Psychometric Research Studies:
Development of the Complex Reasoning Test

Presenter: **Scott H. Oppler, HumRRO**

September 18, 2020

Overview

- Background
- Project Objectives
- Research Plan
- Progress to Date

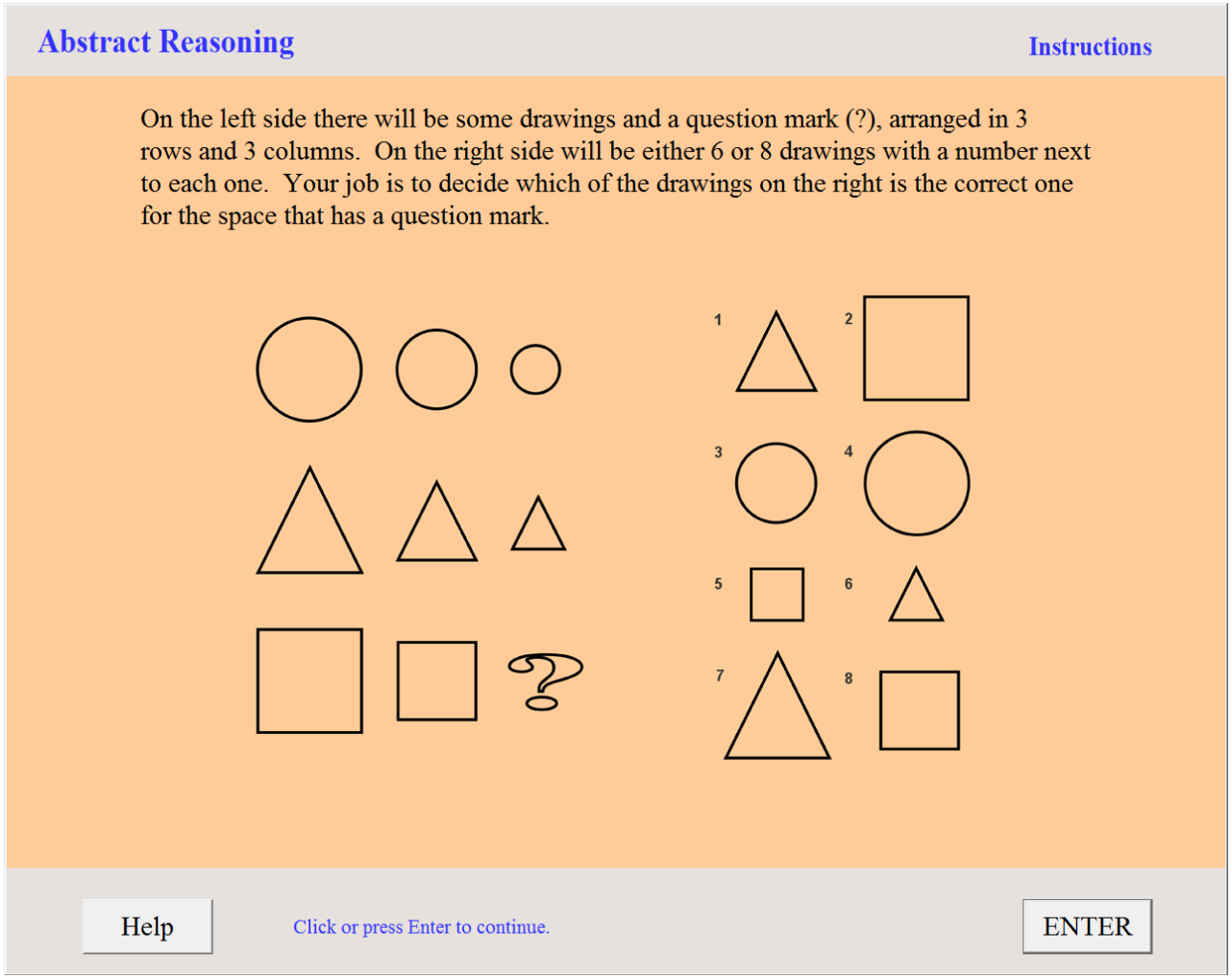
Background

- Fluid intelligence
  - Fluid intelligence has been found to be a strong predictor of training and job success.
  - The 2006 ASVAB Review Panel suggested that DoD consider adding a test of fluid intelligence to better balance the ASVAB's composition (between fluid and crystalized intelligence).
- Potential benefits of adding a test of fluid intelligence to the ASVAB
  - Higher prediction of training and job success
  - Lower susceptibility to compromise
  - Increased qualification rates for non-native and non-heritage English speakers

