



This document reports on work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract 2015-1412020002-002, and is subject to the Rights in Data-General Clause 52-227.14, Alt. IV (DEC 2007). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, ODNI, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon.

©2016 The MITRE Corporation.
All rights reserved.

**Approved for Public Release;
Distribution Unlimited. Case
Number 16-0956**

The Assessment of Biases in Cognition

Development and Evaluation of an Assessment Instrument for the Measurement of Cognitive Bias

Abigail Gertner, The MITRE Corporation
Franklin Zaromb, Educational Testing Service
Robert Schneider, Research & Assessment Solutions, Ltd.
Richard D. Roberts, Professional Examination Service
Gerald Matthews, University of Central Florida

Abstract

The Assessment of Biases in Cognition (ABC) is a new standardized assessment of biases in judgment and decision-making behavior that was developed by the MITRE Corporation and the Educational Testing Service (ETS) for the Intelligence Advanced Research Projects Activity (IARPA) Sirius Program. The purpose of the IARPA Sirius Program is to create serious video games designed to train intelligence analysts to improve their explicit knowledge of, and ability to recognize, six well-known cognitive biases and to significantly mitigate the influence of those biases on behavior as a result of this training. The six biases are: (1) confirmation bias (CB), (2) fundamental attribution error (FAE), (3) bias blind spot (BBS), (4) anchoring bias (ANC), (5) representativeness bias (REP), and (6) projection bias (PRO). The first version of the ABC (ABC-1) was developed for the first phase of the Sirius Program to assess mitigation of CB, FAE, and BBS. The second version of the ABC (ABC-2) was developed for use in second phase of the Sirius Program to assess mitigation of ANC, REP, and PRO.

The ABC-1 and the ABC-2 each include one recognition and discrimination (RD) scale and three behavioral elicitation (BE) scales, one for each bias. The RD scales consist primarily of multiple-choice items and are intended to assess declarative knowledge of the biases. The BE scales consist of a variety of innovative tasks intended to evaluate test-takers' procedural knowledge regarding how to avoid committing the targeted biases in judgment and decision-making tasks specifically designed to give test-takers opportunities to commit those biases. Each version of the ABC is administered online using a customized test delivery platform developed by the MITRE Corporation and takes approximately 45 to 60 minutes to complete. The ABC-1 and ABC-2 both include three equated test forms. This facilitated evaluation of bias mitigation training outcomes by making it possible to compare test-takers' performance on one form post-training with their pre-training performance on an alternate, equated ABC test form. This report summarizes the (1) test development process, (2) research conducted during the development and validity evaluation of the ABC, (3) validity argument for the ABC, and (4) suggestions for future research.

This page intentionally left blank.

Executive Summary

The Assessment of Biases in Cognition (ABC) is a new standardized assessment of biases in judgment and decision-making behavior that was developed by the MITRE Corporation and the Educational Testing Service (ETS) for the Intelligence Advanced Research Projects Activity (IARPA) Sirius Program. The purpose of the IARPA Sirius Program is to create serious video games designed to train intelligence analysts to improve their explicit knowledge of, and ability to recognize, six well-known cognitive biases and to significantly mitigate the influence of those biases on behavior as a result of this training. The six biases are: (1) confirmation bias (CB), (2) fundamental attribution error (FAE), (3) bias blind spot (BBS), (4) anchoring bias (ANC), (5) representativeness bias (REP), and (6) projection bias (PRO).

The Sirius Program was divided into two phases. Phase 1 encompassed biases 1–3 and took place between October 2011 and September 2013. The first version of the ABC (ABC-1) was developed for use in the Phase 1 Independent Validation and Verification (IV&V) study to assess mitigation of CB, FAE, and BBS. Phase 2 encompassed biases 4–6 and took place between September 2013 and November 2015. The second version of the ABC (ABC-2) was developed for use in the Phase 2 IV&V to assess mitigation of ANC, REP, and PRO. The ABC-1 and ABC-2 are referred to, collectively, as the ABC.

The ABC consists of two broad classes of items: recognition and discrimination (RD) and behavioral elicitation (BE). The ABC-1 and the ABC-2 each include one RD scale and three BE scales, one for each bias. The RD scales consist primarily of multiple-choice items and are intended to assess declarative knowledge of the biases. The BE scales consist of a variety of innovative tasks intended to evaluate test-takers' procedural knowledge regarding how to avoid committing the targeted biases in judgment and decision-making tasks specifically designed to give test-takers opportunities to commit those biases. To the extent possible, the tasks were grounded in, and adapted to varying degrees from, extant paradigms relevant to each of the six biases. The BE tasks are complex scenario-based assessments that require test-takers to make decisions and solve problems presented in text, video, and/or voice-over audio formats, typically under conditions of uncertainty, time pressure, and/or rewards (and penalties).

Each version of the ABC is administered online using a customized test delivery platform developed by the MITRE Corporation and takes approximately 45 to 60 minutes to complete. The ABC-1 and ABC-2 both include three equated test forms. This facilitated evaluation of bias mitigation training outcomes by making it possible to compare test-takers' performance on one form post-training with their pre-training performance on an alternate, equated ABC test form.

The purpose of this Executive Summary is to provide a relatively brief synopsis of the complete, and rather extensive, ABC technical report. In the sections that follow, we summarize the (1) test development process, (2) research conducted during the development and validity evaluation of the ABC, (3) validity argument for the ABC, and (4) suggestions for subsequent research based on the project described in this technical report.

Test Development

Development of the ABC-1 and ABC-2 included the following steps:

- **Construct Identification.** This process included:

- reviewing literature relevant to the Sirius project biases, including bias description and elicitation, bias mitigation techniques, individual differences in bias susceptibility, correlates of the biases, and illustrations of how the biases relate to the work of intelligence analysts;
 - generating operational definitions of the bias constructs, including their facets, to help ensure the most complete possible coverage of each bias construct; and
 - periodically consulting with a technical advisory group (TAG), subject matter experts (SMEs), and the IV&V team (which included representatives from IARPA, Johns Hopkins University Applied Physics Lab [JHUAPL], and MITRE) in order to clarify the content and boundaries of each bias construct.
- **Development of Item Prototypes.** We developed BE and RD item prototypes using the following sources: (1) operational definitions of each bias or bias facet; (2) our review of the literature; (2) case studies of intelligence analysis; (3) critical incidents adapted from in-depth interviews with several IC SMEs; and (4) input from the TAG and IV&V team.
 - **Cognitive Laboratory Pilot Research.** We conducted two rounds of cognitive lab studies of BE item prototypes with several dozen ETS employees to identify task elements that test-takers found to be unclear, distracting, or too demanding. In addition, we examined both concurrent think-aloud protocols and retrospective descriptions of test responses in order to enhance understanding of conscious decision making and problem solving strategies adopted by test-takers to improve the ABC.
 - **Item Generation.** Following the development and evaluation of item prototypes, we created a pool of over 600 BE and RD items during both phases of the project. The item pool included several dozen scripted scenarios that were filmed and edited by a professional video production company in Louisville, KY, and at the ETS Princeton, NJ, campus with local professional actors and ETS employees.
 - **Item Review.** Items were reviewed by assessment development specialists and SMEs, including the IV&V team and TAG, with respect to criteria such as (a) clarity, (b) lack of ambiguity and vagueness, (c) ensuring that the items do not assume knowledge specific to the intelligence analyst job, and (d) sensitivity to EEOC protected class (e.g., based on gender, race/ethnicity, age) bias and fairness issues. For items that had content specific to intelligence analysis work, additional reviews were performed by Intelligence Community SMEs at MITRE.
 - **Pilot Testing.** Because the constructs targeted for measurement in the ABC were not well understood from an individual differences perspective, we conducted multiple rounds of programmatic research to enhance understanding and measurement of the biases prior to finalizing and evaluating the validity of the ABC scales.
 - **Assembling and Authoring in Test Administration Platform.** We developed a test administration platform specifically to support the authoring and administration of the ABC. The platform was designed for web-based test administration and hosted on a secure web server. The platform was also designed to facilitate the authoring, revision, and exporting of test-taker responses. In general, this test delivery software was designed to accommodate a wide variety of item/task types in the ABC and to maximize usability, flexibility, and security.

- **Final Field Tests.** We administered the ABC-1 and ABC-2 online in separate field tests, each consisting of over 2,000 U.S. adults. The purpose of the field tests was to administer the entire set of tasks/items to a large and representative group of test-takers to evaluate the ABC's psychometric properties (e.g., mean, standard deviation, frequency distribution, reliability metrics, informative correlations with other measures) and validity, and to collect data necessary for creation of equivalent forms for use in the IV&V. We also conducted studies to evaluate the sensitivity of the ABC to surrogate bias mitigation interventions provided by IARPA.

