MILITARY PSYCHOLOGY
2024, VOL. 36, NO. 1, 33–48
https://doi.org/10.1080/08995605.2021.1997500

Development of a situational judgment test to supplement current US air force measures of officership

Laura G. Barron[a], Imelda D. Aguilar[b], Mark R. Rose[a], and Thomas R. Carretta[c]

[a]Air Education and Training Command, San Antonio, Texas, USA; [b]Air Force Personnel Center, San Antonio, Texas, USA; [c]Air Force Research Laboratory, Dayton, Ohio, USA

ABSTRACT
Aptitude requirements for US Air Force officer commissioning include completion of a college degree and minimum scores on the Air Force Officer Qualifying Test (AFOQT) Verbal and Quantitative composites. Although the AFOQT has demonstrated predictive validity for officer training, the Air Force has striven to improve predictive validity and diversity. To this end, a Situational judgment Test (SJT) was added to the AFOQT in 2015. SJT development was consistent with recommendations to broaden the competencies assessed by the AFOQT with the goal of providing incremental validity, while reducing adverse impact for historically underrepresented groups. To ensure content validity and realism, SJT development was based on competencies identified in a large-scale analysis of officership and input from junior officers in scenario and response generation and scoring. Psychometric evaluations have affirmed its potential benefits for inclusion on the AFOQT. An initial study showed the SJT to be perceived as highly face valid regardless of whether it was presented as a paper-and-pencil test (with narrative or scripted scenarios) or in a video-based format. Preliminary studies demonstrated criterion-related validity within small USAF samples, and a larger Army cadet sample. Additionally, operational administration of the SJT since 2015 has demonstrated its potential for improving diversity (i.e., reduced adverse impact relative to the AFOQT Verbal and Quantitative composites). Predictive validation studies with larger Air Force officer accession samples are ongoing to assess the incremental validity of the SJT beyond current AFOQT composites for predicting important outcomes across accession sources.

ARTICLE HISTORY
Received 31 March 2021
Accepted 8 September 2021

KEYWORDS
Situational judgment; officership; diversity

What is the public significance of this article?—This study shows that a Situational judgment Test (SJT) for military officer qualification results in smaller racial and gender differences than verbal and quantitative aptitude tests, and increases overall validity (as an indicator for military field training and academic outcomes) when used in combination with these tests. Military members view the SJT format favorably regardless of whether it is presented in a written or video-based format.

In the personnel selection literature, situational judgment tests (SJTs) have shown to be an important and useful complement to more traditional sign-based predictor instruments (Lievens & Coetsier, 2002). Situational judgment tests place the examinee in a situation that closely resembles or simulates one that may be encountered on the job and elicit their procedural knowledge about how to respond to the stimuli. Content-specific SJTs are based on the idea of behavioral consistency (examinee's performance on the test will be consistent with their future job performance) (Corstjens et al., 2017). This notion of behavioral consistency has been proposed as the most straightforward explanation for the positive predictive validity results of situational judgment tests (Lievens & Coetsier, 2002).

Research has found SJTs have several positive features including validity approaching that of cognitive ability tests with studies showing SJTs having incremental validity above and beyond traditional predictors such as cognitive ability and personality (Weekley & Ployhart, 2006). In addition, SJTs normally exhibit small to moderate racial subgroup differences, substantially less than typically observed for traditional cognitive ability tests (Motowidlo & Tippins, 1993; Pulakos & Schmitt, 1996; Weekley & Jones, 1999, Hough et al., 2001).

CONTACT Imelda D. Aguilar imelda.aguilar@us.af.mil HQ AFPC/DSYX, 550 C Street West Ste 45, JBSA Randolph AFB, TX 78150-4747.

The views expressed are those of the authors and not necessarily those of the United States government, the Department of Defense, or the US Air Force.
This article has been corrected with minor changes. These changes do not impact the academic content of the article.

Impetus for the AFOQT situational judgment test

The Air Force Officer Qualifying Test (AFOQT) is a standardized test that measures verbal and mathematical aptitude and additional aptitudes (perceptual speed, spatial ability) related to specific aviation career fields. It is primarily used to select college graduates for entry-level officer positions in the US Air Force. Composites of certain AFOQT subtests are also used for qualification for aircrew career fields. Although the AFOQT has demonstrated predictive validity for officer training (Roberts & Skinner, 1996), recommendations have been made to broaden the competencies assessed with the goal of providing incremental validity, while reducing adverse impact for historically underrepresented groups. This led to the addition of the Self-Description Inventory (SDI), a Big Five personality test, in 2005 (see Kantrowitz et al., 2019), and the Situational judgment Test (SJT) in 2015.

Prior to the development of the SJT, the Air Force Personnel Center Strategic Research and Assessment Branch (AFPC/DSYX) directed an initiative to identify competencies required to be an effective Air Force officer across occupational specialties, and to link these core requirements to potential predictors that could be quantified and considered as assessment measures for future versions of the AFOQT. Based on a combination of focus groups with subject-matter experts (SMEs), and surveys administered to 4,436 officers, this initiative identified seven broad core competencies required for effective Air Force officer performance across specialties: (i) Displaying Integrity, Ethical Behavior, and Professionalism, (ii) Leading Others, (iii) Decision-Making and Managing Resources, (iv) Communication Skills, (v) Leading Innovation, (vi) Mentoring Others, and (vii) Pursuing Personal and Professional Development. All seven competencies were evaluated as important beginning at the lieutenant or captain level (i.e., during the first obligated period of service) (Lentz et al., 2009a). Based on a review of existing AFOQT measures, SJT development was recommended to address the officership performance requirements identified (Lentz et al., 2009b). To help ensure the SJT captured a broad range of situations that officers encounter, these officership competencies were used to guide solicitation of critical incidents.

However, rather than taking a strictly construct-oriented approach (e.g., relying on psychologists as SMEs with theoretical knowledge of target competencies; see Whetzel et al., 2020), we aimed to develop an assessment that would produce an overall SJT score based on the extent to which applicant judgment converged with the judgment of experienced officers across common situations. We believed that a strictly construct-oriented approach might be overly transparent to test takers (posing a potential threat to test security if applicants were coached on which competencies the test questions were designed to measure), and instead generated and scaled response options based on overall officer-rated effectiveness rather than the extent to which each response demonstrated a high level of a single competency.

Article overview

In this paper, we first describe the development of AFOQT SJT content (i.e., identification of scenarios, response alternatives, and scoring key) over three phases. We then describe two studies conducted to guide selection of SJT items from overlength SJT versions. In Study 1, conducted in an enlisted (Basic Military Training) sample, we sought to evaluate psychometric properties and compare subgroup differences relative to those found on the Armed Services Vocational Aptitude Battery (ASVAB). In Study 2, we obtained initial evidence of criterion-related validity in small samples from each US Air Force accession source, using this data to prioritize and select the final items for inclusion on the operational AFOQT SJT that was implemented in 2015. In Study 3, we evaluated face validity perceptions, assessing whether a video-based or script-based SJT format may provide additional advantages beyond the standard written SJT version. In Study 4, we administered the AFOQT SJT to an Army Reserve Officer Training Corps (ROTC) sample to evaluate criterion-related validity in a larger officer cadet sample. Finally, in Study 5, we document subgroup differences on the SJT as compared to the traditional cognitive tests on the AFOQT, based on the full population (Air Force officer candidates) that have been administered the test operationally since 2015.

SJT content development

The development of the AFOQT SJT was a three-phase process. Each phase is described in the sections that follow.

Phase 1 (scenario development)

Phase 1a: SME Identification of Critical Incidents. In a series of focus groups, 79 captains (O-3s) described critical incidents (Flanagan, 1954) experienced or observed of others at the lieutenant or captain (O1-O3) levels. Participants were asked to provide a particularly effective or particularly ineffective example of each of the seven competencies, including the situation, task, action, and result. Feedback from initial focus groups indicated that the participants had trouble identifying particularly effective or ineffective incidents involving the competency of Pursuing Personal and Professional Development, with many participants indicating that there was little variability among officers because all simply followed the mandated training/educational requirements. As a result, later focus groups focused only on the six other competencies. We note that this may be an area for more in-depth exploration in development of future AFOQT versions.

Phase 1b: Scenario Editing. The situations and tasks described in the 500+ critical incidents were then edited to create 100 SJT item stems; described actions were retained as possible response options to supplement response options to be identified in Phase 2. The editing process focused on (i) combining critical incidents to remove duplicative content, (ii) removing incidents that required specific technical expertise, seemed to have only one obvious way to address the problem, or did not include important details needed to meaningfully decide upon an effective course of action, and (iii) editing language to be more accessible to applicants without knowledge of specific Air Force terminology and procedures, and to generally improve written clarity.

Phase 2 (response generation)

We next sought input from two groups – current Air Force officers (SMEs) and newly enlisted accessions (non-SMEs) – to identify potential responses options that would likely be viewed as plausible, and yet varied in actual (SME-rated) effectiveness. First, in a series of focus groups, 22 captains and 31 1st lieutenants were asked to provide a response to half of the 100 scenarios, resulting in approximately 25 incumbents' responses per scenario. To provide additional variability in responses, 100 enlisted Airmen undergoing Basic Military Training (BMT) were asked to identify potentially effective responses to the same scenarios using modified written instructions. Each Airman was asked to provide a response to 10 of the 100 scenarios, resulting in approximately 10 novice responses per scenario. Responses were then summarized with the goal of creating 5–7 meaningfully distinct strategies for responding to each of the scenarios. Eighty-six scenarios with 5–7 response alternatives survived this process.

Phase 3 (scoring key development)

Phase 3a: SME Response Effectiveness Ratings. Though SJTs can be scored based on various methods, research has suggested that scoring keys based on input from SMEs (such as supervisors and/or high-performing incumbents) are likely the preferred method. Specifically, such research has found that an SJT scored based on SME effectiveness ratings has greater validity for predicting job performance than keys based on theory, contrasts between experts and novices, or even contrasts between high-performing and low-performing incumbents (Bergman et al., 2006).

Given the demonstrated validity of scoring keys based on SME judgment, we sought input from Air Force junior officers who had performed at a high level. High-performing incumbents were identified based on the award of Distinguished Graduate (DG) from Squadron Officer School (SOS), an eight-week course required of all captains, and representing the first opportunity for officers to be assessed relative to their peer group in a common setting. The individual award of DG recognizes those in the top 10% based on (i) flight commanders' evaluations of leadership, team building, problem solving, communication skills, and physical conditioning, (ii) peer recognition, and (iii) academic testing. This award has been shown to be strongly linked to officer promotion decisions (Bruns & Eichorn, 1993).

In May 2013 e-mail invitations to complete the online survey were sent to all 845 current USAF captains who had been awarded SOS DG (787 e-mails were deliverable). To minimize the survey completion time, the 86 SJT scenarios were divided into three surveys, such that each captain was asked to rate actions involving only one-third of the scenarios; scenario order was randomized. Three hundred sixty-nine of the 787 recipients (46.89%) completed at least a portion of one of the online surveys, and 264 of the 787 recipients (33.55%) completed a full survey, for a sample size of 93–113 incumbents per scenario. Incumbents indicated the single most effective response and the single least effective response of those listed. Additionally, incumbents separately rated the effectiveness of each response alternative using a 1–7 Likert-type scale, anchored *1 = Very Ineffective action to address the situation* and *7 = Very Effective action to address the situation.*

Phase 3b: Analyses to Support Scoring Key Development. To check that the responses for each scenario varied significantly in rated effectiveness, we conducted a repeated-measures ANOVA for each of the 86 scenarios, with response alternative as the factor (with 5–7 levels) and SME rated effectiveness as the dependent variable. All scenarios included response alternatives that overall varied significantly in effectiveness. Next, to identify response alternatives that could be keyed as the most and least effective alternatives for each scenario on the SJT, a series of paired-samples t-tests comparing the effectiveness of response alternatives for each scenario were run. To correct for the family-wise error rate, the Bonferroni correction was applied, such that *t* statistics were evaluated at an alpha level of .05 divided by the number of pairwise contrasts. Response options with the highest mean effectiveness rating were also rated by the greatest percentage of SMEs as the "Most Effective" listed response.

To apply the Motowidlo et al. (1990) scoring method–for a possible score of −1 to +1 for each item ["Select the MOST EFFECTIVE" and, separately, "Select the LEAST EFFECTIVE"]) and total possible score of −2 to +2 for each scenario – we required responses keyed as most effective to have been judged significantly more effective than each of the other retained alternatives. To be keyed as the least effective, we required the response to have been judged significantly less effective than each of the other retained alternatives. For scenarios that contained more than 5 alternatives, alternatives were chosen for exclusion from the operational test that would allow for a clearer contrast between keyed and unkeyed responses. In Study 1, we additionally evaluated an alternative scoring strategy that involved summing the mean SME effectiveness ratings assigned to responses that applicants choose as the Most Likely and subtracting the mean SME effectiveness ratings assigned to responses that applicants choose as the Least Likely (Knapp et al., 2001).

Study 1 initial vetting of overlength SJT in basic military trainee sample

As an initial check on the psychometric properties of the SJT items, and to reduce the pool of SJT items to a smaller number for the operational test, an overlength SJT version was administered to a convenience sample of new enlisted accessions undergoing Basic Military Training (BMT). For each scenario, participants responded to two items: "Select the MOST EFFECTIVE action in response to the situation" and "Select the LEAST EFFECTIVE action in response to the situation." Instructions indicated, "Your responses will be scored relative to the consensus judgment across experienced US Air Force officers."

Method

Newly enlisted Basic Military Trainees served as a convenience sample to evaluate the test; female trainees were oversampled to evaluate gender differences. Participants included 320 trainees (51.3% female); 20 trainees had completed an Associate's degree and 22 had completed a Bachelor degree. Most participants identified as White (72.81%), Black/African-American (15%), and/or Asian (5.94%). Separately, 16.25% identified as Hispanic or Latino.

Given constraints limiting test administration time to 1 hour, half of the BMT sample ($N = 159$) completed the first half of the SJT (scenarios 1–43) and the other half ($N = 161$) completed the second half of the SJT (scenarios 44–86). Test proctors indicated that groups typically finished in 45 minutes to 1 hour, suggesting the need to allow 1 to 1.4 minutes per scenario.

SJT scores were matched to each trainee's pre-accession ASVAB scores (i.e., typically completed approximately 6–12 months earlier at Military Entrance Processing Stations) to provide an initial exploration of overlap between the USAF SJT and traditional cognitive tests (Armed Services Vocational Aptitude Battery subtests), and a comparison of subgroup differences.

Results

Scoring method

The SJT was scored using two methods: (i) −1 to +1 scoring for Most Likely (ML) and Least Likely (LL) based on responses SMEs judged significantly more or less effective than other options (Motowidlo et al., 1990), and (ii) imputed SME mean effectiveness ratings (positive ratings for Most Likely responses and inverse ratings for Least Likely responses) (Knapp et al., 2001). Scores based on the alternate scoring methods were highly correlated. For SJT Version 1 (scenarios 1–43) total scores based on the alternate scoring methods correlated .976; for Version 2 (scenarios 44–86) scores correlated .980. Both methods resulted in total (summed) scores across the 43 scenarios that approximated normal distributions (though scores showed a slight negative skew; see Figure A2). For congruence with AFOQT test-takers' traditional expectation that AFOQT subtest items will contribute equally to their total score, as well as logistical issues based on how the AFOQT is currently administered (i.e., paper-and-pencil answer sheet with five response options), the Motowidlo et al. (1990) method was used in further analyses.