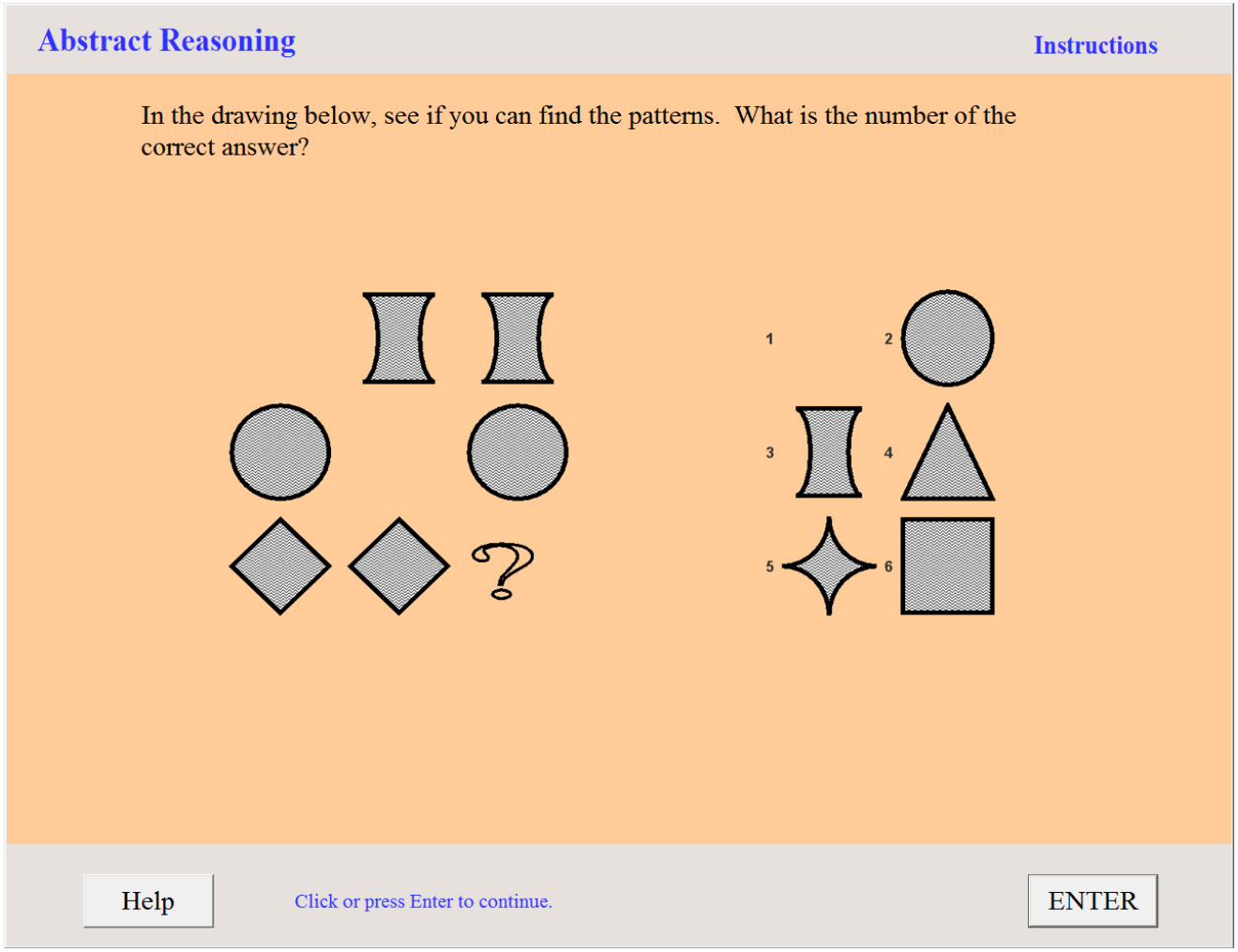
Background

- Earlier work: Abstract Reasoning Test (ART)
  - Developed by Embretson—format similar to Raven's Progressive Matrices items (multiple choice, 6 or 8 response options per item)
  - DPAC commissioned the development of one form (30 items)
  - Administered (for research purposes) to language training applicants (2017)
  - Items were found to be relatively easy, time-consuming

Sample Abstract Reasoning Test (ART) item



Sample Abstract Reasoning Test (ART) item



Background

- DPAC would like to develop a *non-proprietary* Automated Item Generation (AIG) algorithm and difficulty model for a non-verbal test of fluid intelligence (aka, “Complex Reasoning Test”)
  - Improve item development efficiency
  - Reduce/eliminate field-testing requirements
- Desired item properties:
  - Similar to Raven’s Progressive Matrices items
  - Appropriate difficulty for qualifying military applicants into jobs of varying complexity