Preparation and Delivery of Final Test Forms

We developed User Manuals and deployment packages to provide JHUAPL with information necessary to implement the ABC-1 and ABC-2 in the Phase 1 and Phase 2 IV&V studies. The User Manuals describe: (1) the content of the ABC-1 and ABC-2; (2) the scoring process for the ABC scales; (3) test equating methodology to link ABC scores across test forms; and (4) data processing and syntax files created to score the ABC forms. The ABC-1 and ABC-2 deployment packages included: (1) Python scripts and associated files configured to process raw data files from individual test-takers and transform them into a single, master data set; and (2) SPSS syntax files to compute all the scores for the ABC scales.

Overview and Key Findings of the ABC

The table below provides an overview of the contents of the ABC-1 and ABC-2 BE and RD scales, as well as key findings from the pilot test and field test studies. In that table, we refer to two different reliability metrics: internal consistency reliability and test-retest reliability. Internal consistency reliability refers to the extent to which items making up a scale relate to one another statistically (e.g., intercorrelate). It is an index of whether different parts of the scale are measuring similar things. Test-retest reliability refers to the extent to which test-takers maintain the same rank-ordering across different testing occasions. It is an index of the stability of the scale across time. This is important if the scale is intended to measure a relatively enduring trait, such as intelligence or personality. An underlying assumption during the Sirius project has been that the BE and RD scales are also relatively enduring traits. If they were not, then efforts to mitigate the biases would not make sense.

Table 1: Overview of ABC Contents, Key Findings, and Scale Reliabilities

Scale	Facets	Number of Items	Key Findings	Internal Consistency Reliability	Test-Retest Reliability
Confirmation Bias (CB)	<ul style="list-style-type: none"> • Wason Selection • Information Search Decision Making • Evaluation / Weighting of Evidence • Evaluation / Weighting of Questions 	12	<ul style="list-style-type: none"> • Each task elicits CB with substantial variance across test-takers • Correlations between CB tasks represented in the ABC are low • No consistent correlations with background and Big-Five personality variables • Near 0 correlation with cognitive ability (Gf/Gc) 	.49 – .57	.46 – .62
Fundamental Attribution Error (FAE)	<ul style="list-style-type: none"> • Attitude Attribution • Good Samaritan • Quiz Role • Confession • Silent Interview • Attributional Style 	80-82 ratings across 8 items/tasks	<ul style="list-style-type: none"> • Each task elicits FAE with substantial variance across test-takers • Correlations between FAE tasks represented in the ABC are low • No consistent correlations with background and Big-Five personality variables • Near 0 correlation with cognitive ability (Gf/Gc) 	.82 – .85	.50 – .66

Scale	Facets	Number of Items	Key Findings	Internal Consistency Reliability	Test-Retest Reliability
Bias Blind Spot (BBS)	<ul style="list-style-type: none"> • N/A 	8	<ul style="list-style-type: none"> • Most test-takers display BBS • BBS scale has substantial variance across test-takers • BBS scale is relatively unidimensional • BBS results are not unique to a specific bias or bias-type • Higher cognitive workload (NASA-TLX) is associated with less BBS • Cognitive ability (Gf/Gc) and RD are associated with more BBS • Inconsistent correlations with personality measures and background/demographic variables 	.71 – .76	.66 – .73
Anchoring Bias (ANC)	<ul style="list-style-type: none"> • Numerical Priming • Selective Accessibility • Comparative Judgment • Self-Generated Anchor • Focalism 	15-17	<ul style="list-style-type: none"> • Each task elicits ANC with substantial variance across test-takers • Correlations between ANC tasks represented in the ABC are low • No consistent correlations with background and Big-Five personality variables • Small, positive correlations with Cognitive Reflection Test (CRT) and cognitive ability (Gf/Gc) 	.54 – .59	.62 – .67

Scale	Facets	Number of Items	Key Findings	Internal Consistency Reliability	Test-Retest Reliability
Representativeness Bias (REP)	<ul style="list-style-type: none"> • Base Rate Neglect • Sample Size Insensitivity • Conjunction Fallacy • Non-Random Sequence Fallacy 	19	<ul style="list-style-type: none"> • Each task elicits REP with substantial variance across test-takers • Correlations between REP tasks represented in the ABC are low • No consistent correlations with background and Big-Five personality variables • Moderate, positive correlations with RD, CRT, and cognitive ability (Gf/Gc) 	.55 – .66	.60 – .70
Projection Bias (PRO)	<ul style="list-style-type: none"> • False Consensus Effect • Knowledge Projection • Social Projection 	21	<ul style="list-style-type: none"> • Each task elicits PRO with substantial variance across test-takers • Correlations between CB tasks represented in the ABC are low • No consistent correlations with background and Big-Five personality variables • Small, positive correlations with CRT and cognitive ability (Gf/Gc) 	.54 – .61	.55 – .69
Recognition and Discrimination (RD)	<ul style="list-style-type: none"> • N/A 	ABC-1 = 13 ABC-2 = 9	<ul style="list-style-type: none"> • RD is largely unidimensional • RD associated with crystallized intelligence (Gc) markers 	ABC-1 = .79 – .82 ABC-2 = .72 – .80	ABC-1 = .68 – .77 ABC-2 = .61 – .72

Summary of Validity Argument for ABC

The accumulated evidence is consistent with the inference that the ABC is valid for its intended use. Despite the lack of “gold standard” marker tests and bias mitigation interventions, the available evidence indicates that the ABC scales show both convergent and discriminant validity and are sensitive to bias mitigation interventions. Convergent validity refers to evidence that two measures that purport to measure the same thing correlate with one another. Discriminant validity refers to evidence that two scales that purport to measure different things correlate at levels that suggest that the two scales are in fact measuring different things. For example, correlations between the Bias Instrument Coordinating Committee (BICC) scales developed by the Sirius research teams and ABC-2 scales, which were designed to measure the same bias constructs, should show reasonably high correlations between analog scales; for example, the BICC and ABC-2 Representativeness scales should – and did – correlate with one another. On the other hand, BICC and ABC-2 non-analog scales should show lower correlations than their analog scales. The evidence generally supported this conclusion.

Moreover, the extensive literature review conducted for this project enabled us to partition the content domain for each of the six bias constructs measured by the ABC into a set of facets that are both meaningful and comprehensive. That said, we emphasize that validation, especially for novel constructs such as those measured by the ABC, is an ongoing process. While the research record assembled during the course of this project is extensive and supports a solid validity argument, additional validity research is needed to extend our understanding of the constructs measured by the ABC.

Individual Difference Measurement of Biases

The frequency distributions of the individual bias scales indicate that test-takers differ substantially on each bias scale. As such, the ABC appears a promising step in adapting experimental paradigms to individual difference measurement. The overall validity argument suggests that the scales are generally meaningful, especially in their ability to detect changes in test-takers’ (1) bias susceptibility, and (2) knowledge of biases as a result of bias mitigation interventions.

While the RD scales both appear to be relatively unidimensional, the same is not true for the BE scales. With the possible exception of BBS, the BE scale-scores, as well as an overall battery score, are likely best understood as a concatenation of thematically related measures of the Sirius biases rather than unidimensional bias susceptibility measures. That is, they are essentially linear combinations of the items/scales of which they are comprised. Such measures are often referred to as “formative.” This created a trade-off between (1) maximizing capture of content representing the bias constructs, and (2) creating internally consistent, relatively unidimensional BE scales.

Future Research and Potential Applications

Although a great deal of research was done in the course of developing and evaluating the validity of the ABC, the study of bias within an individual difference framework is still largely in its infancy. As such, the research documented in this report can serve as a springboard for many other potential research programs.

The Sirius project encompassed six biases deemed important for intelligence analysis work. It should be noted, however, that there are many more cognitive biases that are worthy of investigation. These might include such constructs as hindsight bias, planning fallacy, and susceptibility to sunk costs, among others.

The validity data base can and should be extended to include criterion-related validation. This would involve identifying and measuring major work performance dimensions and correlating ABC test performance with work performance measures. Work performance may consist of both subjective (e.g., supervisor ratings) and quasi-objective measures (e.g., quantification of errors committed).

Instruments such as the ABC might also be used for training and development purposes as part of a formative assessment system. That is, performance on different components of the ABC might yield a profile of strengths and weaknesses with regard to knowledge of biases and bias susceptibility that would inform training program development. In this way, the ABC could be used not just predictively, but also diagnostically.

Another fruitful area for future research would focus on the bias mitigation interventions. For example, we, in conjunction with JHUAPL, are in the process of conducting formative evaluations of the ABC and the Sirius video game and instructional video interventions to determine what aspects of the best-performing interventions produced the greatest bias mitigation.

Acknowledgments

This work was accomplished in support of the Intelligence Advanced Research Projects Activity (IARPA) Sirius Program, BAA number IARPA-BAA-11-03. This technical data was first produced for the U.S. Government under contract 2009-0917826-016, and is subject to the Rights in Data-General Clause 52.227-14, Alt. IV (DEC 2007). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, expressed or implied, of IARPA or the U.S. Government or any of the authors host affiliations.