Internal reliability by SJT length

As a minimum requirement for an item to be used we required a positive item-total correlation. This minimizes the possibility that the item is assessing extraneous content that is inconsistent with interpersonal and decision-making skills required of officers. All but 2 of the 86 scenarios met this criterion (these excluded items are included in the Appendix). Internal consistency of valid SJTs is often low due to the multi-dimensionality of constructs measured by SJTs (McDaniel et al., 2007). In this study, coefficient alpha was greater than that typically found in the SJT literature (see Campion et al., 2014). Forms comprised of the 25 scenarios with the strongest item-total correlations would be sufficient, with α = .79 for Version 1, and α = .80 for Version 2. See Table 1.

Correlations between the SJT and ASVAB subtests

The SJT literature consistently indicates that, despite the potential usefulness of SJTs for assessing relevant personality traits, SJT scores are at least moderately correlated with general cognitive ability (GCA). Specifically, meta-analysis has found a mean correlation of .17 between GCA tests and SJT scores using behavioral tendency (e.g., ML and LL) instructions, and a mean correlation of .32 for SJTs using knowledge instructions (McDaniel et al., 2007). Consistent with the literature, Armed Forces Qualification Test (AFQT) scores correlated .20 to .23 with the US Air Force SJT, with SJT scores more highly correlated with verbal ability than quantitative ability. See Table 2.

Subgroup differences on SJT as compared to ASVAB

One of the potential advantages of including the SJT is the potential for reduced subgroup differences relative to traditional cognitive tests. As has been well documented, the ASVAB shows substantial subgroup differences, with African-Americans scoring substantially lower than other racial and ethnic groups. Additionally females and Hispanic/Latino examinees have scored somewhat lower than males and White examinees. We found that although female Airmen in our BMT samples ($N = 157$) scored ~0.4 standard deviations (SDs) lower than males ($N = 161$) on the AFQT, female Airmen outscored male Airmen on the SJT by ~0.2 to 0.4 SDs (a statistically significant difference in SJT scores for one of the two SJT versions, $p < .05$). There were no statistically significant race or ethnic score differences on either SJT version. African-American Airmen in our samples ($N = 48$) scored ~0.4 to 0.9 SDs lower than Whites ($N = 233$) on the AFQT, but African-American Airmen scored only ~0.1 SDs lower on the SJT. Findings for Latinos ($N = 52$) relative to non-Latinos ($N = 163$) were inconsistent across the two samples. See Table 3 for Cohen's d values.

Table 1. Study 1 internal consistency (α) reliability estimates.

Test length	SJT version 1 (1–43)	SJT version 2 (44–86)
43 scenarios	.770	.781
Selected 20 scenario test (~24 min)	.778	.790
Selected 25 scenario test (~30 min)	.786	.804
Selected 30 scenario test (~36 min)	.790	.809
Selected 35 scenario test (~42 min)	.791	.802

Table 2. Correlations between SJT total scores and ASVAB scores.

ASVAB component	SJT version 1 (N = 159)	SJT version 2 (N = 161)
Armed Forces Qualifying Test (AFQT)	.197	.227
Paragraph Comprehension (PC)	.159	.229
Word Knowledge (WK)	.301	.217
Arithmetic Reasoning (AR)	.086	.085
Math Knowledge (MK)	.048	.062

Table 3. Standardized mean score differences between subgroups for BMT sample.

Test score	Sample	Female–Male	Latino–Non-Latino	Black–White
AFQT	Sample 1	−0.425	−0.345	−0.884
	Sample 2	−0.400	+0.167	−0.412
PC	Sample 1	−0.213	−0.120	−0.160
	Sample 2	−0.128	−0.009	−0.467
WK	Sample 1	−0.103	−0.287	−0.867
	Sample 2	−0.152	+0.017	−0.619
AR	Sample 1	−0.648	−0.165	−0.743
	Sample 2	−0.495	−0.084	−0.075
MK	Sample 1	−0.319	−0.296	−0.288
	Sample 2	−0.267	−0.125	−0.068
SJT	Sample 1 (scenarios 1–43)	+0.167	−0.035	−0.184
	Sample 2 (scenarios 44–86)	+0.362	−0.367	−0.028

Study 2 vetting of overlength SJT in officer trainee samples

Before new forms of the AFOQT are introduced, the Air Force normally conducts field tests in which a sample of Air Force ROTC (AFROTC) and US Air Force Academy (USAFA) cadets, and Officer Training School (OTS) Basic Officer Trainees who have already (recently) met AFOQT and other officer commissioning minimums are administered an overlength version of the new test form. These results then allow for selection of the final items that appear on the new operational versions. In the field test, two overlength versions were administered; each version included 30 scenarios, including 11 anchor scenarios on both versions, and 19 unique scenarios per version. Anchor items were selected based on relatively higher item-total correlations and relatively lower item-level subgroup differences in Study 1. The two versions were administered to 344 and 310 cadets/trainees, respectively.

To identify which 25 scenarios would appear on each version of the AFOQT Form T, selection was based on item-level criterion-related validation against cadet military training performance outcomes available from each accession source. Specifically, for each individual item ("Select the MOST EFFECTIVE" and, separately, "Select the LEAST EFFECTIVE") we computed a sample-size weighted average of validity coefficients across the three criteria (i.e., validity coefficients based on AFROTC, OTS, and USAFA samples, respectively). Despite the small sample sizes, in order to make necessary scenario deletions to reduce test length, scenarios that did not show a positive relationship across criteria (positive sample-size weighted average validity coefficient) for both corresponding items ("Select the MOST EFFECTIVE" and, separately, "Select the LEAST EFFECTIVE") were prioritized for exclusion.

Method

Participants

Convenience samples of AFROTC and USAFA cadets, and OTS trainees, were administered overlength versions of several AFOQT subtests at USAFA and AFROTC detachments (total *N*s = 344 and 310 for Version 1 and 2, respectively). Test scores did not affect career outcomes for any participants (all participants had already met officer qualification minimums based on the previous AFOQT form). If available, individual participant scores were matched to the performance criteria used by their commissioning source as a potential (limited) indicator of criterion-related validity.

Criteria

Each officer accession source evaluates its cadets and trainees using a different set of officership metrics. The *Basic Officer Training Order of Merit (OM)* and *AFROTC Field Training* rankings are calculated based on a combination of training commander ratings, written and physical fitness test scores, and performance in a Leadership Reaction Course and simulated deployed environment (see Holm Center, 2015). Current rating areas addressed on training performance reports include: Leadership Skills, Professional Qualities, Communication Skills, judgment and Decision Making, and Warrior Ethos. *USAFA Military Performance Average* (MPA) is calculated based on performance in standardized inspections (personal appearance, dorm room), commissioning education, and leadership positions within the USAFA wing structure (across semesters); rating areas include Duty Performance, Professional Qualities (Intrinsic Motivation, Teamwork), and Character (Integrity, Service Before Self, Excellence) (see U.S. Air Force Academy, 2020).

Table 4. Validity coefficients (*r*) for overlength SJT version (field test), by officer accession source.

	Mean	SD	AFROTC field training rank (*N*s = 121 and 99)	USAFA military performance average (*N*s = 60 and 41)	Basic officer training order of merit (*N*s = 53 and 53)
SJT version 1 (all 30 items)	.65	.16	.12	.22†	.25†
SJT version 2 (all 30 items)	.62	.17	.20*	.12	.29*

Independent sample sizes with matched criterion data reported for version 1 and 2, respectively.
**p* < .05 (two-tailed); †*p* < .05 (one-tailed).

Results

Even prior to selection of better-performing items based on the officer cadet samples, results showed a modest positive relationship between the 30-item SJT versions and the aggregate performance criteria that is tracked by each officer accession source. Version 1 was significantly related to USAFA Military Performance Average ($r = .22$, $p < .05$, one-tailed) and Basic Officer Training (BOT) Distinguished Graduate Order of Merit ($r = .25$, $p < .05$, one-tailed). Version 2 was significantly related to AFROTC Field Training class ranking ($r = .20$, $p < .05$, two-tailed) and BOT Order of Merit ($r = .29$, $p < .05$, two-tailed). See Table 4.

Although our focus in item selection was criterion-related validity (given the intended use of an overall SJT score rather than competency-level scores), we conducted some exploratory analysis to map the selected SJT items back to the six Lentz et al (2009b). competencies that had guided critical incident (scenario) generation. We did this in two ways. First, we asked a group of psychologists (three Ph.D. personnel research psychologists) to determine which of the six Lentz et al. officership competencies each SJT scenario most closely corresponded to. Separately, we asked an independent group of industrial/organizational psychologists (one Ph.D. and two masters-level graduates) to review each scenario – together with the corresponding response options and key – to reach consensus on competency categorization. Overall, results based on either method showed that a larger number of selected SJT scenarios corresponded to Leading Others than any other competency, although all competencies were represented. The two approaches generally converged, with Cohen's kappa values ranging from .602 for Communication to .907 for Mentoring Others. See Table 5.

Discussion

Results generally supported the validity of the SJT as likely to be a useful indicator to identify examinees most likely to excel within officer training. Concurrent validity

Table 5. Exploratory mapping of Lentz et al. officership competencies to SJT scenarios.

	Consensus competency correspondence based on review of . . .		
	Scenario only	Scenario±Responses ±Key	Kappa
Leading others	11	9	.863
Mentoring others	6	7	.907
Integrity/Professionalism	7	6	.720
Communication	6	6	.602
Decision-Making/Managing resources	4	7	.684
Leading innovation	3	2	.786
Total	37 scenarios (across SJT versions T1 and T2)		

Scenario-only categorization is based on the consensus across three Ph.D. personnel research psychologists (identification of which of the six target competencies each SJT scenario most closely corresponded to). A second independent group of three industrial/organizational psychologists (one Ph.D. and two masters-level graduates) completed the same categorization task based on review of each scenario *and* the corresponding response options and key. Cohen's kappa values indicate the extent of agreement between these two groups of raters.

coefficients for the SJT were in the range typically found within the broader SJT research literature. However, there were several limitations (including but not limited to the small sample sizes of examinees with criterion data available) that should be addressed in future research on the AFOQT SJT. First, the aggregate criterion data tracked by the accession sources do not directly parallel the constructs the SJT was intended to assess: although some of the rating areas that contribute to the calculation of each accession source's aggregate Distinguished Graduate metric are directly applicable (e.g., commander ratings on Leadership), many scored areas that are included in the aggregate metrics (e.g., physical fitness and personal appearance) are not. Second, the criterion data are clearly limited to what can be observed in a training environment and as such likely assess maximal rather than typical performance. Finally, because all members in the current study had already met officer commissioning minimums on the AFOQT, members did not have any strong incentive to perform well on the SJT. To the extent that these motivational differences introduce construct-irrelevant variance, the observed validity coefficients may underestimate the relationship that would occur in the operational testing environment.

Study 3 applicant reactions and evaluation of alternate SJT formats

Given that the AFOQT is administered to civilians who may not have yet formed a strong impression of the Air Force, and may still be weighing other career options at the time of applying to OTS, we believed it important to evaluate how the SJT would be perceived by examinees. On the one hand, applicants may view SJTs more favorably than traditional cognitive tests (or self-report personality measures) to the extent that they allow applicants to demonstrate how they would respond in scenarios similar to those that they may realistically encounter (see, for example, Chan & Schmitt, 1997). However, it is also possible that applicants would view the SJT negatively if any of the specific SJT scenarios were viewed as implausible or potentially inappropriate because of how the individuals described in the scenarios are treated or portrayed (see Sullivan et al., 2019).

Additionally, given that research on SJTs in other contexts has shown some potential advantages of video-based SJTs beyond written SJTs, we sought to develop and evaluate a video-based version. Specifically, potential advantages of video-based SJTs over written SJTs that have been found in some contexts have included greater face validity perceptions (laboratory studies: Chan & Schmitt, 1997; Richman-Hirsch et al., 2000), further reduced adverse impact on historically underrepresented groups (Chan & Schmitt, 1997), and improved validity for predicting interpersonally oriented performance criteria (Sackett & Lievens, 2006). Notably, not all studies have shown more positive applicant reactions to video-based SJTs than to written SJTs, nor have video-based SJTs demonstrated greater criterion-related validity than written SJTs for all types of relevant performance criteria (Sackett & Lievens, 2006), highlighting the need for evaluation in the specific Air Force context.

Further, given that implementation of a video-based version would be logistically challenging and costly to implement (i.e., would require transitioning from paper-and-pencil administration to a computer-based testing platform), we additionally evaluated a script-based written version that could be administered as a paper-and-pencil test. We hoped that such a format might provide some potential advantages over a traditional written format by making the test more engaging for applicants and potentially provide greater fidelity by presenting verbatim dialogue rather than more abstract description of scenarios. Additionally, the script-based format reduced the reading grade level of the test, which some SJT research has shown to mitigate potential adverse impact (see Whetzel et al., 2020).

Development of script- and video-based SJT formats

Two industrial/organizational psychologists developed written scripts designed to correspond directly with each SJT scenario. The scripts specified the physical location for each scene (e.g., your supervisor's office, break room, etc.), the verbatim words spoken by each character present (e.g., YOU and YOUR SUPERVISOR), and any associated actions (e.g., LAUGHS, SLAMS DOOR LOUDLY, SITS DOWN NEXT TO YOU); certain scenarios included more than one scene, clearly specified on the script ("TWO WEEKS LATER"). While the standard SJT version was written at a Flesch-Kincaid grade level of 9.7, the written scripted format was written at a Flesch-Kincaid grade level of 6.2.

The video-based version was developed by an animation studio directly from the verbatim scripts with the dialogue portrayed by professional voice actors to portray distinct voice types of different characters. The relationship between characters (i.e., supervisor vs. supervisee, highly experienced vs. inexperienced coworker) was apparent from the dialogue in the written script itself (e.g., "Boss . . . " or "I know you've worked in this office for years . . . "). The animations portrayed military members (with a range of diverse skin tones) in uniform with rank insignia removed (rank information did not appear in the written versions to avoid disadvantaging civilian applicants without prior military knowledge). See Figure A3 for example screenshots. The response options presented to examinees remained identical. In the video-based version, the response options were presented in written form after the scenario video, although participants could opt to mouse-over the response options to hear each response option (A-E) read aloud (in a computer-generated voice) if desired.

Participants

As in Study 1, Basic Military Trainees served as a convenience sample to evaluate the test. Participants included 286 female and 300 male trainees. Overall 36.54% completed at least one year of college; 10.98% had completed four years of college. Participants typically identified as White (79.36%), Black/African-American (16.03%), and/or Asian (4.61%). Separately, 20.40% identified as Hispanic or Latino.

Procedure

Participants were randomly assigned to complete one of the three 25-scenario SJT formats: standard written (*N* = 192), script-based (*N* = 168), or video-based (*N* = 165). After the SJT, all participants completed a survey to gauge their reactions to the test. Participants also self-rated themselves on competencies the SJT was designed to assess (Communication, Decision Making and Management, Leading Others, and Displaying Professionalism). While we recognized that self-reported performance ratings have major limitations (i.e., inaccuracy due to lack of self-awareness, over-confidence, impression management, etc.), we nonetheless sought some initial (albeit highly imperfect) indication of whether the alternate response formats might show any promise for increased validity beyond the standard written format.

Survey measures

Examinee reactions

Participants completed an adapted version of applicable Bauer et al. (2001) applicant reaction scales: *Propriety* (extent to which questions avoid bias and are deemed fair and appropriate; for example, "The content of the Situational Judgment Test did not appear to be prejudiced"), *Job-Content Relatedness* (extent to which test appears to measure content relevant to the job situation; for example, "The content of the Situational Judgment Test was clearly related to the job of an Air Force officer"), and *Predictive Job-Relatedness* (extent to which test appears to be valid for the job, e.g., "I am confident the Situational judgment Test can predict how well an applicant would perform on the job as an Air Force officer"). These used a five-point Likert-type scale, anchored 5 = *Strongly Agree*, 4 = *Agree*, 3 = *Neither Agree Nor Disagree*, 2 = *Disagree*, and 1 = *Strongly Disagree*. Internal reliability of the reaction scales ranged from α = .836 to .858.

Self-rated officership competencies

Self-rated proficiency on the competencies the SJT was designed to assess was measured using a five-point Likert scale (1 = Very Low to 5 = Very High) with six to eight associated behavioral statements per competency ("Read each statement and indicate the level at which you see yourself"). These included: *Communication* (six items, α = .759, e.g., "Appropriately express thoughts and opinions"); *Decision-Making and Management* (eight items, α = .825, e.g., "Make sound decisions based upon facts and available evidence"); *Leading Others* (six items, α = .781, e.g., "Understand when to follow and when to lead"); and *Displaying Professionalism* (six items, α = .812, e.g., "Accept responsibility for own actions, regardless of potential consequences"). An overall aggregate self-rating across the officership competencies was also computed (26 items, α = .920). The behaviors were consistent with the specific behaviors identified in earlier officership job analysis (Lentz et al., 2009a) which were provided to SMEs as example competency behaviors when eliciting critical incidents during test development.

Table 6. Applicant reactions by SJT format.

Variable	Standard (1)		Video (2)		Script (3)		ANOVA		Comparison group	Cohen's *d*
	Mean	*SD*	Mean	*SD*	Mean	*SD*	*F*	*p*		
Job-related content	3.99	0.72	3.93	0.71	3.87	0.67	1.46	.234	1 vs. 2	0.08
									1 vs. 3	0.17
									2 vs. 3	0.09
Job-related prediction	3.08	0.84	3.01	0.80	3.03	0.82	0.33	.722	1 vs. 2	0.08
									1 vs. 3	0.06
									2 vs. 3	0.02
Propriety	4.24	0.89	4.25	0.84	4.20	0.74	0.21	.810	1 vs. 2	−0.02
									1 vs. 3	0.05
									2 vs. 3	0.07

Standard *N* = 204; Video *N* = 189 (188 for Propriety); Script *N* = 192.

Results

Psychometric properties of alternate test formats

SJT difficulty was similar across the three versions. Applying the Motowidlo et al. (1990) scoring method, mean scores were .542 (*SD* = .146) for the standard written version, 0.492 (*SD* = .178) for the script-based version, and .533 (*SD* = .138) for the video-based version. Item-level means for the standard written version correlated .93 with item means for the script-based version and .92 for the video-based version. Internal reliability of the script-based version (α = .754) was somewhat higher than observed for either the standard written (α = .663) or video-based version (α = .627) in this study.

Administration time

Overall, the standard written format was substantially less time-consuming to administer than either the script-based or video-based format [*F*(488) = 79.70, *p* < .001]. Mean time to complete the 25-scenario SJT was 38.42 minutes (*SD* = 9.22) for the standard written version, 45.38 minutes (*SD* = 5.60) for the video-based version, and 48.89 minutes for the script-based version (*SD* = 7.97).

Relationship with officership ratings

Despite the major limitations of any self-reported performance criteria, we found a modest significant relationship of scores on the standard written SJT to self-ratings of Communication and Decision Making (*r*s = .162 and .193, respectively). In contrast, there was no evidence of any significant relationship of self-ratings to script-based or video-based SJT scores. Aggregating across the four competencies, overall officership competency self-ratings were correlated .150 with standard written SJT scores (*p* < .05), .076 with script-based SJT scores, and .075 with video-based SJT scores.

Applicant reactions

There were no significant differences by format for any of the applicant reaction measures [*Job-Related Content: F*(583) = 1.46, *p* = .234; *Job-Related Prediction: F* (583) = 0.33, *p* = .722; *Propriety: F*(582) = 0.21, *p* = .810]. Regardless of format, the SJT was endorsed as fair and appropriate, with a mean Propriety score of 4.24 (*SD* = .89) for the standard written version. The content was also endorsed as highly job-relevant (standard written version: *M* = 3.99, *SD* = 0.72). Ratings of perceived job-related prediction were more neutral or mixed. See Table 6.

Discussion

Overall, results found no evidence that would strongly favor use of a video-based or script-based version over use of the standard written version. While the script-based SJT showed somewhat greater internal reliability than either the standard written or video-based version, the script-based format was also the most time-consuming to administer – taking approximately 10 minutes longer to administer a 25-scenario SJT than the standard written version. As such, gains in internal reliability could likely be achieved simply by increasing the length of the standard written SJT. Additionally, while neither the script-based nor video-based format related to self-ratings on officership competencies, scores on the standard written SJT format were significantly related to two of the (more cognitively-oriented) competencies the SJT was designed to assess: Communication and Decision Making. While self-ratings are certainly not a preferred criterion measure, this pattern of findings did not lead us to expect that a video-based or script-based format would markedly improve SJT validity.

Importantly, participants generally had favorable reactions to all three SJT formats. SJTs introduce interpersonal content that (unlike the generic content of AFOQT quantitative or spatial tests) may be sensitive and, if not developed carefully, may be viewed as potentially inappropriate because of how the individuals described in particular scenarios are treated or portrayed. Our results affirmed that the SJT content was viewed as fair and appropriate, with the vast majority of participants endorsing positive agreement rather than neutral ratings on the propriety of the test (i.e., "the content of the test seemed appropriate," and "did not appear to be prejudiced"). Further, although the participants were not privy to the details of how the SJT was developed, they recognized the content as strongly job-related (i.e., "it would be obvious to anyone that this test is related to the job of an Air Force officer").

While this study did not include a sufficiently large sample of African-American or Hispanic participants to evaluate potential differences in adverse impact associated with the standard written vs. video-based version, previous research has suggested some evidence that video-based SJTs can reduce adverse impact beyond written ones (e.g., Chan & Schmitt, 1997). Although practical considerations (i.e., the need for computer-based testing platform to administer a video-based SJT) ultimately precluded further evaluation of a video-based SJT, we hope that this may be an area for future exploration if the Air Force moves to a computer-based testing platform in the future.

Study 4 criterion-related validation in an army officer cadet sample

Given that the initial sample sizes available to evaluate criterion-related validity within a US Air Force officer cadet sample were quite small, we coordinated to experimentally administer the SJT, along with the AFOQT Verbal and Quantitative subtests, to a much larger Army ROTC sample – comprised of cadets who participated in the 2016 Army ROTC Cadet Leadership Course.

Method

Overall, 4,907 Army ROTC cadets were administered the AFOQT SJT, Verbal (Reading Comprehension, Word Knowledge, Verbal Analogies), and Quantitative subtests (Arithmetic Reasoning, Math Knowledge) on an experimental basis during the annual Cadet Leadership Course in Fort Knox, Kentucky (individual results did not affect any career outcomes). The Cadet Leadership Course is a prerequisite to become an Army officer through ROTC, with most cadets attending in the summer between their junior and senior year of college after having contracted to join the Army. By gender, 21.8% of the Army sample was female. Overall, 8.9% of the sample identified as Hispanic/Latino; separately, 7.4% identified as Asian and 11.7% identified as African-American. Matched criterion data was available for 3,589 cadets who were scheduled to become commissioned officers in 2017.

Criteria

The analysis evaluated the Army ROTC Order of Merit Score (OMS), which is intended as an indicator of future officer in-unit performance, and used to assign cadets to branch preferences. The primary OMS components are undergraduate Grade Point Average (GPA), Combined Performance Measurement Systems (PMS) Assessments, and the Army Physical Fitness Test (APFT) (U.S. Army Cadet Command, 2020). Of the OMS components, only the PMS assessments are intended to measure officership competencies or experiences that bear similarity to those that the SJT was designed to assess. Inputs to the aggregate PMS score include: (a) detachment commander ratings of potential and performance (ROTC CDT CMD Form 67–10-1), (b) Platoon Tactical Officers' ratings based on performance in Advanced Camp at Fort Knox (CC Form 1059), and (c) additional points for completion of various training opportunities, extracurricular activities, and paid work experience while in college. Factors that detachment raters are directed to consider include: cadet Character, Presence, Intellect, Leadership, Development of Others, and Achievement.

Results

The analyses found the SJT demonstrated useful criterion-related validity as an indicator for ROTC OMS and undergraduate GPA. For OMS, the validity of the SJT (r = .24) was comparable to the AFOQT Quantitative (r = .25) and Verbal (r = .31) composites. When the SJT was combined with the Verbal and Quantitative composites, validity for the OMS criterion modestly, but significantly, increased from r = .326 to .337 (see regression results in Table 7). For GPA, the validity of the SJT (r = .19) was higher than the AFOQT Quantitative composite (r = .14) and more comparable to the Verbal composite (r = .23). When combining the SJT with the Verbal and Quantitative composites, validity for the GPA criterion significantly increased from r = .235 to .250 (see regression results in Table 7).

Study 5 subgroup differences on operational AFOQT Form T

Between 2015 and 2020, over 70,000 individuals have taken the operational version of the AFOQT Form T (T1 or T2). These include AFROTC and USAFA cadets, who typically take the AFOQT in their sophomore or junior year, civilian applicants for Officer Training School, and active duty enlisted Air Force members who may apply for officer accessioning. Examinees

Table 7. Incremental validity of SJT over AFOQT verbal and AFOQT quantitative scores for ROTC order of merit scores and grade point average (study 4).

Model	Variable	β	b	SE	Model R^2	Adj. R^2
Criterion: OMS						
Model 1	AFOQT V	.244***	.111	.008	.1062	.1057
	AFOQT Q	.127***	.064	.009		
Model 2	AFOQT V	.256***	.116	.008	.1030	.1025
	SJT	.108***	9.18	1.542		
Model 3	AFOQT Q	.199***	.100	.008	.0880	.0875
	SJT	.170***	14.51	1.42		
Model 4	AFOQT V	.201***	.091	.009	.1138	.1131
	AFOQT Q	.120***	.060	.009		
	SJT	.099***	8.45	1.52		
Criterion: GPA						
Model 1	AFOQT V	.214***	.043	.004	.0553	.0547
	AFOQT Q	.039*	.009	.004		
Model 2	AFOQT V	.186***	.038	.004	.0619	.0613
	SJT	.099***	3.78	.696		
Model 3	AFOQT Q	.099***	.022	.004	.0438	.0432
	SJT	.158***	6.01	.648		
Model 4	AFOQT V	.172***	.035	.004	.0626	.0618
	AFOQT Q	.031*ns*	.007	.004		
	SJT	.097***	3.69	.697		

N = 3587. ****p* < .0001; **p* < .05 (two-tailed).

Table 8. *Subgroup differences on operational AFOQT form T SJT* (*N* = 71,225).

	AFOQT SJT	Verbal composite	Quantitative composite	Physical science	Spatial tests	Perceptual speed	Aviation information
Black	0.58	0.95	0.85	0.87	1.03 (BC) 1.15 (IC)	0.82	0.88
Hispanic	0.22	0.40	0.35	0.28	0.22 (BC) 0.28 (IC)	0.26	0.34
Asian	0.34	0.49	−0.26	0.11	0.15 (BC) 0.40 (IC)	0.10	0.54
Female	0.02	0.29	0.46	0.66	0.49 (BC) 1.08 (IC)	0.15	0.81

All differences are statistically significant, *p* < .05.

have included 18,934 female applicants (52,183 males; 108 unreported); 10,683 applicants identifying as Hispanic or Latino (59,021 non-Hispanic; 1,521 unreported); 7,702 applicants identifying as African-American and 3,983 applicants identifying as Asian.

Analysis of the 2015–2020 AFOQT Form T data (*N* = 71,225) indicated the SJT score differences were small to medium in magnitude for Hispanic/Non-Hispanic and White/Asian comparisons (Cohen's *d* = .22 and .34, respectively), generally in line with findings for other AFOQT subtests. Moderate SJT score differences were observed in White vs Black/African-American comparison (Cohen's *d* = .58), substantially lower than the score differences observed for other AFOQT Verbal and Quantitative subtests. Notably, although female examinees score somewhat lower than male examinees on all traditional cognitive AFOQT subtests, gender differences on the SJT were extremely minimal (*d* = .02 favoring male applicants). In summary, compared to the subgroup differences across the cognitive composites on the AFOQT Form T, the SJT shows substantially less adverse impact on historically underrepresented groups overall. See Table 8.

Our findings are generally in line with broader meta-analytic results across other populations, although the observed Black/White score difference for Air Force applicants (*d* = .58) was slightly higher than meta-analytic evidence has found in other settings (*d* = .38; k = 62; *N* = 42,178; Whetzel et al., 2008). Similarly, meta-analytic evidence in other settings shows that women often obtain slightly higher SJT scores than men (*d* = –.11; k = 63; *N* = 37,829; Whetzel et al., 2008). These slight differences may be a function of many factors given that prior meta-analysis aggregated across (a) student, incumbent, and applicant samples, (b) video-based and written formats, and (c) SJTs designed to assess distinct constructs. More active US Air Force recruitment of members of historically underrepresented groups than in other organizations may also affect these results.

Conclusion and next steps

As described, the AFOQT SJT was developed over a multi-phase process, based on input from over 4,000 Air Force officers to identify core officership competencies that should be assessed when screening officer candidates (Lentz et al., 2009a), and which would be amenable to measurement using an SJT (Lentz et al., 2009b). The SJT development process employed the critical incidents method to identify applicable scenarios, drawing from a combined sample of Air Force officers and novices (newly enlisted Basic Military Trainees) to identify plausible response options that appropriately vary in effectiveness. To establish a scoring key that could withstand scrutiny, Air Force officers who had been identified as Distinguished Graduates when competing in a common environment with other O-3s rated response effectiveness, and only response options that differed significantly in effectiveness based on these SME ratings were retained.