Project Objectives

- Evaluate the viability of developing complex reasoning test items using an existing, non-proprietary item generation tool
  - Sandia Generated Matrix Tool (SGMT), developed by the Sandia National Laboratories, Albuquerque, New Mexico
- Identify the features of these items that impact their item difficulty parameters
- Develop difficulty model to improve the efficiency associated with the development and calibration of these items

Research Plan

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

Research Plan

- Stage 1: Conduct background research
  - Review literature
    - Examine prior attempts to model difficulty of matrix reasoning items
    - Examine Abstract Reasoning Test (ART) evaluation study (2017)
  - Assess item generation capabilities of SGMT
  - Identify relevant research questions

- Stage 2: Design and conduct pilot study
  - Develop data collection and analysis plan (to address research questions identified in Stage 1)
  - Develop preliminary set of complex reasoning items using SGMT (k = approx. 60)
  - Collect and analyze data

Research Plan (cont.)

- Stage 3: Design and conduct full study
  - Develop data collection and analysis plan (to address research questions identified in Stage 1)
  - Develop full set of items (k = 200, including items surviving pilot studies)
  - Collect and analyze data
    - Develop and evaluate difficulty model

Progress to Date

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

Progress to Date

- Stage 1: Conduct background research
  - Examine prior attempts to model parameters of matrix reasoning items [complete]
    - Variance accounted for in item difficulties is appreciable, but has varied widely (R^2 = .40 to .80)
    - Substantially less success found in modeling discrimination or guessing parameters
  - Examine Abstract Reasoning Test evaluation study [complete]
    - Previously briefed to MAPWG/DACMPT (Gao et al., 2019)
    - Form administered to over 2,162 highly qualified examinees (military personnel applying for language training)
    - Items tended to be relatively easy (60% had p-values $\geq$ .80)
    - 25-minute time limit may not have been sufficient
    - Only six (of 30) items had more than three distractors selected by 3% or more of the examinees

Progress to Date

- Stage 1: Conduct background research (cont.)
  - Assess item generation capabilities of SGMT [complete]
    - Two item types: Transformation Items and Logic Items
    - Item generation is not fully automatic; requires user to specify structural elements, but can then be used to develop what are essentially clones
    - Several features no longer function as intended, given outdatedness of software (programmed in early version of Java)
    - Many distractors, as generated, appear to be less than plausible, requiring human intervention to make distractors more attractive
    - Items are not generated in a high-resolution format, necessitating additional human processing
    - However, the tool still allows for the manipulation of a variety of features that have been shown by the developers to influence item difficulty and, therefore, may be adequate for purposes of the pilot research (but probably not for larger-scale needs)

Progress to Date

- Stage 1: Conduct background research (cont.)
  - Assess item generation capabilities of SGMT [complete]
    - It has subsequently been learned that ARI has been granted access to an updated version of the tool (developed for another Government agency) that addresses many of the problems currently associated with the original version
    - Likewise, ARI has also indicated interest in the development of a more thoroughly modernized version of the tool
    - DPAC and ARI have recently begun discussing the possibility of collaborating to more efficiently/effectively develop these item types (i.e., transformation and logic items) and to jointly pursue mutually beneficial research to determine the potential usefulness of such items

Progress to Date

- Stage 1: Conduct background research (cont.)
  - Identify relevant research questions to address in Stage 2 [complete]
    - What is the impact of the two different item types (Transformation and Logic) on the dimensionality of the test?
    - How does the number of response options influence the psychometric properties/timing of the items?
      - What are the pros/cons of reducing the number of response options to 4?
    - To what extent is performance on these items influenced by practice?
    - Can the difficulty of the items be adequately controlled/modeled by systematically varying certain features of the items in order to
      - make the test sufficiently difficult for qualifying military applicants into jobs of varying complexity?
      - avoid (or reduce) the need for calibration with live test-takers for scoring and adaptive item selection?

Progress to Date

- Stage 2: Design and conduct pilot study
  - Develop data collection and analysis plan [in progress]
    - Currently designing pilot study to evaluate: (a) the possible multidimensionality associated with the two items types; (b) the impact of number of response options on the psychometric properties of items and test scores; and (c) practice effects
    - The study will also be used to evaluate the extent to which items can be made sufficiently difficult for military selection and classification purposes
    - Currently planning to collect data from non-military examinees via the internet (i.e., using Mturk); potential examinees will be screened to reflect characteristics of ASVAB examinees
    - A question remains regarding which version of an item generation tool to use to develop the necessary items (DPAC has not yet been granted access to the updated version of the tool)

Progress to Date

- Stage 3: Design and conduct full study
  - Develop data collection and analysis plan [not started]
    - Plan will depend on results of Stage 2 pilot study

Questions?
Comments?

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION