AFRL-RH-WP-TR-2019-0058

REVIEW OF SITUATIONAL JUDGMENT TEST (SJT) PROTOTYPE DEVELOPMENT PROCESS AND MATERIALS

Taylor S. Sullivan
Timothy C. Burgoyne
Rodney A. McCloy
Deborah L. Whetzel

Human Resources Research Organization (HumRRO)

August 2019
Interim Report

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
AIRMAN SYSTEMS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th Air Base Wing Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2019-0058 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//signature///
THOMAS R. CARRETTA
Work Unit Manager
Collaborative Interfaces and Teaming Branch
Warfighter Interfaces Division

//signature//
TIMOTHY S. WEBB
Chief, Collaborative Interfaces and
Teaming Branch
Warfighter Interfaces Division

//signature//
LOUISE A. CARTER
Chief, Warfighter Interfaces Division
Airman Systems Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YY)*
08-19-19

2. REPORT TYPE
Interim

3. DATES COVERED *(From - To)*
10- JAN 18 – 19 APR 19

4. TITLE AND SUBTITLE
Review of Situational Judgment Test (SJT) Prototype Development Process and Materials

5a. CONTRACT NUMBER
FA8650-14-D-6500/FA8650-18-F-6828

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
Taylor S. Sullivan, Timothy C. Burgoyne, Rodney A. McCloy, and Deborah L. Whetzel

5d. PROJECT NUMBER
5329

5e. TASK NUMBER
03

5f. WORK UNIT NUMBER
H0SA (532909TC)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Human Resources Research Organization (HumRRO)
66 Canal Center Plaza, Suite 700
Alexandria, VA 22314-1578

8. PERFORMING ORGANIZATION REPORT NUMBER
2018 No. 061

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Material Command
Air Force Research Laboratory
711th Human Performance Wing
Airman Systems Directorate
Warfighter Interface Division
Collaborative Interfaces & Teaming Branch
Wright-Patterson AFB, OH 45433

Infoscitex Corporation
4027 Colonel Glenn Highway
Suite 210
Dayton, OH 45431

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
711 HPW/RHCC

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RH-WP-TR-2019-0058

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release. 88ABW-2019-4768. Cleared on 23 October 2019

13. SUPPLEMENTARY NOTES
Subcontract number: FPH02-S022/180123

14. ABSTRACT

The U.S. Air Force (USAF) is exploring the use of a situational judgment test (SJT) as a potential augmentation to the Weighted Airman Promotion System (WAPS). SJTs are frequently used to assess complex relational skills, such as interpersonal skills and leadership (e.g., Christian, Edwards, & Bradley, 2010). The Air Force Personnel Center Strategic Research and Assessment Branch (AFPC/DSYX) developed a prototype SJT for use with E-7 (master sergeant) candidates to assess People/Team Competencies listed in Air Force Doctrine Annex 1-1, Force Development (AFPD 36-26; U.S. Air Force, 2015). Administered thus far on an experimental basis only, the USAF requires an evaluation of the materials and procedures used to develop the SJT to determine improvements that might be made before moving forward with additional feasibility studies.
(continued on next page)

15. SUBJECT TERMS
Situational judgment test, SJT, test development, sensitivity review, scoring key, interrater reliability, interrater agreement.

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT:**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** (Monitor)
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Thomas R. Carretta
Unclassified	Unclassified	Unclassified	SAR	24	**19b. TELEPHONE NUMBER** *(Include Area Code)* N/A

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18

14. Abstract

This report summarizes the Human Resources Research Organization's review of the prototype SJT. The primary objectives were to: (a) review the content of the prototype SJT and associated test development materials/protocols, (b) conduct a sensitivity review at the item level to identify items with potentially problematic content, and (c) examine initial subject matter expert (SME) effectiveness ratings for each response option used in scoring key development, including an examination of interrater reliability and agreement of SME ratings. Included is a summary of our findings stemming from this review, as well as associated recommendations to consider moving forward.

TABLE OF CONTENTS

List of Tables

List of Figures

1.0 OVERVIEW

The U.S. Air Force (USAF) is exploring the use of a situational judgment test (SJT) as a potential augmentation to the Weighted Airman Promotion System (WAPS). SJTs are frequently used to assess complex relational skills, such as interpersonal skills and leadership (e.g., Christian, Edwards, & Bradley, 2010). The Air Force Personnel Center Strategic Research and Assessment Branch (AFPC/DSYX) developed a prototype SJT for use with E-7 (master sergeant) candidates to assess People/Team Competencies listed in Air Force Doctrine Annex 1-1, Force Development (AFPD 36-26; U.S. Air Force, 2015). Administered thus far on an experimental basis only, the USAF requires an evaluation of the materials and procedures used to develop the SJT to determine improvements that might be made before moving forward with additional feasibility studies.

The goal of this report is to summarize the Human Resources Research Organization's (HumRRO) review of the prototype SJT. Our review had three primary objectives:

- review the content of the prototype SJT and associated test development materials/protocols,
- conduct a sensitivity review at the item level to identify items with potentially problematic content, and
- examine initial subject matter expert (SME) effectiveness ratings for each response option used in scoring key development, including an examination of interrater reliability and agreement of SME ratings.

Included is a summary of our findings stemming from this review, as well as associated recommendations to consider moving forward.

2.0 FOUNDATION FOR SJT DEVELOPMENT

Due to reductions in the size of the USAF and its operations becoming more connected with other military Services, leadership responsibility and authority will increasingly be pushed down to noncommissioned officers (NCOs). A recent RAND report emphasized the critical importance of developing leadership ability among the first level of senior noncommissioned officer (SNCO) leadership, E-7 master sergeants (Keller et al., 2014). In addition to ongoing training and development efforts, implementing a promotion system that selects for the knowledge, skills, abilities and other characteristics (KSAOs) needed to fulfill these new leadership roles will be critical in identifying Airmen with the greatest potential for success as they progress from the rank of E-6 (technical sergeant) to the rank of E-7 (master sergeant).

The WAPS initially was developed more than 40 years ago and has remained largely unchanged. The RAND report evaluated how well the current WAPS is assessing the KSAOs needed for success at the E-7 level and recommended the addition of an SJT and a promotion board to improve current promotion practices.