Development efforts to date have affirmed the potential benefits of including the SJT on the AFOQT. Studies 1 and 2 demonstrated generally appropriate psychometric characteristics when administered to either enlisted or officer cadet samples. Initial item vetting suggested that, consistent with the broader SJT literature, the USAF SJT was likely to result in reduced adverse impact toward females and African-Americans relative to traditional cognitive tests (i.e., ASVAB subtests). As shown in Study 3, Air Force members perceived the SJT as highly content-valid and appropriate (i.e., not viewed as discriminatory or biased). Most notably, the large-scale results from the operational version of the AFOQT Form T SJT clearly affirm a decrease in adverse impact when compared to the AFOQT verbal and math cognitive tests. Predictive validation studies with larger Air Force officer accession samples are ongoing to assess the incremental validity of the SJT beyond current AFOQT composites for predicting important outcomes across accession sources.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The data that support the findings of this study are available from the corresponding author, Dr. Imelda D. Aguilar, upon reasonable request.

References

Bauer, T. N., Truxillo, D. M., Sanchez, R. J., Craig, J. M., Ferrara, P., & Campion, M. A. (2001). Applicant reactions to selection: Development of the selection procedural justice scale (SPJS). *Personnel Psychology*, *54*(2), 387–419. https://doi.org/10.1111/j.1744-6570.2001.tb00097.x

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, *14*(3), 223–235. https://doi.org/10.1111/j.1468-2389.2006.00345.x

Bruns, J. W., & Eichorn, L. A. (1993). *A comparison of non-performance characteristics with United States air force officer promotions* [Masters thesis]. Air Force Institute of Technology. https://apps.dtic.mil/sti/pdfs/ADA273967.pdf

Campion, M. C., Ployhart, R. E., & MacKenzie, W. I. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance*, *27*(4), 283–310. https://doi.org/10.1080/08959285.2014.929693

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

Corstjens, J., Lievens, F., & Krumm, S. (2017). Situational judgement tests for selection. In H. W. Goldstein, E. D. Pulakos, J. Passmore,& C. Semedo (Eds.),*The Wiley Blackwell Handbook of the Psychology of Recruitment, Selection and Employee Retention* (pp. 226–246). Wiley-Blackwell. https://doi.org/10.1002/9781118972472.ch11

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, *51*(4), 327–358. https://doi.org/10.1037/h0061470

Holm Center. (2015). *AFROTC field training manual (Holm Center T-203)*. http://www.fresnostate.edu/craig/depts-programs/air-force/documents/2015%20FTM.pdf

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. *Internal Journal of Selection and Assessment*, *9*(1&2), 152–194. https://doi.org/10.1111/1468-2389.00171

Kantrowitz, T., Kingry, D., Madaj, C., & Nye, C. (2019). *Navy computerized adaptive personality scales (NCAPS) and self description inventory (SDI) wind down and merger with tailored adaptive personality assessment system (TAPAS)* AFRL-RH-WP-TR-2019-0095. Wright-Patterson AFB, OH: 711 Human Performance Wing, Warfighter Interface Division, Collaborative Interfaces and Teaming Branch. https://apps.dtic.mil/sti/pdfs/AD1091172.pdf

Knapp, D. J., Campbell, C. H., Borman, W. C., Pulakos, E. D., & Hanson, M. A. (2001). Performance assessment for a population of jobs. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification* (pp. 181–235). Erlbaum.

Lentz, E., Horgen, K. E., Borman, W. C., Dullaghan, T. R., & Smith, T. (2009a). *Air force officership survey volume I: Survey development and analyses*. PDRI.

Lentz, E., Horgen, K. E., Borman, W. C., Dullaghan, T. R., & Smith, T. (2009b). *Air force officership survey volume II: Performance requirement linkages and predictor recommendations*. PDRI.

Lievens, F., & Coetsier, P. (2002). Situational tests in student selection: An examination of predictive validity, adverse impact, and construct validity. *International Journal of Selection and Assessment*, *10*(4), 245–257. https://doi.org/10.1111/1468-2389.00215

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology*, *60*(1), 63–91. https://doi.org/10.1111/j.1744-6570.2007.00065.x

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, *75*(6), 640–647. https://doi.org/10.1037/0021-9010.75.6.640

Motowidlo, S. J., & Tippins, N. (1993). Further studies of the low-fidelity simulation in the form of situational inventory. *Journal of Occupational and Organizational Psychology*, *66*(4), 337–344. https://doi.org/10.1111/j.2044-8325.1993.tb00543.x

Pulakos, E. D., & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance*, *9*(3), 241–258. https://doi.org/10.1207/s15327043hup0903_4

Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology*, *85*(6), 880–887. https://doi.org/10.1037/0021-9010.85.6.880

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

Sackett, P. R., & Lievens, F. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology*, *91*(5), 1181–1188. https://doi.org/10.1037/0021-9010.91.5.1181

Sullivan, T. S., Burgoyne, T. C., McCloy, R. A., & Whetzel, D. L. (2019). *Review of situational judgment test (SJT) prototype development process (AFRL-RH-WP-TR-2019-0058)*. Wright-Patterson AFB, OH: 711 Human Performance Wing, Warfighter Interface Division, Collaborative Interfaces and Teaming Branch. https://apps.dtic.mil/sti/pdfs/AD1083950.pdf

U.S. Air Force Academy. (2020). *Military performance appraisal (USAFI 36-2401)*. https://static.e-publishing.af.mil/production/1/usafa/publication/usafai36-2401/usafai36-2401.pdf

U.S. Army Cadet Command. (2020). *Reserve officers' training corps accessions fiscal year 2021 (USACC circular 601-21-1)*. https://www.cadetcommand.army.mil/res/files/forms_policies/circulars/USACC%20Circular%20601-21-1.pdf

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology*, *52*(3), 679–700. https://doi.org/10.1111/j.1744-6570.1999.tb00176.x

Weekley, J. A., & Plolyhart, R. E. (2006). An introduction to situational judgment testing. In J. A. Weekly & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 2–3). Lawrence Erklbaum Associates, Publishers.

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, *21*(3), 291–305. https://doi.org/10.1080/08959280802137820

Whetzel, D. L., Sullivan, T. S., & McCloy, R. A. (2020). Situational judgment tests: A n overview of development situational judgment tests. *Personnel Assessment and Decisions*, *6*(1), 1–16. https://doi.org/10.25035/pad.2020.01.001

Appendix. A1. SME Ratings of Response Options for Example SJT Scenarios (Items Excluded from AFOQT)

SCENARIO 1: You have recently been assigned to lead a section comprised of experienced subordinates, but you do not have a full understanding of the mission and tasks. Your subordinates are not helpful when you solicit ideas and information from them. It is necessary for you to understand your job and the other section members' jobs in order to effectively lead your section and accomplish the mission.

	Mean effectiveness rating (*N* = 106)	SD
Contact the superior who assigned you to the section for further guidance.	3.79	1.26
Contact the individual previously assigned to the section for guidance.	5.30	0.92
Meet privately with the most senior subordinate to discuss the section's mission.	5.75	0.73
Meet individually with each subordinate to get to know them personally.	5.75	0.78
Shadow your subordinates' work efforts to see what they do.	5.42	1.13
Call a section meeting, and emphasize that you need everyone's cooperation in order to help the section succeed.	4.81	1.20

SCENARIO 2: Your commander will be deploying soon and you will be taking temporary command of your unit. At the next roll call meeting you are to inform your unit that the commander will be transferring his command authority to you.