Table of Contents

1	GENERAL INTRODUCTION	16
1.1	Summary Project Description.....	16
2	PHASE 1.....	17
2.1	General Requirements and Goals for the ABC	17
2.2	Literature Review	18
2.2.1	Goals of Literature Review	18
2.2.2	Structure of Literature Review	18
2.2.3	Salient Conclusions from Literature Review	19
2.2.4	Use of Literature Review in Test Development.....	19
2.3	Test Development Process	20
2.3.1	Construct Identification.....	20
2.4	Test Development Process	24
2.4.1	Development of Item Prototypes.....	25
2.4.2	Cognitive Labs	26
2.4.3	Item Writing and Review.....	27
2.4.3.1	Item Generation	28
2.4.3.2	Script Writing and Production of Video-Based and Voice-Over SJTs.....	28
2.4.3.3	Item Review	28
2.4.4	Confirmation Bias.....	29
2.4.5	Fundamental Attribution Error	44
2.4.6	Bias Blind Spot.....	69
2.5	Phase 1 “Pre-Pilot” Studies	69
2.6	ABC-1 Field Test	81
2.6.1	Method.....	81
2.6.1.1	Participants	81
2.6.1.2	Study Design and Procedure.....	81
2.6.1.3	ABC-1 Scale Development	83
2.6.2	Results and Discussion	84
2.6.2.1	Descriptive Statistics	84
2.6.2.2	Reliability Analyses and Results	93
2.6.2.3	ABC-1 Intercorrelations	94
2.6.2.4	Scoring Modifications	95
2.6.2.5	Conclusions Regarding Structure and Individual Difference Measurement of Biases.....	96
2.6.3	Pretest Sensitization Study.....	96
2.6.3.1	Purpose.....	96
2.6.3.2	Method	97
2.6.3.3	Results and Discussion.....	97
2.6.4	ABC-1 Implementation.....	102
2.6.4.1	Development of Equivalent Forms	102
2.6.4.2	Completion Time for ABC-1	103
2.6.4.3	ABC-1 User Manual and Deployment Package	104

2.6.4.4	Identification and Resolution of Implementation Issues	104
2.7	ABC-1 Integrative Summary	105
3	PHASE 2.....	105
3.1	Significant Changes between Phases 1 and 2.....	106
3.1.1	Bias Instrument Coordinating Committee	106
3.1.2	Early and Increased Use of Online Crowdsourcing	106
3.1.3	Empanelled new TAG for Phase 2	106
3.1.4	Updating of Literature Review	107
3.2	Construct Identification	107
3.3	Cognitive Labs.....	110
3.4	Item Writing and Review	113
3.4.1.1	Item Generation	113
3.4.1.2	Script Writing and Production of Video-Based SJTs.....	113
3.4.1.3	Item Review	113
3.4.2	Anchoring Bias.....	113
3.4.3	Representativeness Bias.....	123
3.4.4	Projection Bias.....	133
3.5	Phase 2 Pre-Pilot Studies	143
3.6	ABC-2 Field Test.....	153
3.6.1	Method.....	153
3.6.1.1	Participants	153
3.6.1.2	Study Design and Procedure.....	153
3.6.1.3	ABC-2 Scale Development	155
3.6.2	Results and Discussion	156
3.6.2.1	Descriptive Statistics.....	156
3.6.2.2	Reliability Analyses and Results	163
3.6.2.3	Covariate Study.....	165
3.6.2.4	ABC-2 Intercorrelations	173
3.6.2.5	Correlations between ABC-1 and ABC-2.....	173
3.6.2.6	Conclusions Regarding Structure and Individual Difference Measurement of Biases.....	177
3.6.3	ABC-2 Pretest Sensitization Study	177
3.6.3.1	Purpose.....	177
3.6.3.2	Method	177
3.6.3.3	Results and Discussion.....	178
3.7	ABC-2 Implementation.....	186
3.7.1	Development of Equivalent Forms.....	186
3.7.1.1	Completion Time for ABC-2	186
3.7.2	ABC-2 User Manual	187
3.7.2.1	Identification and Resolution of Implementation Issues.....	187
3.8	Integrative Summary for ABC-2.....	188
4	SOFTWARE PLATFORM FOR COMPUTER-BASED DELIVERY OF TESTS	188

4.1	System Overview	188
4.2	System Architecture.....	188
4.3	Implementation	189
4.3.1	Setup and Configuration	190
4.3.2	Browser Requirements.....	190
4.3.3	Security and Authorization.....	190
4.3.4	Test-taker User Interface	191
4.3.5	Test Authoring.....	192
4.3.6	Instrumentation and Logging	192
4.3.7	Data Management	192
4.3.8	Scoring.....	192
4.4	Data Model.....	192
4.4.1	Top Level Data Model	192
4.4.2	Data model for stem parts and questions.....	193
4.4.3	Data model for stem parts	193
4.4.4	Data model for questions.....	194
4.4.5	Test tracker for participant responses	195
4.5	Data Export Format.....	196
4.6	User Interfaces.....	199
4.6.1	Administrator UI.....	199
4.6.2	Test-taker interface.....	200
4.6.3	Test authoring tool.....	203
4.6.4	Editing an Item	204
4.6.5	Editing Stem Parts	204
4.6.6	Editing Questions	210
5	VALIDITY ARGUMENT AND EVIDENCE FOR THE ABC.....	215
5.1	Propositions.....	216
5.1.1	Proposition 1	216
5.1.1.1	Rationale.....	216
5.1.1.2	Relevant Evidence.....	216
5.1.2	Proposition 2	217
5.1.2.1	Rationale.....	217
5.1.2.2	Relevant Evidence.....	218
5.1.3	Proposition 3	218
5.1.3.1	Rationale.....	218
5.1.3.2	Relevant Evidence.....	218
5.1.4	Proposition 4	219
5.1.4.1	Rationale.....	219
5.1.4.2	Relevant Evidence.....	219
5.1.5	Proposition 5	220
5.1.5.1	Rationale.....	220
5.1.5.2	Relevant Evidence.....	220
5.2	Summary	221

6	FUTURE RESEARCH	221
7	REFERENCES	223
	APPENDIX A OVERALL PROJECT ORGANIZATION	230
	MITRE Project Organization	230
	Program Management	230
	Technical Staff	230
	Subject matter experts	230
	Subcontractors	231
	ETS Project Organization	231
	Scientific Leadership and Staff	231
	Program Management	232
	Assessment Development and IT Staff	232
	Data Analysis	233
	Research Support	233
	Administrative Support	234
	Subcontractors to ETS and Their Roles	234
	Technical Advisory Group and Its Role	234
	APPENDIX B LIST OF ABBREVIATIONS.....	237

List of Figures

Figure 1: Wason Selection Paradigm “Shopping Malls” Task.....	31
Figure 2: Wason Selection Paradigm “Shopping Malls” Task.....	32
Figure 3: Information Search Decision Making Paradigm (“Snack Stand” Task).....	34
Figure 4: Information Search Decision Making Paradigm (“Snack Stand” Task).....	35
Figure 5: Information Search Decision Making Paradigm (“Snack Stand” Task).....	36
Figure 6: Information Search Decision Making Paradigm (“Car Comparison” Task).....	37
Figure 7: Information Search Decision Making Paradigm (“Car Comparison” Task).....	38
Figure 8: Information Search Decision Making Paradigm (“Car Comparison” Task).....	39
Figure 9: Information Search Decision Making Paradigm (“Car Comparison” Task).....	40
Figure 10: Evaluation/Weighting of Evidence Paradigm (“Intelligence Analyst” Task).....	42
Figure 11: Evaluation/Weighting of Questions Paradigm (“HR Department” Task).....	43
Figure 12: Attitude Attribution Paradigm (“Revolutionary” Task).....	45
Figure 13: Attitude Attribution Paradigm (“Revolutionary” Task).....	46
Figure 14: Attitude Attribution Paradigm (“Revolutionary” Task).....	47
Figure 15: Attitude Attribution Paradigm (“Revolutionary” Task).....	48
Figure 16: Good Samaritan Paradigm (Good Samaritan “Woman in White” Task).....	50
Figure 17: Good Samaritan Paradigm (Good Samaritan “Woman in White” Task).....	51
Figure 18: Good Samaritan Paradigm (Good Samaritan “Woman in White” Task).....	52
Figure 19: Good Samaritan Paradigm (Good Samaritan “Woman in White” Task).....	53
Figure 20: Attributional Style Paradigm (“Drew’s Good Day” Task).....	55
Figure 21: Attributional Style Paradigm (“What Causes Things?” Task).....	56
Figure 22: Confession Paradigm (“Sick for Party” Task).....	58
Figure 23: Confession Paradigm (“Sick for Party” Task).....	59
Figure 24: Confession Paradigm (“Sick for Party” Task, continued).....	60
Figure 25: Confession Paradigm (“Sick for Party” Task, continued).....	61
Figure 26: Quiz Role Paradigm (“Trivia” Task).....	63
Figure 27: Quiz Role Paradigm (“Trivia” Task).....	64
Figure 28: Silent Interview Paradigm (“Lying” Task).....	66
Figure 29: Silent Interview Paradigm (“Lying” Task, continued).....	67
Figure 30: Silent Interview Paradigm (“Lying” Task, continued).....	68
Figure 31: Histograms Depicting ABC-1 CB, FAE, and BBS Total Raw-Score Frequency Distributions by Form (1-3).....	90

Figure 32: Histograms Depicting ABC-1 CB, FAE, and BBS Total Raw-Score Frequency Distributions by Form (4-6).....	91
Figure 33: ABC-1 RD Total Raw-score Frequency Distributions.....	92
Figure 34: Self-generated anchoring item in one form of the cognitive labs.	111
Figure 35: Self-generated anchoring item in one form of the Field Trial Tests. The directions have been changed as a result of data from the cognitive labs, regarding the unclear directions. This item now includes “think of your ID as a dollar value,” for greater clarity.	112
Figure 36: Numerical Priming (“Front Load Washer”).....	114
Figure 37: Selective Accessibility (“Candy Jar”).....	116
Figure 38: Comparative Judgment (“Camera”).....	118
Figure 39: Self-Generated Anchor (“Earth Days”).....	120
Figure 40: Focalism (“Focalism Condo”).....	122
Figure 41: Base Rate Neglect (“Vet School”).....	125
Figure 42: Sample Size Insensitivity (“Brain Gain”).....	127
Figure 43: Sample Size Insensitivity (“Brain Gain”).....	128
Figure 44: Conjunction Bias (“Bomb Threat”).....	130
Figure 45: Non-Random Sequence Fallacy (“Retrospective Colored Marbles”).....	132
Figure 46: False Consensus (“Tipping”).....	134
Figure 47: Knowledge Projection (“Himalayas”).....	136
Figure 48: Social Projection (“Thesis”).....	138
Figure 49: Social Projection (“Thesis”).....	139
Figure 50: Social Projection (“Design”).....	141
Figure 51: Social Projection (“Design”).....	142
Figure 52: ABC-2 Field Trial Study Design.....	156
Figure 53: ABC-2 Anchoring Bias (ANC), Representativeness Bias (REP), and Projection Bias (PRO) Raw-Score Frequency Distributions.....	162
Figure 54: ABC-2 RD Total Raw-Score Frequency Distributions.....	163
Figure 55: ABC delivery platform system architecture.....	189
Figure 56: Participant login screen.....	200
Figure 57: Page with Text, Direction, Image, and Audio stem parts. (No Questions).....	201
Figure 58: Direction and Table (with image and text) stem parts, Likert and Semantic Differential questions.....	201
Figure 59: Directions and video stem part. The test taker must watch the entire video before the Continue button becomes enabled.....	202
Figure 60: Direction and Image stem parts and two Single Choice questions.....	202

Figure 61: Test Editor start page.....	203
Figure 62: Editing an Item	204
Figure 63: Editing a Text stem part	205
Figure 64: Editing an Image stem part.....	206
Figure 65: Image Browser	207
Figure 66: Editing Video stem part.....	208
Figure 67: Video file browser.....	209
Figure 68: Editing Table stem part with images and text	210
Figure 69: Editing Single Choice question	211
Figure 70: Editing Multiple Choice question.....	212
Figure 71: Editing Numeric Entry question.....	213
Figure 72: Editing Likert question	214
Figure 73: Editing Semantic Difference question.....	215

List of Tables

Table 1: Facets Associated with Behavioral Elicitation Tests for Each Phase 1 Bias.....	21
Table 2: Facets of Recognition and Discrimination (RD) Tests across Phase 1 Biases.....	24
Table 3: Summary of CB Pre-Pilot Research Studies	71
Table 4: Summary of FAE Pre-Pilot Research Studies	73
Table 5: Summary of BBS Pre-Pilot Research Studies	76
Table 6: Summary of RD Pre-Pilot Research Studies	78
Table 7: Number of Items by Scale and Facets Represented in ABC-1 Field Trial Study Forms.	82
Table 8: Allocation of BE and RD Scales to ABC Test Forms.....	84
Table 9: CB Elicitation Tasks and Paradigms across Forms.....	85
Table 10: FAE Behavioral Elicitation Tasks and Paradigms across Forms.....	85
Table 11: BBS Elicitation Items across Forms.....	87
Table 12: Distribution and Format of RD Items across Biases	88
Table 13: Descriptive Statistics for ABC-1 CB, FAE, BBS, and RD Scales by Form (Raw- scores).....	89
Table 14: Reliability Analysis of ABC-1 Confirmation Bias (CB), Fundamental Attribution Error (FAE), Bias Blind Spot (BBS), and Recognition and Discrimination (RD) Scales.....	93
Table 15: Intercorrelations between ABC-1 Scale-Scores.....	95
Table 16: ABC-1 Confirmation Bias (CB), Fundamental Attribution Error (FAE), Bias Blind Spot (BBS), and Recognition and Discrimination (RD) Pretest and Posttest Scale-scores.....	99
Table 17: Summary of ANCOVA Results for ABC-1 BE and RD Measures: Did the IARPA Video Group Do Better Than the Control Group, Controlling for Pretest Scores?.....	99
Table 18: NASA-TLX ratings following completion of the ABC pretest and posttest.....	100
Table 19: NASA-TLX ratings following completion of the ABC pretest and posttest.....	102
Table 20: Completion times for the ABC-1 (in minutes).....	103
Table 21: Working Definitions of Phase 2 Biases: Behavioral Elicitation of Cognitive Bias Measures.....	108
Table 22: Summary of Anchoring Bias (ANC) Pre-Pilot Research Studies	144
Table 23: Summary of Representativeness Bias (REP) Pre-Pilot Research Studies.....	148
Table 24: Summary of Projection Bias (PRO) Pre-Pilot Research Studies.....	150
Table 25: Summary of Recognition and Discrimination (RD) Pre-Pilot Research Studies.....	152
Table 26: Number of Items and Paradigms Represented in ABC-2 Field Trial Study Forms. ..	154
Table 27: Descriptive Statistics for ABC-2 ANC, REP, PRO, and RD Scales by Form (Raw- scores).....	157

Table 28: ANC Behavioral Elicitation Items and Paradigms across Forms.....	158
Table 29: REP Behavioral Elicitation Items and Paradigms across Forms.....	159
Table 30: PRO Behavioral Elicitation Items and Paradigms across Forms.....	160
Table 31: Distribution of RD Items across Biases and Forms.....	161
Table 32: Reliability Analysis of ABC-2 Anchoring Bias (ANC), Representativeness Bias (REP), Projection Bias (PRO), and Recognition and Discrimination (RD) Scales.....	164
Table 33: Correlations between ABC-2 Scaled Scores and Personality and Cognitive Ability Scales.....	165
Table 34: Zero-Order Correlations between Biases and Relevant Demographic Variables.....	168
Table 35: Zero-Order and Partial Correlations between Cognitive Reflection Test (CRT) and ABC-2 Scale-scores.....	169
Table 36: Correlations between ABC-2 and BICC Scales.....	169
Table 37: Correlations between ABC-2, BICC, and Personality, CRT, and Cognitive Ability Variables.....	171
Table 38: Correlations between ABC-2, BICC, and Demographic Variables.....	172
Table 39: Intercorrelations between ABC-2 Scale-scores.....	173
Table 40: Correlations between ABC-1 and ABC-2 Scale-scores.....	174
Table 41: Correlations between ABC-1 and ABC-2 Scale-scores, Disattenuated for Unreliability.....	174
Table 42: Correlations Between Cognitive Reflection Test (CRT) and ABC-1 and ABC-2 Scale-scores.....	175
Table 43: Zero-order Correlations Between ABC-1 Scale-Scores and Personality and Cognitive Ability Factor-Scores.....	176
Table 44: ABC-2 Anchoring Bias (ANC), Representativeness Bias (REP), Projection Bias (PRO), and Recognition and Discrimination (RD) Pretest and Posttest Scale-scores.....	179
Table 45: Summary of ANCOVA Results for ABC-2 BE and RD Measures: Did the IARPA Video Group Do Better Than the Control Group, Controlling for Pretest Scores?.....	180
Table 46: Summary of ANCOVA Results for Anchoring Bias Facets: Did the IARPA Video Group Do Better Than the Control Group, Controlling for Pretest Scores?.....	181
Table 47: Summary of ANCOVA Results for Representativeness Bias Facets: Did the IARPA Video Group Do Better Than the Control Group, Controlling for Pretest Scores?.....	182
Table 48: ABC-2 pretest and posttest knowledge projection composite raw-scores for IARPA Video, Control Video and No Pretest IARPA Video groups.....	184
Table 49: Summary of ANCOVA Results for Projection Bias Facets: Did the IARPA Video Group Do Better Than the Control Group, Controlling for Pretest Scores?.....	185
Table 50: ABC-2 completion times for three primary ABC-2 test forms in Field Test and Retest studies.....	186

Table 51: Access permissions for each role.....	191
Table 52: Example response data export file	198

This page intentionally left blank.

1 General Introduction

1.1 Summary Project Description

The purpose of the Intelligence Advanced Research Projects Activity (IARPA) Sirius Program, led by Program Manager Dr. Rita Bush, is to create serious video-games designed to train intelligence analysts to improve their explicit knowledge of, and ability to recognize, six well-known cognitive biases and to significantly mitigate the influence of those biases on behavior as a result of this training. The six biases and their definitions are as follows (definitions are those provided in the Broad Agency Announcement [BAA]):

1. **Confirmation bias.** The tendency to search for or interpret information in a way that confirms one's preconceptions. Often preceded by priming.
2. **Fundamental attribution error.** The tendency for people to overemphasize personality-based explanations for behaviors observed in others while underemphasizing the role and power of situational influences on the same behavior (also called attribution bias).
3. **Bias blind spot.** The tendency for an individual to be unaware of their own cognitive biases, even when the individual can recognize cognitive biases in others.
4. **Anchoring bias.** The tendency to rely too heavily or overly restrict one's attention to one trait or piece of information when making judgments. The information in question can be relevant or irrelevant to the target decision, as well as numerical or non-numerical. Includes focalism or the focusing illusion.¹
5. **Representativeness bias.** The tendency for people to judge the probability or frequency of a hypothesis by considering how much the hypothesis resembles available data. Also sometimes referred to as the *small numbers bias*.
6. **Projection bias.** The tendency to unconsciously assume that others share one's current emotional states, thoughts, and values.

The project was conducted in two phases. Phase 1 encompassed biases 1–3 and took place between October 2011 and September 2013, while Phase 2 encompassed biases 4–6 and took place between September 2013 and November 2015. The MITRE Corporation worked in collaboration with the Educational Testing Service (ETS) to develop a new standardized assessment, the Assessment of Biases in Cognition (ABC), to evaluate the effectiveness of the bias mitigation training. The ABC consists of a variety of innovative tasks grounded in, and adapted to varying degrees from, extant paradigms relevant to each of the six aforementioned biases. These tasks are complex scenario-based assessments that require test-takers to make decisions and solve problems presented in text, video, or voice-over audio formats, typically under conditions of uncertainty, time pressure, and/or rewards (and penalties).

Two versions of the ABC were developed for the IARPA Sirius Program. The first version consists of scales measuring three of the six biases listed above: Confirmation bias (CB), Fundamental attribution error (FAE), and bias blind spot (BBS). The second version assesses the remaining three biases: Anchoring bias (ANC), Representativeness bias (REP), and Projection

¹ This definition was revised by IARPA from the original BAA definition prior to the start of Phase 2.

bias (PRO). Each version of the ABC is administered online using a customized test delivery platform developed by the MITRE Corporation and takes, on average, anywhere between 45 and 60 minutes to complete. In addition, each version of the ABC includes three equated test forms that are used for evaluating bias mitigation training outcomes by comparing test-takers' performance on one form of the ABC post-training with pre-training baseline performance measured on an alternate, equated test form.

2 Phase 1

2.1 General Requirements and Goals for the ABC

A guiding principle of our test development was that the ABC should result in scores that are fair, reliable, and valid for its intended use in the Independent Validation and Verification (IV&V) phases of the Sirius Program in accordance with the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council of Measurement in Education, 1999, 2014). One objective was to develop a test instrument that would not disadvantage or be deemed offensive or upsetting to individual test-takers. To serve as a fair test instrument in the IV&V, it was also necessary to develop the ABC without the Sirius performer teams having specific knowledge of ABC content, nor for MITRE and ETS to intentionally create or select test items that would favor any specific bias mitigation approaches adopted by the Sirius research teams.

The development of the ABC presented unique challenges. One fundamental challenge was translating experimental paradigms that have been used to measure bias at the group level into construct-valid, reliable individual measures. Another challenge was measuring constructs that are potentially multi-dimensional, formative, or partially formative. In the case of *formative measurement* (see Bagozzi, 2007; Bollen & Lennox, 1991; Diamantopoulos, & Winklhofer, 2001; Edwards, 2001; Howell, Breivik, & Wilcox, 2007; MacCallum & Brown, 1993), the operational indicators (i.e., items, tasks) are said to *form* the construct. In the case of *reflective measurement*, constructs are said to cause, or determine, their operational indicators, which is what makes the indicators intercorrelate. For example, as we shall discuss, we found that several of the Phase 1 and Phase 2 biases appear to be formative, rather than reflective, constructs. This had substantial implications for measurement and implementation. Such challenges raised important questions about how to best model items in the ABC. For example, what is it about an item that causes people to score differently from one another? What are the most appropriate reliability metrics for the ABC scales? What are the sources of error variance associated with items, and how can they be mitigated? Most critically, it was important not to impose a reflective measurement model on a formative construct. Doing so potentially results in use of inappropriate reliability metrics, inappropriate combination and interpretation of scale-scores, and inappropriate inferences about the validity of the scales and overall ABC measure.

We also had to consider challenges posed by administering the ABC in the IV&V. For example, we were concerned that simply taking the pretest would influence performance on the immediate posttest, regardless of the bias mitigating intervention. For example, taking a pretest could sensitize examinees by causing them to pay attention to aspects of an intervention simply because those aspects were covered in the test (Goldstein, 1991). Similarly, taking the immediate posttest could influence performance on the delayed posttest due to a testing effect. A *testing effect* refers to the finding that being tested on previously acquired knowledge can directly influence (and most often improve) long-term retention of that knowledge to a greater extent

than additional study (e.g., Roediger & Butler, 2011; Roediger & Karpicke, 2006). We addressed these possibilities by conducting experiments that were adapted from the Solomon four-group design (Braver & Braver, 1988; Solomon, 1949), which extends the pre- and posttest control group design to include two additional control groups in which the pretest is not administered. Such experiments conducted with instructional videos produced by IARPA were also designed to measure the sensitivity of ABC measures to bias mitigation training.

There were additional test administration guidelines for the IV&V that informed the development of the ABC. First, we had to ensure that each version and test form of the ABC would measure all Mitigation and Behavioral Elicitation (BE) and Recognition/Discrimination (RD) constructs within 45 minutes to an hour. Second, the ABC was developed as a computer-based, online assessment that could be taken by examinees across a wide-range of testing sites and locations throughout the U.S. Third, in order to facilitate automated scoring of the ABC and avoid the need to hire human raters to score verbal written or oral responses, we only included selected-response item types and constructed response item types that required numeric answers. Last, the IV&V test-taker population was primarily U.S. college students who represent potential future IC analysts, as well as U.S. Intelligence Community (IC) analysts representing different organizations, functions, and experience working in the IC. Therefore, we had to ensure that the ABC was appropriate for these specific test-taker populations.

2.2 Literature Review

We began this project by compiling an extensive review of the extant literature (Gertner et al., 2011), which was then revised at the start of Phase 2 to include the latest thinking and scientific results pertaining to the three Phase 2 biases (Gertner et al., 2013).

2.2.1 Goals of Literature Review

Our original literature review goals were to (a) identify and understand the Sirius project biases, (b) identify and evaluate any individual difference measures of those biases that have been developed, (c) incorporate intelligence community (IC) literature into the review to understand how the Sirius project biases play out in the work of intelligence analysts, (d) identify bias mitigation techniques and determine how effective those have been, (e) identify experimental paradigms that have been used to establish the existence of the Sirius project biases and determine the extent to which they can be adapted for individual differences measurement purposes, (f) gain an understanding of cognitive processes underlying each cognitive bias, and (g) integrate the literature review and formulate conclusions relevant to development and evaluation of the ABC – and, by extension, the ABC-2.

2.2.2 Structure of Literature Review

The literature review begins with an introduction section that discusses previous research examining cognitive biases and their relationship to individual difference variables such as personality and cognitive ability measures. The introduction also provides a brief overview of bias mitigation research; the background and rationale for the Sirius Program; and previous attempts to develop standardized measures of cognitive bias. The literature review then provides an in-depth treatment of each of the six Sirius Program biases, discussing the nature, proposed causes, measurement, and mitigation of the bias. The literature review concludes by discussing the implications of the previous sections for developing the ABC.

2.2.3 Salient Conclusions from Literature Review

The primary observations and conclusions that were drawn from our review of the literature are as follows:

- A great deal of research has been conducted on most of the biases studied in the Sirius Program. In general, the research has shown that these biases exert powerful effects on human cognition and behavior, are largely ubiquitous, and are quite resistant to attempts to mitigate or eliminate them. These biases also tend to occur under conditions that are typical of intelligence analysis work, such as the need to work under significant time pressure, and interpret and weigh the relevance of data and behavior that are frequently ambiguous, incomplete, or duplicitous.
- Research that has addressed the existence and robustness of the biases has been far more conclusive than research on underlying causes (especially cognitive mechanisms) of the biases. In general, there is much debate, but virtually no empirical closure, regarding causes of these biases.
- These biases have been defined in various ways by different researchers. Perhaps the best example of this is the projection bias, which has been defined in several distinct ways that do not align precisely with the IARPA BAA definition. Similarly, the anchoring bias appears to manifest in at least two ways, one that involves System 1 cognition and another involving System 2 cognition². Researchers have suggested that entirely different bias mitigation methods should be used to address these two different manifestations of what had originally been regarded as a unitary bias.
- Most of the literature dealing with cognitive biases has taken place within the experimental tradition, meaning that researchers have not been especially interested in developing measures of individual differences in susceptibility to bias elicitation or individual differences in bias mitigation.