The RAND report enumerated the duties and KSAOs required at the E-7 level and assessed their criticality by conducting a literature review and interviews, rather than conducting a full job analysis. The literature review was on-target; it included well-known review articles and seminal works on leadership. The KSAO categories identified (i.e., cognitive skills, interpersonal skills, business/management skills, strategic skills, personality, and motivation) are consistent with most prevailing models of leadership, and generally align with the USAF Institutional Competency model. The interviews were conducted with two senior enlisted leader advisory bodies and a small sample of wing, group, and squadron commanders (N = 41). Important duty areas and KSAOs were identified by summarizing the percentage of interviewees who mentioned the duty or KSAO during the interview.

- **HumRRO Recommendation:** In the future, additional survey research or SME focus groups could be conducted to formally document the linkages between the competencies in the USAF Institutional Competency model and the duty areas (or better yet, specific tasks) identified in the RAND study. This linkage process would provide key support for the claim that behaviors and situations simulated in an assessment such as the SJT do require target competencies to perform them. This builds evidence for the validity and job-relatedness of the assessment.

3.0 INTENDED USE OF SJT

As stated above, the RAND study evaluated the extent to which the WAPS measures competencies required for effective leadership at the E-7 level and recommended that an SJT be developed to aid in the selection of Airmen with the greatest leadership potential from the technical sergeant (E-6) level to the master sergeant (E-7) level. Thus, scenarios were drafted to be representative of performance at the E-7 level. Based on discussions with USAF staff, our understanding is that it would be desirable to use the SJT to assess competencies for promotion into the E-6 level in addition to or, more likely, instead of into the E-7 level. If this is the case, it will be important to establish evidence for the assessment's validity for the E-6 level, as the scenarios in the SJT prototype originally were drafted to reflect the E-7 level.

- **HumRRO Recommendation:** During future item development efforts, content developers and SMEs participating in activities should be aware that SJT items should be appropriate for the target promotion level, whichever level(s) that may be. During the SJT prototype scoring key development process (described below), SMEs rated the relevance of each scenario for E-7s in their Air Force Specialty Code (AFSC) as well as for E-7s across the Air Force. Therefore, to bolster support for use of the SJT at lower ranks, a similar exercise would need to be conducted to confirm relevance of these prototype items to the technical sergeant level, as well. Because behavioral expectations and performance requirements shift from one level to another, the effectiveness of different behaviors or responses to situations may vary across level. For example, making an independent decision may be a very effective behavior at upper levels of leadership, whereas at lower levels, it may be ill advised to make decisions without consulting peers or superiors for guidance. Therefore, it would be advisable to review and revise the scoring key as needed to articulate a standard that is representative of the target promotion level (i.e., E-6 in the case this is used for promotion to technical sergeant).

4.0 TARGET COMPETENCIES

AFMAN 36-2647 (U.S. Air Force, 2016) provides guidance on how institutional competencies are established, assessed, and used in support of the USAF mission. According to this document, "the vision for Institutional Competency (IC) development is to create the appropriate strategies, policies, and processes required to prepare all Airmen with the appropriate leadership expertise to accomplish assigned airpower missions." The USAF ICs apply to all Airmen and serve as a common framework for understanding the leadership KSAOs required to meet the challenges of the Air Force's dynamic operating environment. The ICs are considered observable, measurable patterns of KSAOs and behaviors needed to perform institutional or occupational functions successfully. The ICs consist of eight competencies clustered into three categories: Organizational, People/Team, and Personal. The competencies are further broken down into 25 sub-competencies. Each sub-competency is further described according to basic, intermediate, proficient, skilled, and advanced levels of proficiency. The SJT prototype was developed to measure five sub-competencies within the People/Team category: "Develops and Inspires Others," "Takes Care of People," and "Diversity" (in the Leading People competency), as well as "Builds Teams and Coalitions" and "Negotiating" (in the Fostering Collaborative Relationships competency).

- **HumRRO Recommendation:** It may be advisable to revisit the decision-making process regarding which competencies the SJT should target to confirm that selected competencies are indeed amenable to measurement via an SJT. Our experience suggests that the Diversity competency can sometimes be overly transparent to measure using an SJT. Critical incidents may bear this out. For example, one incident is about an NCOIC who complains openly about third-country contractors being uneducated and lazy. The "correct" action is to intervene. The challenge would be to develop plausible-yet-incorrect answers (e.g., defer to a supervisor, ignore the problem, talk to a more experienced colleague about what to do). None of these alternatives seems close in effectiveness to the correct answer of intervening.

  On the other hand, it may also be helpful to revisit which competencies are currently *not* being targeted for measurement with the SJT. For example, judgment and decision making (Decision Making is under the Strategic Thinking competency) and self-development (Develops Self is under the Embodies Airman Culture competency) are competencies that are commonly assessed using SJTs. Thus, it may be possible to expand the competency coverage. One first step would be to experiment with critical incident documentation and/or to review the tasks/duties that are typically linked to the behaviors in these competencies to see if they may be amenable to simulation in the context of an SJT, while not being overly transparent in terms of how to respond.

- **HumRRO Recommendation:** In addition, several of the competencies seem to conceptually overlap and may appear even more overlapping when manifested in the context of an SJT scenario. For example, in a situation in which a supervisor needs to coach a subordinate on how to relate to another team member, it may be unclear whether the situation measures Develops and Inspires Others (because coaching and leading are involved), Takes Care of People (because coaching and interpersonal skills are involved), or Builds Teams and Coalitions (because ultimately the supervisor is improving relationships among the team). Therefore, if it does not already exist, we recommend

articulating a clear schema for the types of behaviors that do and do not fall under each target competency, paying close attention to areas of potential conceptual overlap. Considering this schema during item authoring and revision as well as during scoring key development could lead to cleaner measurement of target competencies. That said, a great deal of literature has supported SJTs as being multidimensional, and in both research and practice, it is difficult to produce an interpretable multiple factor structure underlying SJTs (e.g., Ployhart, 2006). Thus, we often caution against reporting scores or candidate results at the competency level, even if the SJT initially was developed to measure distinct competencies.

- **HumRRO Recommendation:** AFMAN 36-2647 describes in-role competency assessments being used for professional development and competency gap analysis purposes. It may be helpful to explore the types of internal competency assessments being administered by the Institutional Competency Development Programs (ICDPs) and determine if (a) these relate to or overlap with content featured on the SJT, or (b) they contain content that would be appropriate to feature on the SJT. In addition, documentation suggests that the virtual Force Development Center, located on the Air Force Portal site, is a clearinghouse of leadership development resources. If these resources were organized by competency or subcompetency, or linked to the development of certain competencies, it may be helpful to determine if the materials contain any vignettes or content developed to define behaviors and/or situations associated with the competencies and sub-competencies being targeted by the SJT. Well-recognized resources from these entities or programs could be helpful source materials, references, or points of comparison when developing SJT content.

5.0 PROTOTYPE SJT DEVELOPMENT PROCESS

When developing the prototype SJT for use with E-7 (master sergeant) candidates, AFPC/DSYX used a multi-part development process comprising three phases: (a) critical-incident gathering, (b) item development (which included a content validation process), and (c) scoring key development. We describe each step of the development process below and provide associated recommendations for future SJT development efforts.

5.1 Phase I: Critical Incident Gathering

The critical incident technique (CIT) is a set of procedures for collecting observations of behavior that have some critical significance and meet some pre-defined criteria (Flanagan, 1954). The incidents gathered can be used for a variety of purposes (e.g., solve practical problems, detect procedural errors, develop psychological principles), and they are often used as a starting point in behavioral assessment development. Critical incidents can be gathered in various ways, but the process typically involves having respondents describe experiences in a structured format. Two of the most common response formats are the STAR (Situation, Task, Action, and Result) and ABC (Antecedent, Behavior, and Consequence) methods.

During prototype USAF SJT development, Phase I consisted of gathering critical incidents. Using the STAR method, Airman Advancement Division (AAD) development teams (i.e., SMEs assigned to develop Specialty Knowledge Test/Performance Fitness Exam [SKT/PFE] exams) independently generated critical incidents as well as alternative effective and ineffective actions that might have been taken in response to the incident. The SMEs then shared their critical incidents with the group, and other group members generated additional effective and ineffective alternative actions. Finally, the AFPC/DSYX team aggregated the critical incidents by dimensions and used the incidents as a starting point to generate draft SJT items.

The process described above is consistent with best practices for use of the critical incident technique (CIT). Based on our experience, HumRRO has recently adopted several additional practices that could be considered for inclusion in future critical incident workshops:

- **HumRRO Recommendation:** One of the most common issues we have experienced when collecting critical incidents is that many of the situations documented ultimately do not involve judgment, but rather rely almost exclusively on knowledge of a certain policy or procedure to resolve the situation. That is, there may not be enough complexity involved to be able to generate numerous potential actions that one could take to address the situation. One strategy we have used is to add a section on the critical incident form that asks SMEs to explicitly identify the judgments or decisions they had to make to address the challenging situation. Explicitly stating that this is a feature of the types of situations you are looking for can generally improve the quality of the incidents in terms of richness and level of complexity.
- **HumRRO Recommendation:** One of the primary goals of the CIT is to gather examples of situations that are *realistic* and *representative*. We have found that using the term "critical" to describe the types of situations we are looking for can sometimes give SMEs the impression they should document only the rarest or extraordinary circumstances they have experienced in their careers. To the contrary, having SMEs focus on more commonly encountered experiences, as long as they are challenging and required judgment, can help ensure that the types of situations documented are representative of those an applicant may expect to face on the job. To address this, we often tell SMEs we are documenting "performance incidents" or "behavioral incidents" to convey that the situations they document do not have to be extraordinary but should focus on relevant behavioral competencies or performance dimensions.

5.2 Phase II: Item Development

During the prototype Air Force SJT development, Phase II consisted of having the AAD development teams review the draft SJT scenarios to ensure they were realistic and clearly worded. They also generated additional detailed response options to accompany the scenarios.

We reviewed the Air Force SJT items against HumRRO item-writing guidelines and industry best practices. Below we provide observations and recommendations regarding the scenarios and response options.

5.2.1 Scenario and Response Option Review

In general, SJT *scenarios* should describe key features of the situation clearly and concisely (e.g., 4-6 sentences). The description should be clear and realistic, and language and terminology (including acronyms) should be understood by examinees. The examinee should not have to make assumptions about the situation, and responding to the situation should not require knowledge of a specific department, organizational unit, or policy. Across the entire SJT, scenarios should be consistent with verb tense and the perspective of the actor (i.e., second person, third person).

From a content perspective, the scenarios appear to have a great deal of richness. They are appropriate in length and are written in active voice, which makes them easier to read and comprehend. As is relatively common across SJTs, the situations described are ongoing (i.e., the actors are in the midst of the challenging situation), and the examinee is addressed directly (i.e., the scenarios are written in second person).

- **HumRRO Recommendation:** The reading level of the items should be systematically evaluated and documented. There are several off-the-shelf tools that can be used to determine the reading level of a text passage. For example, the Flesch-Kincaid Reading Level feature is available in the Spell Check function in Microsoft Word. We typically aim not to exceed an 8th grade reading level, even in situations where the applicant pool comprises primarily college graduates. Research has shown that items written at a higher reading level tend to exhibit greater subgroup differences in examinee performance.

SJT *response options* describe actions that can be taken to address the situation described in the scenario. The response options should be clear and concise (i.e., we usually recommend keeping them at one sentence), and they should include an appropriate amount of detail. They also should be similar in length, structure, and specificity.

- **HumRRO Recommendation:** Each response option in the Air Force SJT lists several actions that could be taken (i.e., in a sequence). In general, options should not be "double-barreled" and include a sequence of actions (i.e., do this and then do that). It can be hard to avoid this, because the responses need to be rich enough that they are actions people would actually take. However, if the response option has multiple parts, it can be very difficult to judge its overall effectiveness, because one part may be effective and another part ineffective. This is true for both examinees completing the SJT and SMEs who provide ratings of relevance, effectiveness, and the like. Thus, we recommend that each option be limited to reflect a single, discrete action that states what should be done in general, or what should be done first.
- **HumRRO Recommendation:** Depending on the type of scoring approach that will ultimately be used, it can be helpful to gather input from item writers on the options' effectiveness. For example, we have asked item writers to rate the effectiveness of the options as they write them to ensure there is a range in effectiveness across the set of items. When planning to use "pick most effective" and/or "pick least effective" response formats, we have asked item writers to identify not only the most and least effective options, but also the *next-most* effective option and *next-least* effective option, and to provide rationales for why most is better than next most and least is worse than next least.

Gathering the ratings and information from item writers *during the development phase* tends to lead to more careful writing of response options. We therefore suggest implementing this type of process during item development. These ratings and rationales should be kept on file in the event an item key is challenged down the line.

