. J. (1995). An exploratory and confirmatory factor-analytic investigation of item wording effects on the obtained factor structures of survey questionnaire measures. *Journal of Management*, *21*(6), 1177–1193. https://doi.org/10.1177/014920639502100609

Short, J. C., McKenny, A. F., & Reid, S. W. (2018). More than words? Computer-aided text analysis in organizational behavior and psychology research. *Annual Review of Organizational Psychology and Organizational Behavior*, *5*(1), 415–435. https://doi.org/10.1146/annurev-orgpsych-032117-104622

Society for Industrial and Organizational Psychology. (2018). Principles for the validation and use of personnel selection procedures. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *11*(S1), 1–97. https://doi.org/10.1017/iop.2018.195

Speer, A. B. (2018). Quantifying with words: An investigation of the validity of narrative-derived performance scores. *Personnel Psychology*, *71*(3), 299–333. https://doi.org/10.1111/peps.12263

Speer, A. B. (2021). Scoring dimension-level job performance from narrative comments: Validity and generalizability when using natural language processing. *Organizational Research Methods*, *24*(3), 572–594. https://doi.org/10.1177/1094428120930815

Speer, A. B., Tenbrink, A. P., Wegmeyer, L. J., Sendra, C. C., Shihadeh, M., & Kaur, S. (2022). Meta-analysis of biodata in employment settings: Providing clarity to criterion and construct-related validity estimates. *Journal of Applied Psychology*, *107*(10), 1678–1705. https://doi.org/10.1037/apl0000964

Sriurai, W., Meesad, P., & Haruechaiyasak, C. (2010). *Hierarchical web page classification based on a topic model and neighboring pages integration*. arXiv. https://doi.org/10.48550/arXiv.1003.1510

Tesluk, P. E., & Jacobs, R. R. (1998). Toward an integrated model of work experience. *Personnel Psychology*, *51*(2), 321–355. https://doi.org/10.1111/j.1744-6570.1998.tb00728.x

Uniform Guidelines on Employee Selection Procedures. (1978). Employee selection procedures—Adoption by four agencies of uniform guidelines–1978. *Federal Register*, *43*(166), 38290–38315. https://www.govinfo.gov/content/pkg/CFR-2011-title29-vol4/xml/CFR-2011-title29-vol4-part1607.xml

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education*, *2*(1), 319–330. https://doi.org/10.28945/331

Van Iddekinge, C. H., Arnold, J. D., Frieder, R. E., & Roth, P. L. (2019). A meta-analysis of the criterion-related validity of pre-hire work experience. *Personnel Psychology*, *72*(4), 571–598. https://doi.org/10.1111/peps.12335

Walker, D. D., van Jaarsveld, D. D., & Skarlicki, D. P. (2017). Sticks and stones can break my bones but words can also hurt me: The relationship between customer verbal aggression and employee incivility. *Journal of Applied Psychology*, *102*(2), 163–179. https://doi.org/10.1037/apl0000170

Wasko, L. E., Putka, D. J., Legree, P. J., & Kilcullen, R. N. (2019). *Validation of measures for predicting leader development and assessment course performance* (Technical Report No. 1375). U.S. Army Research Institute for the Behavioral and Social Sciences. https://apps.dtic.mil/dtic/tr/fulltext/u2/1080161.pdf

Wolfe, R. A., Gephart, R. P., & Johnson, T. E. (1993). Computer-facilitated qualitative data analysis: Potential contributions to management research. *Journal of Management*, *19*(3), 637–660. https://doi.org/10.1177/014920639301900307

Zhang, C., Yu, M. C., & Marin, S. (2021). Exploring public sentiment on enforced remote work during COVID-19. *Journal of Applied Psychology*, *106*(6), 797–810. https://doi.org/10.1037/apl0000933

Zhang, N., Wang, M., Xu, H., Koenig, N., Hickman, L., Kuruzovich, J., Ng, V., Arhin, K., Wilson, D., Song, Q. C., Tang, C., Alexander, L., III, & Kim, Y. (2023). Reducing subgroup differences in personnel selection through the application of machine learning. *Personnel Psychology*. Advance online publication. https://doi.org/10.1111/peps.12593

Received November 7, 2020
Revision received June 5, 2023
Accepted July 21, 2023 ■