Analysis Plan

Project Name: Describing candidate assessments and hiring outcomes
Project Code: 2107
Date Finalized: September 9, 2022
Date updated: August 5, 2024

Note: This is an updated version of a prior pre-analysis plan which was posted on 9/9/2022. The analysis plan document was updated on August 5, 2024 to remove instructions and an internal document link that are used for internal preparation only. None of the analysis plan content was changed with this update.

Project description

This descriptive evaluation will explore *hiring outcomes across candidate assessment tools* by examining the use of different assessment tools by DOI and key hiring outcomes over the past two years. The evaluation aims to support the requirement of the Executive Order on Modernizing and Reforming the Assessment and Hiring of Federal Job Candidates that agencies shall "continually evaluate the effectiveness of different assessment strategies to promote and protect the quality and integrity of their hiring processes." It will also support DOI's Assessment Practices Working Group's goal of sharing and evaluating strategies and successful practices for assessing candidates for common occupations and hiring scenarios.

The descriptive evaluation can illustrate how different hiring practices have been used by DOI and what priority outcomes look like within each practice. This study will not provide evidence of the relative effectiveness of adopting different assessment practices related to any hiring outcomes, instead descriptive results will provide a baseline understanding of what hiring outcomes are for searches that use each type of assessment. Further research will be necessary to understand whether any observed differences in hiring outcomes between assessment types are due to differences in the assessment types themselves, or to unobserved aspects of the hiring process, such as hiring manager preferences.

Program background

The Executive Order (EO) on assessment and hiring is designed to change the application and screening process that has, historically, relied heavily on self-assessments of technical competencies ("self-assessments"). With self-assessments, applicants have an incentive to rate themselves as highly as possible, and human capital specialists who are tasked with comparing self-reported qualifications to supporting information on applicants' resumes may not have the subject-matter expertise or enough information to accurately discern which applicants are the most qualified among those with the same self-rating.

The Office of Management and Budget (OMB) has issued guidance on the EO that calls for the increased use of skills- and competency-based assessments in the federal hiring process. Specifically, agencies should, "not rely solely on candidate self-evaluations of their qualifications (e.g., resumes and occupational questionnaires). Applicants must clear other assessment hurdles in order to be considered qualified and eligible for preference and referral." Bureaus and offices within DOI have responded to this guidance by taking two approaches: (1) using subject-matter expert (SME) to manually evaluate candidate resumes; and (2) using USA Hire scored assessments of general competencies.

Bureaus and Offices within the DOI began to implement E.O. Compliant Hiring Assessment around January 2021. Prior to implementation the vast majority of hiring actions used only the self-assessment of technical competencies; after January 2021 most eligible positions use either the SME manual assessment or USA Hire scored assessment methods in addition to a self-assessment questionnaire.[1]

How hiring officials create vacancy announcements and choose assessment methods:

Hiring actions are developed within bureaus and offices at DOI. The vast majority of hiring actions at DOI that are covered by the EO recruit for a single hire within the hiring office. In some cases a hiring manager may recruit for multiple positions from the same hiring action. A vacancy announcement may mention that multiple positions are being hired within the office, but DOI generally does not recruit for multiple positions placed in multiple offices or bureaus from a single vacancy announcement for positions covered by the EO. In some instances a hiring official will create two vacancy announcements for one or more positions: one for current Federal employees, and one that is open to the public.

As a job announcement is being prepared, hiring officials work with HR specialists to identify core job competencies, develop the position description, and choose a method for assessing applicants. Hiring managers have considerable leeway in choosing the assessment method to use for a given job announcement. They may choose a particular assessment based on the job requirements, their previous knowledge of the applicants, estimated size of the applicant pool, or the urgency of needing a position filled. This discretion prevents us from evaluating the relative effectiveness of each hiring method. Ultimately, however, most positions use self-assessment questionnaires for technical competencies. Since the implementation of the EO, many hiring managers in DOI bureaus have chosen to pair self-assessments with either the SME panel method or the USA Hire scored assessment method.

Subject-matter expert (SME) panel manual assessments: Hiring officials can choose to convene subject-matter experts to review applications and score applicants based on evidence of qualifications for the position. Manual assessments by SMEs may involve one or more SMEs rating each applicant and providing a consensus rating. The SME ratings are then forwarded to an HR specialist who reviews the ratings and self-assessment ratings to create a list of certified qualified applicants (the "cert list").

[1] Current implementation of the EO does not apply to supervisory positions or temporary or seasonal positions.

USA Hire scored assessments: USA Hire is the OPM run assessment and evaluation service that is a part of USA Staffing, the federal government's integrated talent acquisition system. USA Hire has developed 118 assessments of general competencies that can be scored in order to determine applicants' eligibility for a specific job. Each assessment is tied to a particular job series (e.g., "2210: Information Technology Specialist") and grade level. These assessments measure things like reasoning skills, decision-making ability, math skills, interpersonal skills, stress tolerance, etc.

Scored assessments are implemented separately from the application process. Once an applicant submits their application, they receive an email invitation to complete the assessment, which they must complete within 48 hours. If an applicant previously completed the same assessment within the previous 6 months, they do not receive an invitation to complete a new assessment; instead, their score on file for that applicant will be applied to the new JOA application. Applicants cannot retake the assessment to achieve a higher score except after their current score expires after 6 months. Following the close of the vacancy announcement, applicant scores are forwarded with application materials to the HR specialist. If multiple scored assessments are used, or the USA Hire assessments are paired with self-assessment ratings, the scores are weighted and combined and the HR specialist creates the cert list of the highest rated candidates.

Why hiring outcomes may differ across candidate assessment methods:

Assessment methods may differ in how candidates move through the screening process, the personnel involved in assessing candidates, and how candidate competencies are assessed. Hiring managers and HR officials also vary across bureaus and offices within an agency, and may have a range of preferences for how they process and screen candidates. This descriptive study is useful to examine hiring outcomes within each of these methods, and to understand the current state of hiring processes and hiring outcomes in the Department of Interior more generally. Although this descriptive study cannot determine causal links between assessment methods and hiring outcomes, it provides a starting point to understand hiring outcomes for various assessment types. Each assessment method may vary in which candidates are included as qualified or excluded from consideration. Assessment methods could differ in their ability to rank candidates on qualifications and fitness for the position, which may affect the likelihood that a suitable candidate is chosen, or whether the search is canceled because no qualified candidates could be identified from the applicant pool.

Differences in the amount of time and labor required to rate candidates between assessment methods could affect the time needed to generate a list of qualified candidates and make selection decisions. It is also possible that hiring officials vary in how long it takes them to screen and select candidates, which could affect which assessment method they choose.

A secondary goal of this study is to examine whether hiring outcomes are, on average, different across hiring assessments. Results from this descriptive study can inform future evaluations which can examine the potential causes of differences and compare the hiring processes directly, should differences be present.

Evidence type

This is a descriptive evaluation, employing quantitative statistical analysis of administrative data. Due to data structure and availability, all analyses will explore means and proportions of outcome

variables of interest to gain a baseline understanding of how hiring outcomes may vary across candidate assessment methods. This study cannot answer questions about the effectiveness of different assessment methods in influencing hiring outcomes (i.e., we will not make any causal claims with these analyses), but the description of outcomes can provide a useful starting point for generating hypotheses about causal relationships.

Preregistration details

This Analysis Plan will be posted on the OES website at oes.gsa.gov after taking possession of the data and is therefore not blind to outcomes.

We include both confirmatory analyses and exploratory analyses in this analysis plan. However, only the confirmatory analyses will be re-analyzed by a second analyst who is blind to the results in the first analysis, in accordance with the OES project process.

Hypotheses

There are three primary research questions:

1. How are candidate assessments being implemented at DOI?
2. How long does the hiring process take and how often are positions filled for each candidate assessment method?
3. Does the representation of female, people of color, and veteran candidates change at key stages of the recruitment process within each assessment method?

These questions aim to independently describe hiring outcomes at DOI for three common methods of assessing applicant qualifications. However, this descriptive study cannot be used to evaluate the relative effectiveness of different candidate screening mechanisms (i.e., the study cannot differentiate the causal effects of using one type of assessment over another from any other factor that may differ between the assessment methods, such as hiring manager preferences or applicant pools).

Data and data structure

This section describes data that will be analyzed, including changes that will be made to the raw data with respect to data structure and variables and a description of the outcomes.

Data:

Data on JOAs and applications will be extracted from recruitment data systems by DOI and the Office of Personnel Management (OPM). JOAs and applications for this study will include hiring actions for 10 DOI critical occupations[2] within five DOI bureau/offices.[3] JOAs included in the data

[2] 0401: General Natural Resources Management and Biological Sciences Series; 0025: Park Ranger; 0462: Forestry Technicians; 1315: Hydrology; 0810: Civil Engineering; 1801: General Inspection, Investigation, Enforcement, and Compliance Series; 1350: Geology series; 1811: Criminal Investigations ; 0028: Environmental Protection Specialist; 0802: Engineering Technical

[3] U.S. Geological Survey (USGS), National Park Service (NPS), Bureau of Land Management (BLM), Fish & Wildlife Services (WS), Bureau of Reclamation (BOR).

will include permanent, non-supervisory positions that are delegated examining (Public Non Status) posted after July 2020. These parameters were chosen based on DOI's goals of building evidence for priority job series by bureaus that typically recruit the most positions each year.

JOA-level records will include fields describing an announcement or vacancy number (a unique JOA identifier), the bureau posting the JOA, a job series number, JOA characteristics (e.g., grade level, duty station, competitive/open-to-the-public vs. merit promotion announcement), fields indicating the types of candidate assessments used (self-assessment questionnaire, USA Hire/scored assessment, SME assessment), JOA dates (date opened, closed, date selection made or search canceled), and search result (e.g., offer made or canceled).

DOI will also extract applicant records for JOAs included in the study from the USA Staffing Applicant Flow data system. Each JOA typically has between 30 and 50 applicants, although this can vary greatly. Applicant data will include fields describing the date the application was submitted; demographic data associated with the applicant (including gender, age, race, and ethnicity); Veteran status indicator; indicators for whether the candidate met minimum qualifications, was included on the list of certified eligible candidates, was offered a position, and accepted an offer. Applicant data may also include self-assessment ratings where applicable and overall rating scores (on a 0-100 scale).

An important feature of the applicant data is self-reported demographic information. Applicants can submit this information voluntarily when they create their USA Jobs profile, and can opt out of providing that information with the application; HR and hiring managers do not see this information during the screening process, so opting out does not affect screening and selection outcomes. Past research suggests that between 60 and 70 percent of applicants submit their demographic information.[4]

The primary unit of analysis will be at the JOA level. JOA records are structured using one row per JOA with assessment type, bureau, job series, external/merit, other identifying fields, dates, search outcome. These data will include JOAs posted from July 2020 to present; permanent (non-seasonal) non-supervisory positions; internal/merit and external recruitment. Data describing the candidates who applied to the JOA are structured as one row per applicant-JOA combination with JOA#, applicant demographic fields (if reported by applicants), and dates and applicant status at each stage of the screening process.

Data source(s):

The evaluation will utilize administrative data from DOI recruitment systems to examine JOA's and applicant characteristics. Queries from DOI's USA Staffing Applicant Flow Data yields observations of vacancies posted and applicants to those vacancies. Data queries are performed by staff at OPM under the direction of DOI CHCO staff.

Total number of observations:

The current evaluation will focus on 10 mission-critical occupational series recruited across five major bureaus at DOI: Bureau of Land Management; Bureau of Reclamation; National Park

[4] See: https://oes.gsa.gov/projects/USAJOBS-demographics/.

Service; Fish and Wildlife Service; and U.S. Geological Survey. The study period includes JOA and applicant observations from July 2020 through June 2022.

Approximately 2,100 JOAs were posted between July 2020 and May 2022 for the 10 occupational series in the five bureaus. This includes vacancy announcements where both an internal and external JOA are posted for the same position. We will consider multiple JOAs posted for the same position as a single hiring action. After matching JOAs for the same position we expect the number of observations to be between 1,100 and 1,500 vacancies.

We expect an average of about 50 applicants per vacancy, although this number may be highly variable across vacancies. If 70% of applicants provide their demographic information, we would expect candidate-level information on 38,500 to 52,500 applicants. These applicants are not likely to be unique; many applicants may apply for multiple vacancies and show up in the data multiple times.

Outcomes to be analyzed:

The primary outcomes that will be observed for this evaluation is the length of time needed to reach hiring milestones, whether or not a search successfully results in an offer to a candidate, and the diversity of the candidate pool throughout the hiring process. Outcomes are observed with JOAs as the unit of analysis.

- *Certification time:* Duration of time (in days) between the closing of the JOA application period and the date that a certificate of eligible applicants is created.
- *Post-certification time to selection:* Duration of time (in days) between the receipt of the cert list by the hiring manager and the first offer to a candidate (for successful searches) or cancellation of the search (for unsuccessful searches).
- *Recruitment success:* Binary indicator of whether or not an offer was made to a candidate from the cert list. Success indicates at least one offer made; failure indicates no offers made and cancellation of the search/cert list.
- *Female applicants:* Proportion of applicants that identify as female among completed applications, certified eligible list, and those selected for the position.
- *Applicants of color:* Proportion of applicants who identify as non-White or Hispanic among completed applications, certified eligible list, and those selected for the position.
- *Veteran applicants:* Proportion of applicants that claim Veteran's preference among completed applications, certified eligible list, and those selected for the position.

The three outcomes describing the candidate pool will be observed at each stage of the recruitment process for each demographic group. For example, we will observe three outcomes for female applicants: proportion of completed applications, the proportion of certified eligibles, and the proportion of selected candidates.

Two secondary outcomes will look at the size of the candidate pool for JOAs that use each assessment tool: Number of completed applications and number of applicants on the list of certified eligibles.

Imported variables:

No variables will be imported for the main analyses of EO-compliant assessment tools. All data for the evaluation will be drawn from administrative data received from DOI.

Transformations of variables:

The variable that identifies the assessment method and the JOA-level outcome variables will be created from fields in the JOA-level data:

- Assessment Type: Four fields in the data, all of which are dichotomous (yes/no) indicators, allow us to categorize JOAs by assessment type. These will be used to construct three dichotomous variables: "Self-Assessment Only," "Manual Assessment," and "USA Hire Assessment." Table 1 describes the four fields in the data and how they will be used to code the three assessment type variables.
- Certification time: Indicates the number of days from the close of the JOA application period to the date when a certificate of eligible applicants (the "cert list") was issued. It will be calculated by subtracting the date in the "Announcement Close Date" field from the "Certificate Issue Date" field.
- Post-certification time to selection: Indicates the number of days from the date the certificate of eligible applicants (the "cert list") was created to the date when a candidate was selected. It will be calculated by subtracting the "Certificate Issue Date" field from the "Certificate Review Returned Date" field.
- Recruitment success: Indicates whether a candidate was selected for an official offer from the cert list. =1 if the "Certificate Status" field = "Selection made," =0 otherwise.

Applicant-level outcomes are created from fields in the USA Staffing Applicant Flow data. We will report on applicant-level outcomes if these data are provided in the data extract from DOI:

- Female applicants: A dichotomous indicator for Female will be created based on the gender identification field, where Female = 1 if Female gender is selected and = 0 otherwise.
- Applicants of color: A dichotomous indicator for applicants of color will be created based on the race category field and the ethnicity fields, where applicant of color = 1 if indicates any non-White race OR indicates Hispanic ethnicity.
- Veteran applicants: A dichotomous indicator for Veteran's status will be created based on the "Veterans Preference Adjudicated" field, where Veteran's status = 1 if "Veterans Preference Adjudicated" includes a preference code and = 0 otherwise.

Table 1. Coding of assessment types

Field in JOA-level data	Field description	Assessment type variables		
		Self-assessment only = Yes if:	*Manual Assessment = Yes if:*	*USA Hire Assessment = Yes if:*
Assessment Used[1]	=Yes if any assessment was used; =No otherwise.	=Yes	=Yes	=Yes
Assessment Questionnaire Used[2]	=Yes if a questionnaire where applicants self-report qualifications was used; =No otherwise.	=Yes	=Yes or No	=Yes or No
Manual Assessment Used	=Yes if a manual SME resume review was used; =No otherwise.	=No	=Yes	=No
USA Hire Used	=Yes if a scored USA Hire questionnaire was used; =No otherwise.	=No	=No	=Yes

Notes:

[1] JOAs where "Assessment Used" = "No" are excluded from the analysis.

[2] In practice all JOAs typically use an assessment questionnaire. JOAs that use a manual assessment or USA Hire assessment typically use those in addition to the questionnaire.

Transformations of data structure:

We will collapse the JOA records to unique hiring actions. Hiring officials will often post two JOAs for the same hiring action: one JOA for applicants from the public, and one JOA for applicants who are current Federal employees. For the purpose of observing JOA-level outcomes (time to selection, recruitment success), these JOAs will be considered as a single hiring action and collapsed to a data set with one row per hiring action for JOAs with identical bureau, occupational series, grade, announcement open and closing dates, and assessment methods. Typically these JOAs will be identical except for "Vacancy Announcement Type" will be either "DE" or "ST" indicating whether the announcement is open to the public or only to Federal employees, respectively. We will consult with DOI to identify any JOAs that may need to be matched within hiring action in a different manner.

Applicant-level data will need to be matched to JOA-level data to associate applicant characteristics with JOAs that used different assessment methods. The JOA- and applicant-level data uses a unique 8-digit numeric "vacancy number" to identify JOAs. Applicant-level information will be joined with JOA-level data based on this unique identifier.[5]

[5] The field "Announcement Number," an alpha-numeric string, is also a unique identifier for JOAs.

Data exclusion:

The descriptive study will focus on hiring activity at five DOI bureaus (U.S. Geological Service, National Park Service, Bureau of Land Management, Fish and Wildlife Service, Bureau of Reclamation) and ten "mission-critical" occupational series as defined by DOI. JOAs posted for other bureaus or series will not be included. Further, only permanent (not seasonal or temporary), non-supervisory positions will be included. Current guidance on the use of EO-compliant assessments does not apply to seasonal or temporary positions, or for supervisory positions; thus, JOAs for these positions will be excluded.

JOAs that do not have any assessment type associated with it will be excluded from the analysis. These include JOAs where the field, "Assessment Used?" = "No."

Treatment of missing data:

Missing data may arise in three ways: missing dates for JOAs, hiring actions that haven't been closed, and applicants who do not report demographic data.

JOAs with missing dates: Date fields are used to code the two outcomes related to the duration of recruitment: certification time and post-certification time-to-selection. These rely on date fields for announcement close date, certificate issue date, and certificate review returned date. Announcement close date is required to post a vacancy announcement; we do not expect this field to be missing for any observations. Certificate issue date may be missing if the recruitment was abandoned prior to a cert list being issued, or if candidates are still being assessed for the vacancy and a cert list hasn't been created yet (e.g., for more recently closed vacancies). We will only calculate the certification time outcome for JOA observations where the certificate issue date is non-missing.

Post-certification time-to-selection will similarly be calculated only for hiring actions where the certificate issue date and certificate review returned date fields are non-missing (i.e., only for hiring actions where a cert list was created and a selection was made).

Open hiring actions: Hiring actions with missing certificate issue date or certificate review returned date may be either still open and considering candidates, or canceled/abandoned hiring actions. We do not expect there to be any fields to indicate whether a hiring action has been canceled. For the "recruitment success" outcome variable, we will code a success only if the "certificate status" field is coded as "selection made" (as opposed to "no selection made" or "issued"), and will be coded as not a success otherwise. To indicate the potential for open hiring actions that have not been canceled or abandoned, we will tabulate the number of hiring actions by assessment method that do not have a certificate issued or do not have a selection made.

Missing demographic data: Applicants who choose not to submit demographic information as part of their application will be included in the analyses and coded categorically as "unreported" for the characteristic that the applicant does not report. Candidates that report some (but not all) demographic characteristics will be included in the analyses as non-missing for the demographic characteristic that they report. For example, if a candidate reports being Male but does not respond to race or ethnicity questions he would be coded as "not Female" for the gender analysis (and included in the denominator of the gender proportion calculation), but coded as "unreported"

for the race and ethnicity analysis (and not included in the denominator for the race/ethnicity proportion calculation). For each demographic characteristic we will report the proportion of applicants that did not report that characteristic.

For all types of missing data we will tabulate the number of observations that have missing data by the appropriate categories. We do not plan to impute any missing data but will conduct sensitivity tests on results as needed to determine the degree to which uncertainty about missingness may be driving the results.

Descriptive statistics, tables, and graphs

We are interested in describing how DOI is implementing EO-compliant candidate assessments, summarizing priority hiring outcomes for each assessment method (self, manual, and USA Hire), and tabulating the representation of demographic characteristics as candidates move through the application process, given the assessment method.

We plan to describe the current state of assessment use across DOI. This analysis will report the utilization rates for each of the three assessment types from July 2020. Analysis will provide the proportion of use for each month, for each bureau, and by job series. We will execute three analyses that examine 1) the rate of use for all three assessment types across DOI, 2) the rate of use for all three assessment types for select bureaus in DOI, and, 3) the rate of use for all three assessment types across select DOI job series, each by month.

The findings of the assessment-level descriptive analyses will be displayed using ridgeline plots created in R statistical software. Ridgeline plots allow comparison by overlaying distributions among different data groups and within the same data groups. Using these plots we can map the use of assessments over time and compare trends across bureaus and job series.

Figure 1. Sample illustration of ridgeline plots

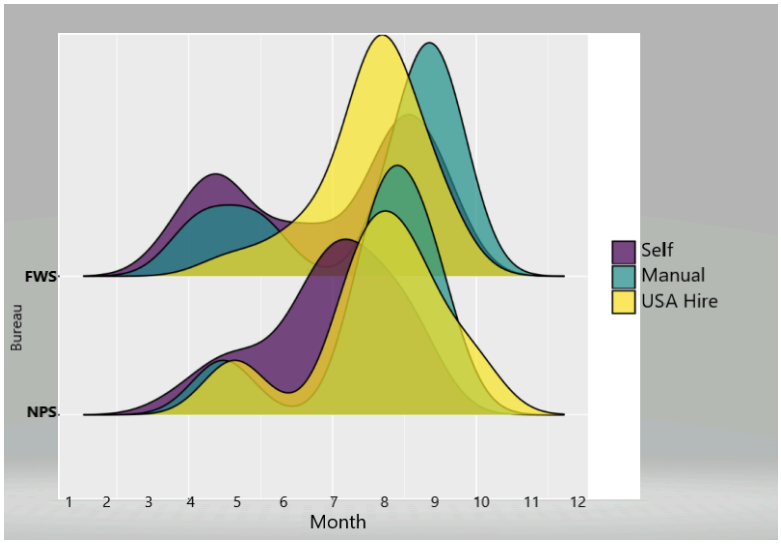

Table 2 describes the outcomes we will summarize for hiring actions. Mean outcomes reported include certification time, post-certification time to selection, and the rate of successful searches.

We will report the proportion of applicants who are female, non-White or Hispanic, and Veterans at each stage in the process – applications submitted, certified eligible, and selected – for all JOAs that use a given assessment method.

Table 2. Example table of mean outcomes and applicant characteristics for a given assessment method. There will be three of these tables, one for each assessment method.

JOA-level outcomes			
	Certification time (# of days)	Post-certification time to selection (# of days)	Recruitment success (%)
Mean			
St. Dev.			
# hiring actions			
% of hiring actions where a cert list was issued			
% of hiring actions with a cert list issued but no decision recorded			
Applicant-level characteristics			
	Application	Cert list	Selection
Female (%)			
Did not report gender (%)			
Non-White or Hispanic (%)			
Did not report race or ethnicity (%)			
Veteran (%)			

For the applicant-level characteristics, we will sum the number of applicants that indicate each demographic characteristic at each stage of the recruitment process. For each cell in the table we will calculate the proportion of applicants that fall in each category using the total number of applicants that provided demographic information among applicants at that stage in the process. For example, suppose 100 people applied for a position and 70 provided demographic information. If 35 applicants indicated that they are female, the female applicant proportion would be 0.5.

For each cell in table 2 we will tabulate the proportions by pooling applicants across all JOAs that use each assessment type. This tabulation method gives equal weight to each applicant who applies for a position using any assessment type.

$$P_{c,s,a} = \frac{\sum_i Y[C_i=1]_{s,a}}{\sum_i Y[C_i=1 \mid C_i=0]_{s,a}},$$

where C_i is an indicator of whether an individual applicant indicates demographic characteristic c, observed at recruitment stage s for JOAs using assessment method a. We will also report for each demographic characteristic (gender, race/ethnicity) the proportion of applicants at each stage that did not report that characteristic.[6]

Statistical models and hypothesis tests

This section describes the statistical models and hypothesis tests that will make up the analysis — including any follow-ups on effects in the main statistical model and any exploratory analyses that can be anticipated prior to analysis.

Statistical models:

No statistical models will be estimated for the descriptive analyses.

Confirmatory analyses:

Analyses to answer the first two research questions (How are different candidate assessments being implemented at DOI? How long does the hiring process take and how often are positions filled for each candidate assessment method?) will report counts and averages, but will not involve confirmatory analyses. The third research question (Does the representation of female, people of color, and Veteran candidates change at key stages of the recruitment process for vacancies that use each assessment method?) will present summaries of demographic profiles at each stage in the recruitment process and test the following hypothesis:

Hypothesis 1: The proportion of applicants who are female, non-White or Hispanic, or Veterans does not change at different stages in the recruitment process (from application to cert list to selection) for JOAs that use a given assessment method.

[6] Veteran's status is not separately self-reported with other demographic characteristics. Part of the application process asks applicants if they claim Veteran's status, which is then adjudicated by HR specialists based on documentation provided by the applicant. All applicants will have a Veteran's status reported.

This hypothesis describes, for a given assessment method, the proportion of applicants in a demographic category at each stage in the recruitment process and whether differences from the previous stage in the process are likely to be observed by chance.

For each assessment method we will calculate the proportion of applicants in each demographic group who completed an application, were included on the list of certified eligible applicants, and were selected for a position. We will then compare the proportions within demographic groups and assessment method across the three recruitment stages. We will not conduct any tests comparing representation between assessment methods.

For example, among JOAs that use manual assessments we will compare the proportion of Female applicants on the certified eligible list to the proportion of completed applications submitted by Females. Similarly, for the same JOAs that use manual assessments, we will compare the proportion of selected candidates who are Female to the proportion of Female applicants on the certified eligible list.

We will conduct a multinomial goodness-of-fit test of the difference in proportions for each comparison within demographic group–assessment method combination. We will calculate three chi-squared test statistics for each group–assessment combination: cert list proportion to applicant proportion; selected proportion to cert list proportion; and selected proportion to applicant proportion. The chi-squared test statistic is calculated as:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} , \qquad (1)$$

where k is the number of demographic categories (k=2 in our case, e.g., female or not female), O_i is the observed number of people in the sample with demographic characteristic k, and E_i is the expected number of people in the sample with demographic characteristic k given the frequency of that characteristic in the population. The calculation of E_i depends on which of the three comparisons we are testing. For comparisons to the population of applicants (either the proportion of a demographic characteristic on the cert list or among selected candidates), $E_{i,app} = \frac{d_{i,app}}{N_{app}}$ where $d_{i,app}$ is the number of applicants with characteristic k among N_{app} total applicants. For the comparison of selected candidates to the cert list population, $E_{i,cert} = \frac{d_{i,cert}}{N_{cert}}$, where $d_{i,cert}$ is the number of applicants on the cert list with characteristics k among N_{cert} total applicants on the cert list.

The goodness-of-fit chi-squared test statistic is used because each test involves a sample from a known population, and we want to test whether the observed (sample) proportion of the demographic characteristic of interest is significantly different from the expected proportion given observed proportion in the population. That is, we treat the population from which the sample is drawn as a well-defined and known distribution. For example, suppose 100 people apply for a position and 40 of the applicants are female. From these applicants 20 are included on the cert list and five of these are female. If gender is unrelated to the likelihood of an applicant making

the cert list, we might expect eight of the 20 cert list applicants to be female. The goodness-of-fit test estimates the likelihood of observing five female candidates on the cert list given that we expect eight candidates on the cert list based on the proportion of female candidates among all applicants. In this case $\chi^2 = 1.875$ (d.f. = 1), which would not reject the hypothesis that the frequency of females on the cert list is the same as the population of applicants (p = .17).

Exploratory analysis:

In order to facilitate future research, we will conduct an exploratory analysis of variance (ANOVA) for the JOA-level outcomes (certification time, post-certification time to selection, and recruitment success) observed for the three assessment methods. The primary categorical variable used for comparison in the one-way ANOVA is assessment type (self assessment, manual assessment, or USA Hire). Mean differences across assessment types will be examined using a one-way analysis of variance (ANOVA). An F test will test the null hypothesis that the three assessment methods do not have significantly different outcomes. If the ANOVA detects significant differences in outcomes between assessment methods, we will follow the ANOVA with a Tukey's post-hoc test to describe pairwise differences in outcomes between assessment methods.

The ANOVA will allow us to examine whether mean hiring outcomes are statistically different across assessments (i.e., unlikely to be observed due to chance). However, the exploratory analysis cannot provide evidence about whether any of the assessment methods are more or less effective than the others or whether any differences are due to the use of a particular assessment, rather than by differences in hiring manager preferences or applicant pools, for example.

We will conduct an exploratory analysis of the primary outcomes by bureau. Five bureaus will be included in the descriptive evaluation: U.S. Geological Service, National Park Service, Bureau of Land Management, Fish and Wildlife Service, Bureau of Reclamation. We will report mean outcomes for certification time, post-certification time-to-selection, and recruitment success by assessment method for each of the five bureaus and report the results of a one-way ANOVA for each of the bureaus. We will also report the cross-tabulations of number and proportion of applicants at each stage in the process by demographic group and assessment method for each of the five bureaus.

Inference criteria, including any adjustments for multiple comparisons:

For all of the statistical tests we will use a rejection region of p=0.05. Tests that return p<=0.05 will be considered statistically significant, and we will infer that the hypothesis can be rejected.

The cross-tabulation analysis of applicant proportions for demographic groups will involve three chi-squared tests for each demographic group within each assessment type, resulting in a total of 27 tests. We will adjust inferences for multiple comparisons within demographic group–assessment method combinations using the Holm-Bonferroni adjustment. That is, we will adjust for multiple comparisons in nine different families of tests, each of which will involve three comparisons.

In the exploratory ANOVA, if we reject the null hypothesis for the one-way ANOVAs (using an F-test) we will conduct a Tukey's post-hoc analysis, which will compare the means across

subgroups of data in order to identify where the differences exist and adjust inferences for multiple comparisons.

Limitations:

The study is composed of descriptive analyses; observed differences in outcomes cannot be attributed to differences in the assessment methods themselves. The study does not attempt to control for several confounding factors that may account for any observed differences in outcomes among comparison groups. Inferential statistics are used to help discern when estimated differences in outcomes are or are not likely due to chance, but we cannot infer the cause of differences or whether observed or unobserved factors may explain differences.

We also are limited in the conclusions we can draw about the processes that underlie each assessment method and how those processes relate to observed outcomes. We do not observe how candidates are screened for the JOAs and candidates in the data set (e.g., how much time is spent on reviewing resumes and questionnaire responses), so we cannot attribute observed differences in outcomes to specific characteristics of the assessment processes.

Finally, this study can only describe the use of assessment tools and recruitment outcomes for a subset of DOI bureaus and job series. Although this subset accounts for a large share of hiring at the agency, results may not be generalizable to other bureaus or job series. Similarly, limiting the analysis to permanent and non-supervisory positions limits how conclusions can be extended to other types of positions.

Link to an analysis code/script:

N/A