- **HumRRO Recommendation:** During item development, five response options were drafted to accompany each scenario. Because scenarios would be retained only if four response options "survived" subsequent pruning activities (described below), drafting only five options per item may have been overly limiting. We suggest trying to draft double the number of response options than will ultimately be retained (i.e., eight options per item) to account for the fact that some options will have to be dropped for various reasons. Although it can certainly be challenging to develop such a large number of response options, it is worth the investment of time and effort to avoid situations in which an entire scenario must be dropped due to an insufficient number of response options. HumRRO will provide guidelines, tips, and strategies for generating response options in the training manual we develop for Air Force SJT development.

5.2.2 Bias and Sensitivity Review

A primary goal of assessment development is to ensure that tests are both valid and reliable. There is no way to achieve perfect validity and reliability, but careful assessment development and scoring can certainly enhance both test properties. There are various threats to validity and reliability, including things that cannot be entirely controlled (e.g., cheating, an examinee's temporary mental or physical condition during the testing session). On the other hand, some threats can be controlled, such as the exclusion of assessment content that is (a) irrelevant to the target population or job, or (b) poorly constructed.

Along these lines, a primary threat to validity that can and should be excluded is test content that exhibits bias. Two very common issues create bias in assessment content. First, if the language used in the items is specific to a particular culture, examinees from different cultures may be penalized or disadvantaged. Concerns regarding language include slang, colloquialisms, complex grammar, or words not commonly used. Second, there is the concern regarding representation of members of different subgroups. This includes over- or under-representing a culture or group, or using stereotypes (negative or positive beliefs about that group's characteristics) to represent a specific culture or group. This type of bias can create a threatening environment for the examinee, which can cause that examinee to underperform on the test.

As part of our review of the SJT prototype, HumRRO conducted a bias review of the items administered in the pilot test. An internal expert who is knowledgeable about and sensitive to group representation and cultural and language differences reviewed each item, focusing principally on three issues:

- Test materials should not contain any language, roles, situations, or contexts that could reasonably be considered offensive or demeaning to any population group.
- A total test form or pool of items should generally be balanced (or neutral) in cultural and gender representation. Strategies to accomplish this are to ensure inclusion of culturally

diverse passages within each form and/or to ensure that all passages depict themes applicable to all groups.

- No test material should contain elements extraneous to the job-related content and skills being assessed as part of the test specifications. Such extraneous material could provide an unfair advantage or disadvantage to population groups.

Table 1 summarizes bias-related issues by scenario (i.e., item) that were flagged during our review.

- **HumRRO Recommendation:** Although it may seem that the introduction of bias and stereotyping language would be obvious, most often it is unintentional. Therefore, it is advisable not only to train item writers on the meaning of bias and strategies to avoid it, but also to perform a dedicated review to identify test content that could be potentially biased *prior to* pilot testing items. Conducting the bias review prior to pilot testing provides an opportunity for item writers to incorporate feedback to revise items that may have been flagged. If edits are made to address bias-related issues *after* pilot testing, test developers cannot be certain the edits would not affect the revised items' performance and statistical properties. In HumRRO's experience, most bias-related concerns flagged during bias review can be addressed through making rather minor tweaks to the wording or context of the SJT items (e.g., substituting a more familiar vocabulary word, changing the gender of a character). However, some items should ultimately be dropped (and not pilot tested) if bias concerns cannot be addressed through item revision.
- **HumRRO Recommendation:** We suggest formally training both item writers and SME reviewers in strategies to mitigate and detect bias. During training, it is also important to explain the impact that test bias can have on examinee perceptions, sponsoring program reputation, and overall assessment reliability and validity. HumRRO will include a bias and sensitivity review guide in the training manual to be developed in conjunction with this effort. The review guide will describe test bias, explain why it is important to avoid, identify different types of item construction issues that create bias, and provide a comprehensive list of factors to consider as SMEs or item writers review items for bias. All feedback should be carefully documented and a system for implementing feedback should be put in place to ensure items are revised before pilot testing.

Table 1. Item-Level Feedback from Bias Review

Scenario ID	Issue	Description
I	Gender representation	Being "gossipy" is more often a stereotype of women than men. Suggest making both characters men in this scenario.
II	Language (slang)	While the terms may be part of common AF jargon, "on standby" and "in case" are idioms according to the American Heritage Dictionary. Consider changing Option D to "ready and available" and "if," respectively.
VII	Language (slang)	The term "24/7" could be considered an idiom. Consider replacing with "full-time" or "around the clock"
X	Language (complexity)	Rewrite, "Several Airmen have approached you…" to "Several Airmen have complained that…"

X	Sensitivity	Not all speech impediments may be able to be “improved.” If an examinee has a speech impediment that cannot be improved, Option D may be offensive and/or disheartening.
XI	Language (slang)/ Group representation	"Worker bee" is slang and could be seen as pejorative in this context.
XII	Language (slang)	The term “dead end” is an actual, non-slang phrase, so the quotation marks can be removed.
XVI	Language (slang)	“Blew ‘over the limit’” is slang and potentially subject to multiple interpretations. Suggest using less colloquial term (e.g., “intoxicated while driving” or “drunk driving”).
XVII	Language (slang)	The term “cool off” in Option C could be considered slang.
XXII	Language (complexity)	Consider changing the word “relay” to “explain” in Options B and D for simplicity.
XXIV	Gender representation	The example of the wife taking care of a family obligation expresses traditional, stereotypical gender roles.
Overall	Gender representation	There are approximately 23 references to men (across all stems and options), while there are only 7 to women. Of the 7 instances of women appearing in the test, they play roles such as wife, mother, “gossip,” someone who “tattles” on her supervisor, and a victim of harassment. Many of the themes and roles women play conform to traditional gender stereotypes. Consider trying to balance gender across the SJT item pool, and ensure that items do not propagate or reinforce traditional gender stereotypes for either gender.

5.3 Phase III: Scoring Key Development

During prototype SJT development, Phase III consisted of having an independent group of SMEs (i.e., not the AAD development teams who developed the SJT content) provide ratings that were used to evaluate the SJT’s content validity and to develop the scoring key. Specifically, AFPC/DSYX recruited senior NCOs E-7 through E-9 (nonpersonnelists) from as many AFSCs as possible to participate in this phase. During this process, the SMEs read the SJT scenarios and accompanying response options and were asked to:

- rate on a Likert scale the relevance of each scenario to their own AFSC and to MSgts across the Air Force in general,
- identify the single most effective response option and least effective response option,
- rate on a Likert scale (1 through 7) the effectiveness of each response option, and
- provide feedback on any wording changes or realism of ONLY the response options.

These ratings, taken together, were used to prune items and response options and to develop the provisional scoring key for the final experimental version of the prototype SJT used in the pilot study.

5.3.1 Content Validation

Content validity is measured by the degree to which the items on an exam are representative of the range of knowledge and skills required for acceptable performance. There are many ways to

accumulate content-related validity evidence for an assessment, but most approaches typically involve asking SMEs to evaluate or rate the items for job relevance. During Phase III of the Air Force prototype SJT development, 49 SMEs read through the SJT and, using a scale from 1 = Completely Irrelevant to 7 = Extremely Relevant,[1] rated the relevance of each scenario to MSgts in their own AFSC, as well as to MSgts across the Air Force population in general.

Table 2 presents a descriptive summary of these ratings across the 60 items evaluated prior to pruning items ahead of the pilot test. With respect to relevance to Air Force overall, the 60 scenarios were rated at or above a 5 by at least 50% of the SMEs. At least 40% of SMEs rated all 60 scenarios at or above a 5 in terms of relevance to their AFSC.

Table 2. Summary of SME Ratings of SJT Item Relevance (Pre-Pruning)

Relevance	Mean	SD	Min	Max
To AFSC	5.47	0.58	3.73	6.61
To Air Force	5.86	0.37	4.57	6.67

Table 3 presents a descriptive summary of SME relevance ratings across the 25 items retained after pruning for pilot testing. With respect to relevance to the Air Force overall, the 25 post-pruning scenarios were rated at or above a 5 by at least 69% of the SMEs. At least 57% of SMEs rated all 25 post-pruning scenarios at or above a 5 in terms of relevance to their AFSC.

Table 3. Summary of SME Ratings of SJT Item Relevance (Post-Pruning)

Relevance	Mean	SD	Min	Max
To AFSC	5.60	0.37	4.92	6.16
To Air Force	5.91	0.24	5.39	6.26

Perspectives on validity of scores on selection and promotion-related assessments have more recently embraced a unitarian view of validity. That is, any source of evidence that can be brought to bear on the predictive inference of interest (i.e., scores on the SJT will be predictive of future job performance) is of value. HumRRO has adopted a framework for evaluating evidence for the validity of scores that tends to be more sensitive to the nuances of the content and response formats of simulation-based assessments such as SJTs. This framework is based on examining evidence for linkages among (a) assessment content (e.g., scenarios and response options), (b) competencies targeted by the assessment, and (c) the target job or role. Establishing evidence for multiple types of linkages provides support for the claim that scores from the assessment will predict examinees' future performance in the Air Force.

- **HumRRO Recommendation:** Aside from having SMEs rate the relevance of the types of scenarios reflected in the SJT items to their AFSC and to the Air Force overall, there are other activities that can be conducted moving forward to bolster support for the content-related validity evidence of the SJT. First, although job relevance is arguably one of the most important aspects of content validity, the extent to which the items reflect and measure the target competencies (i.e., construct-related validity evidence) is also

[1] Scale points 2 through 6 did not have anchors.

important—particularly if there is any intent to report scores at the competency level or give candidates feedback on their relative or absolute standing on competencies. Thus, it can be helpful to conduct a “retranslation activity" in which SMEs sort the SJT scenarios (or even the actions reflected in the response options) into the competency (or competencies) they believe are reflected in the scenario. Substantiating these links between assessment content and the assessment’s target competencies provides additional evidence in support of the assessment’s validity.

5.3.2 Key and Distractor Identification

A primary consideration when it comes to the validity of an SJT regards the defensibility of the scoring key. Unlike traditional, knowledge-based multiple-choice questions, there is often not an objectively correct, “by-the-book” answer to SJT items. Instead, examinees must exercise their subjective judgment to evaluate various potential courses of action in a given situation to determine the appropriate way to respond. This is often viewed as a strength of SJTs, in that they measure something that cannot be adequately measured by traditional knowledge tests. It also means, however, that due diligence must be taken to ensure a solid basis for the SJT’s scoring key.

A variety of approaches can be used to screen and evaluate SME ratings to determine the scoring key. For the Air Force prototype SJT, SMEs identified the response options they considered most and least effective, and they provided effectiveness ratings for each response option. We have found this approach to be beneficial, because these response formats do not always point to the same key. That is, the option with the highest SME effectiveness rating is not always the option with the highest number of SMEs categorically judging it to be the most effective. This inconsistency largely stems from the fact that SMEs can provide more than one option the same effectiveness rating to indicate a “tie” between the options, whereas they must select one and only one option to be the most (or least) effective. After pruning options that rival the most or least effective keyed response, we typically see much higher agreement between the keys generated by the “pick the most/least effective” and “rate the effectiveness” formats. In the case of the USAF SJT, the most effective option based on highest mean effectiveness rating was the same as the most effective option based on categorical judgments 88% of the time pre-pruning when ties *were* permitted among categorical judgments for most effective. That is, the most effective option based on ratings was the same as *any* option that tied for most effective based on categorical judgments. When ties were <u>not</u> permitted (i.e., when there was no clear categorical judgment key because more than one option received the most judgments as most effective) the ratings and categorical judgments aligned only 82% of the time pre-pruning. The least effective option based on lowest mean effectiveness rating was the same as the least effective option based on categorical judgments 93% of the time pre-pruning when ties were permitted, and 90% of the time pre-pruning when ties were <u>not</u> permitted. After pruning, the keys aligned 100% of the time for most and least effective options. Thus, pruning the response options to remove options that rivaled the key in effectiveness (this process is described in more detail below) did bring these types of judgment into alignment when it comes to informing the scoring key.

- **HumRRO Recommendation:** To help interpret SMEs’ most and least effective option selections in light of the Likert-scale effectiveness ratings they provide, it would be helpful to ask SMEs to provide a rationale for why they selected an option as most or

least effective. This process can often point to ambiguities in option wording and can help identify the frame of reference SMEs used to interpret the option. Once an item's key is ultimately determined, these rationales can also be retained on file to support the defensibility of the SJT key in the event an item is challenged during administration.

The Air Force used the SME effectiveness ratings to determine which options should be keyed as most and least effective. To do this, they first computed descriptive statistics (i.e., mean and standard deviation) on the Likert scale effectiveness ratings for each option. The option with the highest mean effectiveness rating was set as the key for the most effective response, and the option with the lowest mean effectiveness rating was set as the key for the least effective response. In our review of the materials, it was unclear which rules or procedures were followed when two options tied for the most or least effective during this process.

- **HumRRO Recommendation:** The process used to identify the keyed response options was appropriate. HumRRO has used a similar process in the past for SJTs we have developed. If not already in place, we suggest implementing a clear and simple set of rules to identify the most and least effective keys in the event of a tie in effectiveness ratings. For example, we have used the following set of rules:
    1. If an option has the highest (or lowest) mean effectiveness rating for an item, flag it as a potential key for most (or least) effective (1st pass).
    2. If more than one option has the highest (or lowest) mean for an item, choose the option(s) with the lowest standard deviation (2nd pass).
    3. If more than one option has the highest (or lowest) mean, and those options have the same standard deviation, randomly choose one of those options as the keyed response (final). (Note that the non-selected rival key will necessarily be dropped based on steps described below.)

After keying the most effective and least effective response options, the next step was to identify viable distractors. A response option in any given item was deemed as a viable distractor if and only if its mean effectiveness rating was significantly different from the mean effectiveness ratings of *both* the most effective and least effective response options based on a dependent-sample (i.e., matched- or paired-sample) *t*-test ($p < .01$, one-tailed). Each scenario had to have at least two viable distractors to be retained. Scenarios with fewer than two viable distractors were dropped. After setting the key and dropping non-viable options for all scenarios to be retained, some scenarios still had more than two viable distractors. In our review of the materials, it was unclear which rules or procedures were used to determine which distractors to drop versus retain in these situations. In addition, there seemed to be several deviations from what we inferred to be guiding principles in terms of pruning response options and scenarios. For example, some options were retained despite not being significantly different from a most or least effective key, other options were dropped despite being viable (presumably to maintain consistency in having four options per item), and entire scenarios were dropped in several cases despite having a sufficient number of distractors. It is unclear what drove these decisions.

- **HumRRO Recommendation:** The overall process used to identify viable distractors was appropriate and rigorous. HumRRO has used a similar process in the past for SJTs we have developed. If not already in place, we suggest implementing a clear and simple set

of rules to guide response-option pruning when more than two distractors are viable. For example, we have used the following set of rules:

1. Determine if any of the viable distractors are redundant (e.g., highly similar wording, high conceptual overlap). If an option is conceptually redundant with one of the keys, drop that option. If two distractors are redundant, then proceed to the subsequent steps to determine which of the redundant distractors to drop. If there are multiple pairs of overlapping responses, do the same for each pair.
2. Determine if any of the options are of subjectively low quality: Is the option poorly written? Does the option not make sense based on knowledge of the item's intent? Is the option too transparent (i.e., no one would rationally select it)? Drop all low-quality options.
3. Try to sample the range of the scale with the remaining distractors. For example, if remaining distractors have effectiveness ratings of 2.5, 2.7, 3.5 and 4.0 after taking the preceding steps, pick only one of the lower two, as they are very close together on the effectiveness scale. Make sure redundant options identified in step 1 have been removed by this point.

5.3.3 Evaluation of Interrater Reliability and Agreement of SME Ratings

As part of our review of the prototype Air Force SJT, HumRRO also evaluated the interrater reliability and agreement of the SME ratings used to determine the scoring key. As with other high-stakes tests used for selection purposes, before validity evidence is established (i.e., evidence that the test predicts job performance or other criteria of relevance to the Air Force), evidence must first be established that the test is reliable, meaning that the test has stability/consistency in measurement.

The SME judgments have two primary sources of systematic variance: variance due to response options and variance due to raters (SMEs). Variance due to response options – the target of measurement – is considered *true score variance* in the language of classical test theory. This is "good" variance, in that we want to appropriately reflect the differences in effectiveness or relevance among the SJT response options. Variance due to raters, however, can be considered a source of error variance (depending on the type of information the analyst seeks to report). Ideally, all raters would be replicates of one another, providing identical (and accurate) judgments. In truth, raters often differ systematically from one another. The differences in their judgments can contribute artifactual variability to the judged relevance or effectiveness of the response options.

Intraclass correlation coefficients (ICCs; McGraw & Wong, 1996) are constructed from different variance components estimated from the ratings data to obtain estimates of interrater reliability (IRR) and interrater agreement (IRA). IRR reflects whether judges (here, SMEs) rank-order targets (here, the SJT item response options) in a manner that is relatively consistent with (i.e., the same as that observed for) other judges (Lebreton & Senter, 2008). In contrast, IRA refers to the absolute consensus in ratings, rather than just the relative rank order of the rated targets. IRA estimates are used to assess whether judges' scores are interchangeable or equivalent in their absolute value (Lebreton & Senter, 2008). IRR and IRA are separate indices. A set of ratings can

evidence high or low IRR and high or low IRA (see Figure 1; note that an example of low IRR and IRA is not shown).

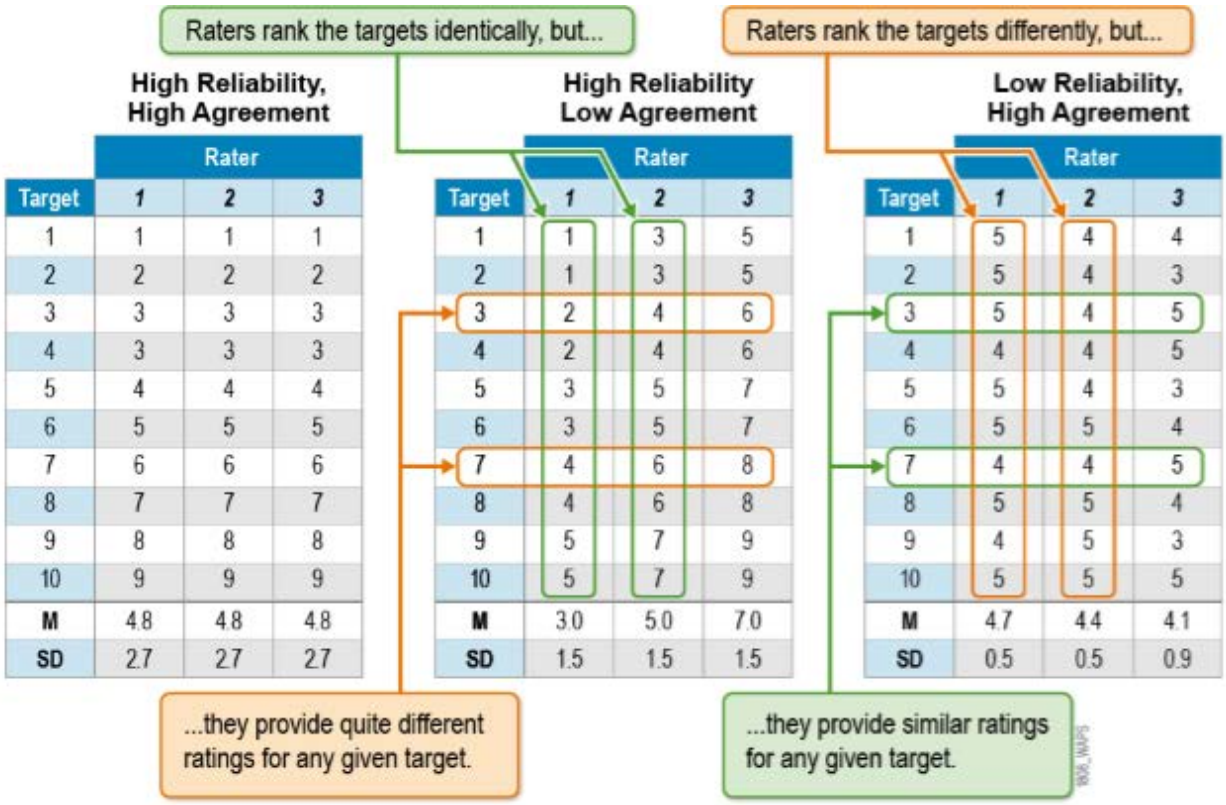

Figure 1. Examples of high and low interrater reliability and agreement.

ICC (C, *k*) is a measure of IRR, as it concerns the consistency (C) among SMEs' (of which there are *k*) ratings (Bliese, 2000; James, 1982; LeBreton, Burgess, Kaiser, Atchley, & James, 2003; McGraw & Wong, 1996). ICC (C, *k*) considers how similarly raters rank order the targets of measurement rather than the similarity of magnitude of those ratings for the targets and thus does not consider variance due to raters to be measurement error. ICC (C, *k*) generally ranges from 0 to 1, with a value of 0 indicating no consistency in rank order of measurement and a value of 1 indicating perfect consistency in rank order of measurement (Lebreton & Senter, 2008). An ICC with a value of .700 or higher is a typical cutoff used to justify aggregation of ratings (e.g., to determine an option's mean effectiveness rating across SMEs). However, more stringent cutoffs (e.g., > .900) are recommended, for example, when evaluating the reliability of ratings used to establish scoring keys during the development of assessments that will be used to inform high-stakes decisions such as selection (Lebreton & Senter, 2008).[2]

We computed ICC (C, *k*) for the overall relevance ratings, AFSC relevance ratings, as well as for effectiveness ratings both before and after item pruning to assess IRR among SMEs. Reliabilities are reported only on a pre-pruning basis for relevance ratings because pruning, by design,

[2] Other researchers believe all data should be used for aggregation, not just the data from more homogeneous groups (Carron et al., 2003; Cole, Bedeian, Hirschfeld, & Vogel, 2011).

restricts range in relevance and attenuates reliability. Thus, this process resulted in a total of four ICC (C, *k*) estimates. As shown in Table 4, the ICC (C, *k*) values all exceed .700, and those for the option effectiveness ratings, upon which the scoring key is based, exceed the .900 threshold. The ICC (C, *k*) increases after pruning for effectiveness ratings. The effectiveness rating IRR levels provide support for aggregating ratings to inform the scoring key.

Table 4. Interrater Reliability Estimates

	Overall Relevance Ratings	AFSC Relevance Ratings	Option Effectiveness Ratings	
			Pre-Pruning	Post-Pruning
ICC (C, *k*)	.873	.903	.972	.980

ICC (A, *k*) is a measure of IRA, as it concerns the absolute agreement (A) among SMEs' (of which there are *k*) ratings (McGraw & Wong, 1996). Unlike ICC (C, *k*), ICC (A, *k*) considers idiosyncratic rater differences to be measurement error. That is, ICC (A, *k*) includes rater variance as part of the error term, meaning that it reflects the absolute differences in scores, not just the relative rank order. With an additional source of error variance now considered, ICC (A, *k*) values will be lower than (and at best, equal to) ICC (C, *k*) values. We computed ICC (A, *k*) for the overall relevance ratings, AFSC relevance ratings, as well as for the effectiveness ratings both before and after item pruning to assess IRR among SMEs. Again, agreement is reported only on a pre-pruning basis for relevance ratings because pruning, by design, restricts range in relevance and attenuates reliability. As shown in Table 5, the ICC (A, *k*) increases ever so slightly for option effectiveness ratings after pruning response options.

Table 5. Interrater Agreement Estimates

	Overall Relevance Ratings	AFSC Relevance Ratings	Option Effectiveness Ratings	
			Pre- Pruning	Post- Pruning
ICC (A, *k*)	.769	.839	.970	.978

SMEs also were tasked with indicating which response option was most effective and which response option was least effective. These judgments are *categorical judgments*, in that raters either agree or disagree that a certain response option was most (or least) effective. Several types of indices assess the extent to which raters agree when making categorical judgments. The simplest way to represent agreement is to report the percentage of raters who agree (i.e., percentage agreement). As shown in Table 6, the percentage agreement on which responses were the most and least effective increased after item pruning. However, even after pruning only 62.9% of SMEs agreed on which response option was most effective, and 64.1% agreed on which response option was least effective. Considering the Air Force prototype SJT is scored using "most effective and least effective" response keys, this is a relatively low level of agreement.

Table 6. Agreement for SMEs' Categorical Judgments

	Percentage Agreement		Krippendorff's Alpha		Gwet's AC(1)	
	Pre-Pruning	Post-Pruning	Pre-Pruning	Post-Pruning	Pre-Pruning	Post-Pruning
Most Effective	56.3	62.9	.265	.323	.276	.344
Least Effective	56.7	64.1	.289	.321	.347	.344

Although intuitive to understand, percentage agreement is an imperfect measure of the agreement of categorical judgments because it does not account for chance agreement. Some level of agreement is expected to be observed just by chance, and thus percentage agreement can misrepresent true levels of agreement. Chance agreement can be thought of as agreement that occurs when raters agree on a rating due to one or both raters giving a random rating. The important point is that chance agreement can inflate the overall agreement probability and should not contribute to a measure of actual agreement between raters (Blood & Spratt. 2007).

To assess percentage agreement above and beyond chance agreement, we computed Krippendorff's Alpha and Gwet's (2008) AC (1), which are measures of agreement that control for chance agreement. They also are reported in Table 6. Both indices range from 0 to 1, with a value of 0 indicating an absence of agreement and a value of 1 indicating perfect agreement (Hayes & Krippendorff, 2007). Krippendorff's Alpha and Gwet's AC (1) calculate chance agreement differently, so we have chosen to report both indices.

In sum, SMEs' effectiveness ratings tended to exhibit high agreement, but the SMEs' categorical judgments about which options are most and least effective showed much lower agreement.

- **HumRRO Recommendation:** Given these results, we recommend that the Air Force use effectiveness ratings as the basis for scoring the prototype SJT until more research has been done to investigate why the categorical judgments of most/least effective are not more aligned amongst SMEs.

5.3.4 Revising Response Option Wording

During Phase III of the USAF prototype SJT development, SMEs also were asked to provide feedback on wording changes to the response options associated with the items. We have found that altering the wording of a response option even slightly can alter perceptions of the option's effectiveness. Thus, it is possible that making even trivial changes to options at this point could have changed or invalidated the scoring key associated with that item.

- **HumRRO Recommendation:** We recommend having SMEs carefully review the wording of the items prior to collecting effectiveness ratings and categorical judgments to inform scoring key development. It is best to make any edits, whether trivial or substantive in nature, prior to the scoring key development stage, because the changes could change the key itself. If item scenarios or options must be revised either concurrent with or after key development, care should be taken to ensure the changes would not substantially alter the scoring key. If this concern arises, the item should be dropped (not revised) until the key can be verified and/or revised as needed.

6.0 SUMMARY

The U.S. Air Force is exploring the use of a situational judgment test as a potential augmentation to the WAPS. AFPC/DSYX developed a prototype SJT for use with E-7 (master sergeant) candidates. This report summarized HumRRO' s evaluation of the materials and procedures used to develop the SJT to determine what types of changes or improvements to the SJT development process might be warranted before moving forward with additional feasibility studies for the inclusion of an SJT in WAPS. Detailed recommendations were presented throughout this report. The list below summarizes our recommendations:

- Document linkages between the competencies in the Air Force Institutional Competency model and the duty areas (or better yet, specific tasks) identified in the RAND study to build evidence for the content validity and job-relatedness of the SJT.
- Gather SME content validity ratings to bolster support for the use of the SJT at lower ranks, if the SJT will be used for promotion into those ranks.
- Re-evaluate the competencies targeted by the SJT to determine if they are truly amenable to measurement via an SJT.
- Articulate a schema for the types of behaviors that do and do not fall under each of the target competencies, paying close attention to areas of potential conceptual overlap. Use this scheme to determine the types of situations that would elicit each competency.
- Explore the internal competency assessments currently being administered by the Institutional Competency Development Programs (ICDPs) and determine if these relate to or overlap with content featured on the SJT, or if they contain content that would be appropriate to feature on the SJT.
- Ask SMEs to explicitly identify the judgments or decisions they had to make to address the challenging situations they describe in their critical incident documentation.
- Instead of using the term "critical incidents," tell SMEs they are documenting "performance incidents" or "behavioral incidents" to convey that the situations they document do not have to be extraordinary or "critical" in nature, but should focus on relevant behavioral competencies or performance dimensions.
- Systematically evaluate and document the reading level of the items.
- Write response options such that they are not double-barreled and reflect a single, discrete action rather than a sequence of actions.
- Have item writers rate response option effectiveness and designate most/least effective, providing rationales to support their designations, *during* the response-option drafting process.
- Draft twice as many response options as needed to account for item pruning that will take place after key development and pilot testing.
- Conduct a bias review prior to scoring key development and pilot testing to revise items while they are still "under development."
- Train item writers and bias reviewers in strategies to mitigate and detect bias.

- Conduct a retranslation activity in which SMEs sort the SJT scenarios (or even the actions reflected in the response options) into the competency(s) they believe are reflected in the scenario to generate additional evidence in support of the assessment's content validity.
- Ask SMEs to provide a rationale for why they selected an option as most or least effective during the scoring key development process. Retain these rationales on file in the event an item is challenged.
- Implement a clear and simple set of rules to follow to identify the most and least effective keys in the event of a tie in effectiveness ratings.
- Implement a clear and simple set of rules to guide response option pruning when more than two distractors are viable.
- The Air Force should use effectiveness ratings as the basis for scoring the prototype SJT until more research is done to investigate why the categorical judgments of most/least effective are not more aligned amongst SMEs.
- Refrain from making non-trivial edits to item scenario or option wording after the scoring key has been developed.

7.0 REFERENCES

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349-381). San Francisco: Jossey-Bass.

Blood, E., & Spratt, K. F. (2007). *Disagreement on agreement: Two alternative agreement coefficients*. SAS Global Forum.

Carron, A. V., Brawley, L. R., Eys, M. A., Bray, S., Dorsch, K. D., Estabrooks, P. A., Hall, C. R., Hardy, J., Hausenblas, H., Madison, R., Paskevich, D., Patterson, M. M., Prapavessis, H., Spink, K. S., & Terry, P. C. (2003). Do individual perceptions of group cohesion reflect shared beliefs? An empirical analysis. *Small Group Research, 34*, 468-496.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63*, 83-117.

Cole, M. S., Bedeian, A. G., Hirschfeld, R. R., & Vogel, B. (2011). Dispersion-composition models in multilevel research: A data-analytic framework. *Organizational Research Methods, 14*, 718 –734. doi: 10.1177/1094428110389078

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327-358.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61(1),* 29-48.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1(1),* 77-89.

James, L. R. (1982). Aggregation in estimates of perceptual agreement. *Journal of Applied Psychology, 67,* 219-229.

Keller, K. M., Robson, S., O'Neill, K., Emslie, P., Burgette, L. F., Harrington, L. M., & Curran, D. (2014). *Promoting airmen with the potential to lead: A study of the Air Force master sergeant promotion system* (RR-581-AF). Santa Monica, CA: RAND Corporation. As of May 25, 2018: https://www.rand.org/pubs/research_reports/RR581.html

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11(4),* 815-852.

LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K. P., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods, 6(1),* 80-128.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1(1),* 30-46.

Ployhart, R. E. (2006). The predictor response process model. In J. A. Weekley & R. E. Ployhart (Eds.) *Situational judgment tests: Theory, measurement and application* (pp. 83-105). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

U.S. Air Force. (2015, December 22). *Total force development and management* (AFPD 26-26). Washington, DC: Author.

U.S. Air Force. (2016, September 15). *Institutional competency development and management* (AFMAN 36-2647). Washington, DC: Author.