	Mean effectiveness rating (*N* = 105)	SD
Ask the outgoing commander to make the announcement.	5.13	1.34
Prepare a written message, and ask the outgoing commander to provide input prior to the announcement.	3.95	1.37
Explain the accomplishments of the outgoing commander, and emphasize that you will carry on his vision of the unit.	4.87	1.26
Explain that you will be relying on unit members for advice and guidance.	4.37	1.30
Explain that you are confident in your ability to lead because you are confident in the team.	5.24	1.27
Explain the need for flexibility during this time of adjustment.	4.19	1.27

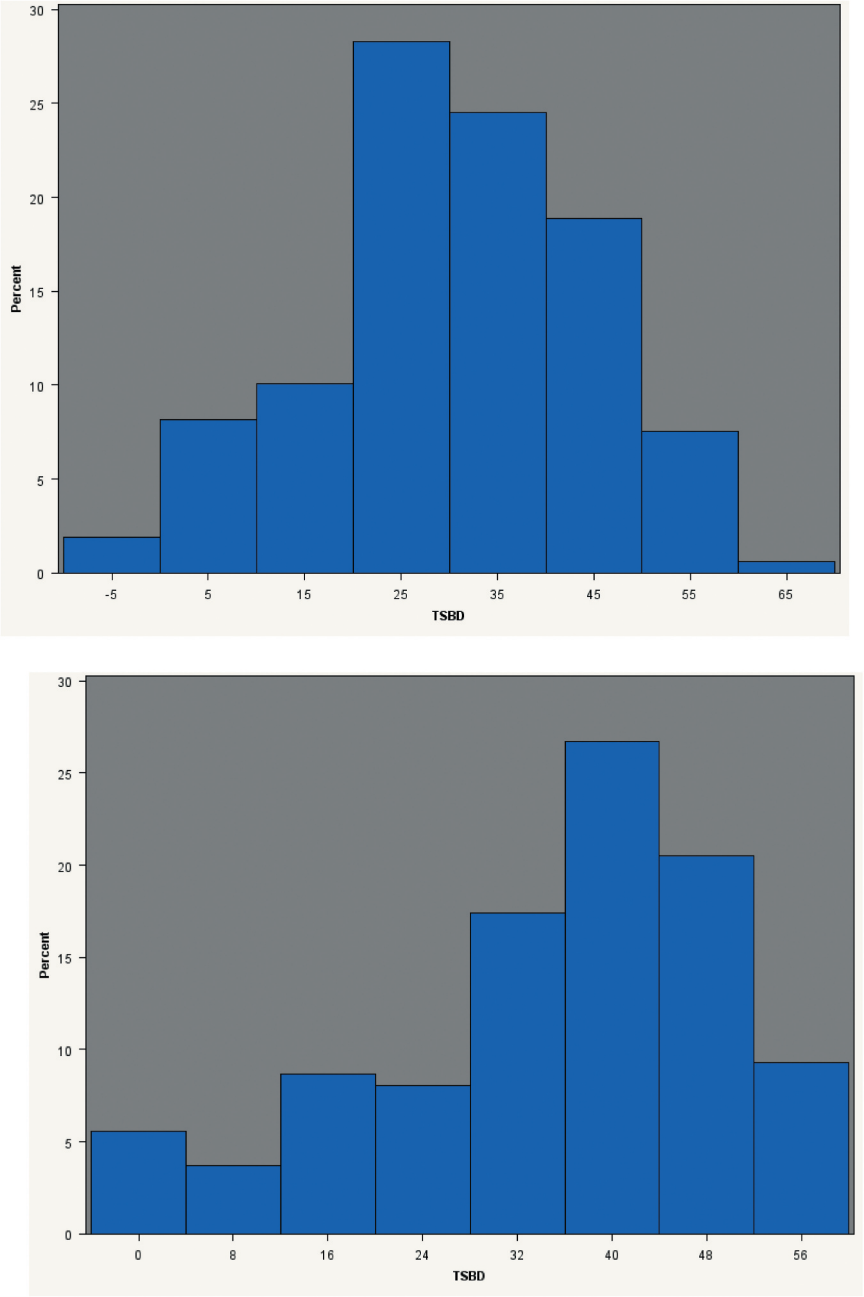

Figure A2. Distributions of study 1 SJT scores (versions 1 and 2, respectively).

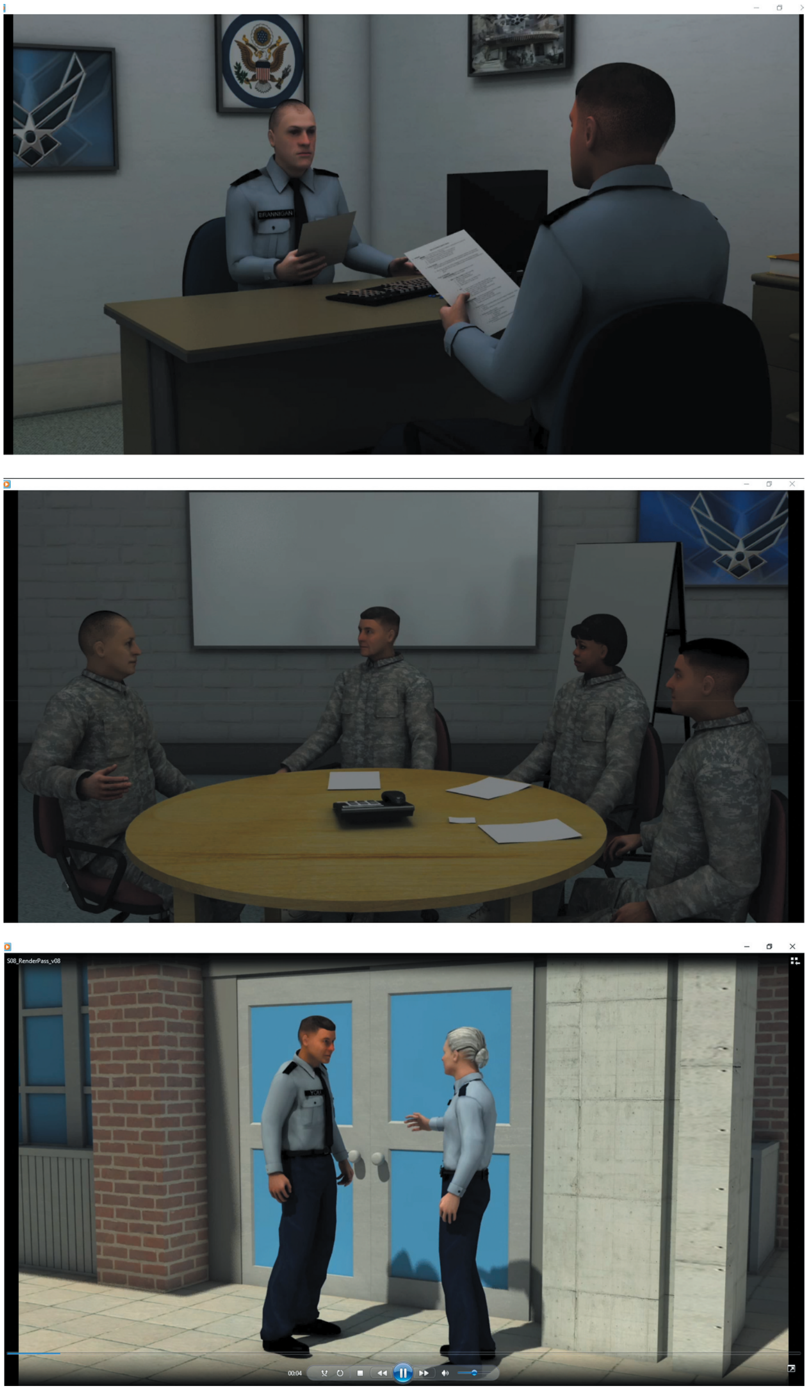

Figure A3. Example video-based SJT screenshots.