2.2.4 Use of Literature Review in Test Development

The literature review was used in several ways during the development of the ABC. First, it helped inform the drafting of a document summarizing the measurable content domain for each bias. Second, we sketched the initial set of task prototypes designed to elicit and measure each of the 6 biases by adapting established paradigms identified in the literature review. Last, observations and conclusions drawn from the literature review were important for drafting the ABC-1 and ABC-2 research plans, as well as designing of first round of pre-pilot research studies in both Phases 1 and 2 of the Sirius Program.

² *System 1 thinking* includes automatic mental processes and affective reactions that occur quickly and effortlessly, often without conscious attention or even awareness. This type of thinking underlies routine, well-learned activities without having to consciously focus attention on the steps required to perform those activities. By contrast, *System 2 thinking* occurs more slowly and involves conscious, effortful mental deliberation (See Gertner et al., 2011, 2013).

2.3 Test Development Process

Below is a list of the key steps involved in the development of the Phase 1 ABC (ABC-1) Behavioral Elicitation (BE) and Recognition and Discrimination (RD) Tests.

- Construct Identification
- Development of Item Prototypes
- Cognitive Labs
- Item Review
- Pilot Testing
- Assembling and Authoring in Test Administration Platform
- Final Field Test
- Preparation and Delivery of Final Test Forms

The steps were conducted in both sequential and iterative fashion. We discuss the details of each step in the subsequent sections.

2.3.1 Construct Identification

A routine part of test development is specification of what the test is intended to measure. At a minimum, this should include: (a) definition of the constructs to be measured by the test; and (b) a listing of its facets, or sub-constructs. Consistent with the definitions in the Sirius BAA, and the literature review conducted for Phase 1 of this project (Gertner et al., 2011), we defined the Phase 1 bias constructs as follows:

- **Confirmation bias.** The tendency to search for or interpret information in a way that confirms one's preconceptions.
- **Fundamental attribution error.** The tendency to underestimate the degree to which situations determine others' behavior and to overestimate the degree to which others' dispositions (including attitudes and beliefs) determine their behavior.
- **Bias blind spot.** The tendency for individuals to assume and recognize the existence and operation of bias to a greater extent in other people's behavior than in their own behavior; and the tendency for individuals to be unwittingly biased in their perceptions of their own attributes (including bias susceptibility), actions, predictions, and decisions.

Table 1 shows the facets we proposed for each of the ABC bias constructs based upon our review of the research literature in order to specify in more detail definitions of the bias constructs. We adopted a bootstrap approach, whereby we sought to capitalize on useful distinctions made in the literature and combine them with insights derived from item prototype development, empirical results of our pre-pilot research and Field Test studies, and input from the TAG, Subject Matter Experts (SMEs), and the IV&V team in order to clarify the boundaries of each construct throughout the test development process. Our goal was to be overinclusive with regard to facets making up the constructs in the ABC (Loevinger, 1957) so as not to run the risk of excluding facets necessary to provide a complete operationalization of the construct under investigation. By extension, we also produced an assessment that was fairer to the Sirius Program research teams.

Table 1: Facets Associated with Behavioral Elicitation Tests for Each Phase 1 Bias

Confirmation Bias (CB)	Fundamental Attribution Error (FAE)	Bias Blind Spot (BBS)
<p><u>CB Facet 1:</u> Tendency to access prior knowledge or beliefs or to generate/adopt an initial hypothesis that conforms to previously-formulated hypotheses, attitudes, or beliefs</p>	<p><u>FAE Facet 1:</u> Tendency to overestimate the impact that a person's personality, attitudes, or beliefs have on their actions</p>	<p><u>BBS Facet 1:</u> Tendency to underestimate or to not recognize one's susceptibility to cognitive bias</p>
<p><u>CB Facet 2:</u> Tendency to seek evidence consistent with a previously-made position or decision, or with a previously formulated hypothesis; tendency to overlook, ignore, or discount disconfirming evidence</p>	<p><u>FAE Facet 2:</u> Tendency to lack awareness of situational constraints on others' behavior</p>	<p><u>BBS Facet 2:</u> Tendency to underestimate one's bias susceptibility relative to one's peers</p>

Table 1: Facets Associated with Behavioral Elicitation Tests for Each Phase 1 Bias

Confirmation Bias (CB)	Fundamental Attribution Error (FAE)	Bias Blind Spot (BBS)
<p>CB Facet 3: Tendency to misinterpret or distort evidence that disconfirms a previously formed hypothesis or that is inconsistent with a prime</p>	<p>FAE Facet 3: Tendency to have unrealistic expectations for appropriate conduct, given certain "strong" situational influences</p>	<p>BBS Facet 3: Tendency for people to unwittingly assume that their own perceptions reflect objective reality, and to assume that perceptions of reality different from their own reflect bias on the part of others ("naïve realism")</p>
<p>CB Facet 4: Tendency to assign greater weight to evidence that is consistent with an initially formed hypothesis or prior belief than to disconfirming evidence</p>	<p>FAE Facet 4: Tendency to make incomplete corrections to dispositional attributions when exposed to information that favors situational attributions (e.g., when a situation is made more salient)</p>	<p>BBS Facet 4: Tendency for people to be either under- or overconfident regarding the accuracy of their judgments.</p> <p>In this context, under- or overconfidence includes:</p> <ul style="list-style-type: none"> • Under/Overestimation of one’s actual ability, performance, level of control, or chance of success (e. g, on a specific test) • Under/Over-placement: When people believe themselves to be worse/better than others, such as when a majority of people rate themselves below/above the median • Under/Over-precision: Excessive un/certainty regarding the accuracy of one’s beliefs (e.g., providing an overly wide or narrow confidence interval when making a quantitative judgment about the accuracy of one's test responses)

Table 1: Facets Associated with Behavioral Elicitation Tests for Each Phase 1 Bias

Confirmation Bias (CB)	Fundamental Attribution Error (FAE)	Bias Blind Spot (BBS)
<p>CB Facet 5: Tendency not to revise an initially formed hypothesis to be consistent with new evidence that is consistent with a different hypothesis (i.e., tendency not to think in a Bayesian way)</p>		

The reason for developing facets was more practical than theoretical. It was important that we developed measures that encompassed the full measurable construct domain in order to formulate a coherent and convincing validity argument. Clear operational definitions also provided a roadmap for item writing and review.

Facets were developed using multiple methods. The methods that we used differed for the three biases to the extent that their respective literatures provided different information. In the case of the confirmation bias, we adapted a process model proposed by Klayman and Ha (1987) that seemed to make sense as a point of departure. In the case of other biases, the partitioning of the content domain was largely based on a parsing of the definitions of the constructs. We, of course, adhered closely to the descriptions of the content domain of each bias provided by IARPA in the BAA.

The recognition and discrimination (RD) measures, while applying to the same three Phase 1 biases, assessed constructs that were different from the behavioral elicitation measures and, as such, had different definitions and facets. Unlike the behavioral elicitation of cognitive bias measures, the definitions of the recognition and discrimination measures were essentially the same for each bias, except that when they were operationalized in each test, they were populated with content specific to each of the three Phase 1 biases.

Our definitions of the RD constructs were as follows:

Recognition of Cognitive Bias: Knowledge of a given bias’ definition, key characteristics, illustrative examples, and relevance to judgment and decision-making tasks similar to those faced by intelligence analysts.

Discrimination among Cognitive Biases: Ability to distinguish and identify a given bias from among other biases; knowledge of areas of overlap between that bias and the other two Phase 1 biases; ability to distinguish the effects of that bias from the effects of other biases on various judgment and decision-making tasks similar to those faced by intelligence analysts.

Table 2 shows facets of the recognition and discrimination (RD) construct in the ABC.³

Table 2: Facets of Recognition and Discrimination (RD) Tests across Phase 1 Biases

Recognition of Cognitive Bias	Discrimination among Cognitive Biases
<p>RD Facet 1: Knowledge of definitions, labels, key characteristics, and examples (whether described statically or depicted dynamically in scenario-based items) of this bias; ability to identify both prototypical and more peripheral instances of this bias (to the extent that those peripheral instances do not overlap conceptually with other Phase 1 biases)</p>	<p>RD Facet 4: Knowledge of the primary differences between this cognitive bias and the other two Phase 1 biases</p>
<p>RD Facet 2: Knowledge of how this bias can impact intelligence analyst-type tasks</p>	<p>RD Facet 5: Ability to identify this bias when it is embedded among the other two Phase 1 biases (as well as additional biases)</p>
	<p>RD Facet 6: Knowledge of the existence and nature of the overlap between this bias and both of the other two Phase 1 biases</p>
<p>RD Facet 3: Knowledge of conditions that make behavior reflecting this bias more and less likely to occur</p>	<p>RD Facet 7: Ability to distinguish peripheral instances of this bias from peripheral instances of other Phase 1 biases, so long as they do not overlap conceptually</p>
	<p>RD Facet 8: Ability to distinguish how this bias can impact intelligence analyst-type tasks from how other Phase 1 biases can impact intelligence analyst-type tasks</p>

Having partitioned the Phase 1 content domain, we proceeded to develop items intended to operationalize the content. That process is described in the next sections.

2.4 Test Development Process

The Phase 1 Behavioral Elicitation (BE) and Recognition and Discrimination (RD) tests of CB, FAE, and BBS were developed iteratively in multiple steps. We developed BE and RD items and

³ We initially separated the (1) Recognition and (2) Discrimination aspects of the RD construct because it was not known whether they were empirically separate facets of the overall construct. Subsequent empirical analyses showed that recognition and discrimination were, in fact, part of the same construct, and this distinction was no longer deemed necessary.

scoring rubrics representing all bias types to measure the content domain specified in Tables 1 and 2.

2.4.1 Development of Item Prototypes

We formulated a number of item prototypes. Many of the items⁴ consisted of complex scenario-based assessments with text and graphic image stimuli and required test-takers to respond under conditions of uncertainty, time pressure, and/or rewards and penalties.

BE item prototypes were developed using a variety of sources, including:

- Adaptation of individual-difference analogs of experimental paradigms (e.g. Wason card selection task [Fischer et al., , 2011]; Quiz-Role paradigm [Gawronski, 2003; Ross, Amabile, & Steinmetz, 1977]);
- Adaptation of extant scientific literature relevant to the construct of interest (e.g., overconfidence as it relates to bias blind spot [Moore & Healy, 2008; Pallier et al., 2002; Stankov & Lee, 2008]);
- Informed professional judgment (required, for example, for adaptation of some experimental paradigms and rational/empirical justification of probable construct validity of item prototypes);
- Abstraction from intelligence analyst performance domains relevant to one or more biases (c.f. critical incidents adapted from in-depth interviews with several IC SMEs; Fingar, 2011; Heuer, 1999; Hoffman, Henderson, Moon, More, & Litman, 2011; Hutchins, Pirolli, & Card, 2007; Keibell, Muller, & Martin, 2010; Plous, 1993; Tecuci, Schum, Boiceu, Marcu, & Hamilton, 2010; Williams, 2010);
- Description or depiction of social interpretation and prediction activities relevant to a particular bias (cf. Gawronski, 2003; Ross et al., 1977; Snyder & Frankel, 1976); and
- Description of one of the construct or facet definitions of a given bias.

Sections 2.4.4 - 2.4.6 describe in greater detail the task prototypes developed for each of the Phase 1 bias constructs. For each BE task prototype, ETS assessment development specialists designed PowerPoint mock-ups and programmed functional tasks for administration in cognitive labs (described in Section 2.4.2) and pre-pilot research studies (described in Section 2.5). In the case of video and voice-over situational judgment tests (SJTs), scripts were written and filmed

⁴ The distinction between an "item" and a "task" is occasionally slightly blurry. In general, items refer to short text- or video-based item stems coupled with a single set of multiple-choice, selected-response options. Tasks refer to more elaborate BE measures that typically require participants to view multiple screens consisting of text- or video-based stimuli, usually involving a series of decisions about how much, or what type of, information to access to make one or more ratings or decisions. The distinction is primarily based on convention, with items referring more to the type of stimuli typically found in many traditional psychometric tests (e.g., Likert-type personality items, multiple-choice items often found in intelligence and achievement tests); and tasks referring to adaptations of social and cognitive paradigms found in the experimental psychology literature. In both cases, however, the effect on scoring is the same: whether an item or task, the outcome is used as an operational indicator of either recognition and discrimination, or behavioral elicitation, of biases measured in this project.

(or audio-recorded in the case of the voice-over items) in collaboration with CML to provide necessary content and context.

RD items were, in general, not as complex as BE items. We began by developing a large number of items for possible inclusion in the RD test. 100 RD items were written by the ABC development team, based on:

- Our professional and scientific knowledge of the Phase 1 biases based on extensive review of the extant literature (Gertner et al., 2011)
- Case studies of intelligence analysis (Beebe & Pherson, 2011)
- Critical incidents adapted from in-depth interviews with several IC SMEs

The 100 items comprising this initial item pool were both text-based and video-based, and represented all Phase 1 biases. Attention check items were also included so that we could screen our data to maximize the quality of the data used in the test development process. We also included items the correct answer for which was “None of the above.”

These BE task prototypes and RD items were reviewed by ETS staff to ensure that their content conformed to established testing standards for fairness and sensitivity to test-takers. In addition, they were reviewed by MITRE, IARPA, and Johns Hopkins University Applied Physics Lab (JHUAPL), the TAG, and, to the extent possible, IC SMEs. Based upon these reviews, we modified the original item prototypes and developed new item prototypes.

2.4.2 Cognitive Labs

Cognitive laboratories ("*cognitive labs*," for short) refer to a class of small-scale laboratory-based research studies that use verbal reporting and probing techniques as well as other observational and qualitative methods in order to collect information about the psychological processes and response behaviors test-takers exhibit while attempting to answer test questions. For our purposes, cognitive labs incorporated techniques adapted from usability and think aloud studies. When conducted early in the test development process, cognitive labs can provide rich qualitative data about particular test items that allow for adjustments to be made to those items prior to development of the items on a large scale. In addition, the results from cognitive labs may indicate that the test items are tapping cognitive processes consistent with the constructs we are targeting, thereby contributing evidence relevant to the overall validity argument (Kane, 1992; Messick, 1995). Such evidence is referred to as *response process validity*. This type of data collection is a unique way to learn otherwise unknown information during a typical assessment. Detailed and individual reactions to items can provide helpful information on item difficulty, novel item types or item format issues, difficulties for specific groups of test-takers, and test-taker preferences (e.g., Almond et. al., 2009; Johnstone, Bottsford-Miller, & Thompson, 2006; Martiniello, 2008; Sato, Rabinowitz, Gallagher, & Huang, 2010).

For behavioral elicitation test items, we developed interviewer protocols that detailed the instructions that study participants follow while performing the task, as well as instructions that the interviewer/experimenter followed in providing any verbal prompting for participants and coding behavioral observations. One type of protocol required participants to talk out loud while they were performing a given task, verbalizing any thoughts that came to mind, in order to provide real-time information about participants' conscious thoughts, problem solving strategies, and emotional reactions (Birns, Joffre, Leclerc, & Paulsen, 2002; Ericsson & Simon 1993;

Leighton, 2004). The interviewer/experimenter's role was to avoid interrupting the participant, except for encouraging him/her to "keep talking" when the participant was quiet for a period of time. Following the session, the interviewer asked the participant targeted questions, retrospectively probing for additional information about their thoughts and experiences working on the task.

Another type of protocol instructed the interviewer/experimenter to retrospectively probe participants' thoughts and reactions immediately after they completed a task or series of tasks (Dumas & Redish, 1993; Ericsson & Simon 1993). Because the process of thinking aloud while performing a behavioral elicitation task might confound the measure of the target bias by, for instance, increasing the level and quality of their introspection, the interviewer/experimenter alternated using concurrent think aloud and retrospective probing methods when conducting cognitive labs for each of the behavioral elicitation tasks. Participants' verbalizations, facial expressions/body movements, and user-computer interactions (e.g., key presses, mouse movements) were recorded using Morae software (see <http://www.techsmith.com/morae.asp>).

We conducted two cognitive lab studies during Phase 1. The first round, which was conducted with 13 ETS employees for 12 BE item prototypes, focused on usability concerns. Do examinees understand the task requirements? Are there any particular task elements or features that facilitate or hinder task performance? We made modifications to the BE items based upon our observations from this initial study. In the second round of cognitive lab studies, which was conducted with 9 ETS employees and graduate student interns for 10 BE item prototypes, we continued to examine usability concerns, but in addition, we examined thinking strategies reflected in the verbal protocols.

In general, participants found the task instructions and requirements to be clear and the task designs to be appealing and engaging. Study participants at the University of Cincinnati were also given a usability survey based on Finstad's (2010) four-item Usability Metric for User Experience (UMUX), and they too gave very high usability ratings for these and other item types. We also identified task elements that participants still found to be unclear, distracting, or too demanding. Verbal protocols from concurrent thinking aloud as well as retrospective verbal accounts of response behaviors also indicated that conscious decision making and problem solving strategies varied considerably across tasks and participants. Both qualitative and quantitative analyses showed no indication that participants were performing the BE tasks differently when given concurrent think aloud instructions as compared to being given retrospective questions alone. Interestingly, no participants reported any specific knowledge or awareness of underlying aims of the assessments.

2.4.3 Item Writing and Review

Each round of item generation involved item writing based on the prototypes described above; followed by item review; and, for items administered in a multimedia format, videotaping, editing, and programming.

2.4.3.1 Item Generation

Following the development and evaluation of the BE task prototypes, we created clones⁵ of those prototypes with the objective of substantially increasing the item pool for the ABC-1. We developed over 500 BE items and 100 RD items in two rounds.

2.4.3.2 Script Writing and Production of Video-Based and Voice-Over SJTs

Working closely with CML, we wrote and videotaped a total of 91 scripted scenarios both for BE and RD items in two rounds of film production. 25 of the scripted scenes were filmed at the ETS Princeton, NJ, campus with local professional actors and ETS employees. The remaining scenes were filmed in Louisville, KY, also with local professional actors and ETS employees.

Scripts were written, filmed, and edited with the following objectives and logistical constraints in mind:

- Because of timing constraints on the ABC, the videos were kept short, and none were longer than 45 seconds.
- We filmed “clones” and multiple takes and slight variations of scenes with the same or different actors in order to increase the pool of candidate items and permit exploratory investigation of scene features that enhanced bias elicitation and user acceptability.
- However, we utilized our pool of actors in such a way as to accomplish as much as possible with the fewest number of actors.
- That said, actors were carefully selected for scenarios for maximum effectiveness; for example, scripts calling for dominant individuals were matched to actor characteristics; the same actor was not used for two or more related scenarios if inclusion of those actors would be confusing or inappropriate.

In order to generate additional clones of video-based SJTs with good psychometric properties, we developed 14 voice-over audio recordings of scripts that were written subsequent to the second round of film production.

2.4.3.3 Item Review

Item review was a structured process in which psychological testing and subject matter experts reviewed items with respect to criteria such as (a) clarity, (b) lack of ambiguity and vagueness, (c) ensuring that the items do not assume knowledge specific to the intelligence analyst job, and (d) getting a sense of the difficulty of each item to help ensure adequate variance in the scales we developed for each construct. In addition, the IV&V team and TAG reviewed items during each of the three TAG meetings held at ETS during Phase 1. For items that had content specific to intelligence analysis work, additional reviews were performed by IC SMEs at MITRE.

Items were also reviewed for sensitivity to EEO protected class bias and fairness issues. ETS requires a formal, documented Fairness Review by specially trained staff for compliance with

⁵ A clone is an item or task that is structurally the same as the prototype item or task. Clones are designed to have measurement properties that are, as nearly as possible, identical to the prototypes from which they were derived, and to leverage prototypes shown through pilot test research to be functioning well in the ABC.

ETS Fairness policies for all assessments and for all other materials intended for use by more than 50 people outside of ETS. Fairness review is the stage in which the item is reviewed for bias and sensitivity content concerns.

The Fairness Review process helps ensure that diverse audiences will understand the test or assessment materials and not be offended by them. An important aspect of fairness is treating people with impartiality regardless of such characteristics as gender, race, ethnicity, or disability that are not relevant to the test being given. In addition, the Fairness Review process helps ensure that only construct relevant factors affect test takers' scores. Test items that cause group differences because of construct-irrelevant factors are not fair. For example, unless it is part of the construct being measured, all culturally specific content should be removed or replaced, so not to cause bias in test use.

In the next series of subsections, we describe in some detail the paradigms used to generate item prototypes to operationalize the BE constructs. In addition, we include screenshots of illustrative items within each paradigm. It should be noted that other paradigms were considered based on our partitioning of the BE content domains. However, some paradigms were dropped due to excessive length, complexity, and/or poor psychometric properties in pilot testing.

2.4.4 Confirmation Bias

Wason Selection Paradigm

ABC-1 tasks representing the “Wason Selection Paradigm” were modeled on the classic experimental paradigm known as the Wason card-selection task (Wason, 1966, 1968), which has been frequently used by researchers to demonstrate individuals' tendency to test hypotheses by considering confirming rather than disconfirming evidence. In this task, participants see an array of four cards, each of which has a letter on one side and a number on the other side. Each of the cards shows a letter and number that are either: a vowel, a consonant, an even number, or an odd number. Participants are asked which cards one would have to turn over to determine the truth or falsity of the following statement: "If a card has a vowel on one side then it has an even number on the other side." Given this set of cards, one can determine the rule to be false by finding either the card showing the vowel “A” or the card showing the number “7” to be inconsistent with it, or one can determine the rule to be true by finding both of these cards to be consistent with it. However, individuals are most likely to select only the card showing a vowel or the card showing a vowel and the one showing an even number. They seldom select either the card showing a consonant (which would give them no useful information) or the one showing an odd number (Cosmides, 1989; Evans, 1982; Evans, Newstead, & Byrne, 1993; Tweney & Doherty, 1983; Wason & Johnson-Laird, 1972). In other words, individuals tend to adopt a strategy in which they seek to confirm the terms of the rule rather than falsify the rule by, for instance, testing the possibility that a vowel may be behind the odd number card.

The Wason Selection Paradigm items that were developed for the ABC-1 also involve testing the truth or falsity of a proposition, but they included several features designed to make the problems more realistic than the Wason card selection task. First, each proposition was framed as a probabilistic, rather than a causal, relationship (e.g., “If a community has a shopping mall, there is an increased likelihood of a flu outbreak.”). Second, individuals must select among eight information icons that correspond to facts about eight cities listed in a table before submitting a final answer. For example, an exhaustive test to determine the truth or falsity of the statement,

“If a community has a shopping mall, there is an increased likelihood of a flu outbreak,” would require test takers to sample/click every “Mall” and every “No Flu” icon on the map. Moreover, since the probabilistic framing of the proposition changes which information is relevant as compared to causal framing (e.g., “If a community has a shopping mall, there will be a flu outbreak.”), it also becomes important for test-takers to learn about relative outcomes for communities that do and do not have shopping malls. What percentage of relevant icons would a person sample before s/he is convinced there is a general rule? Confirmation bias would be indicated by the tendency to sample a greater number of “Mall” icons than other icons. Extreme confirmation bias might be shown when participants sample each and every “Mall” icon, and nothing else.

We developed three other variants of this task that are formally similar to the “Shopping Mall” problem, but require participants to test the truth or falsity of the following propositions: 1) If a team is playing on its home field, it has an increased likelihood of winning a baseball game; 2) Cities that lower taxes tend to see industrial job growth; and 3) Towns where PCT is used tend to see higher incidence of Chisolm syndrome in children born there. CB was measured as the proportion of information icons sampled that represent confirmatory selections (e.g., sampling “Mall” icons).

Do communities with shopping malls tend to experience flu outbreaks?

Directions: Use information in the table to figure out if the theory stated below is true or false.

Note that clicking on one or more of the **Show** buttons to reveal a second piece of information is necessary to determine if the theory below is true or false.

You will be scored on both the correctness of your answers AND the number of **Show** buttons you click on, so use as few clicks as necessary to find the answer.

Click on one or more of the Show buttons in the table to determine if the theory below is true or false.

Theory: If a community has a shopping mall, it has an increased likelihood of experiencing a flu outbreak.

- True False



	1st piece of information	2nd piece of information
Northwood	Mall	Show
Clear Lake	No Flu	Show
Hampton	Flu	Show
Osage	No Mall	Show
Charles City	No Mall	Show
Waverly	Mall	Show
Sumner	No Flu	Show
Ridgeway	Flu	Show

Submit



Figure 1: Wason Selection Paradigm “Shopping Malls” Task

Do communities with shopping malls tend to experience flu outbreaks?

Directions: Use information in the table to figure out if the theory stated below is true or false.

Note that clicking on one or more of the **Show** buttons to reveal a second piece of information is necessary to determine if the theory below is true or false.

You will be scored on both the correctness of your answers AND the number of **Show** buttons you click on, so use as few clicks as necessary to find the answer.

Click on one or more of the Show buttons in the table to determine if the theory below is true or false.

Theory: If a community has a shopping mall, it has an increased likelihood of experiencing a flu outbreak.

True False



	1st piece of information	2nd piece of information
Northwood	Mall	Flu
Clear Lake	No Flu	Mall
Hampton	Flu	Mall
Osage	No Mall	Flu
Charles City	No Mall	No Flu
Waverly	Mall	Flu
Sumner	No Flu	No Mall
Ridgeway	Flu	No Mall

Submit



Figure 2: Wason Selection Paradigm “Shopping Malls” Task

Information Search Decision Making Paradigm

The tasks representing the “Information Search Decision Making Paradigm” were adapted from research studies that have demonstrated confirmation bias in evidence seeking behavior (e.g., Cook & Smallman, 2008; Fischer, Greitemeyer, & Frey, 2008; Fischer et al., 2011; Frey & Rosch, 1984) and involve making an initial decision based on a fictitious scenario and then being asked to seek additional information from a pool of confirmatory and disconfirmatory evidence in order to make a final decision. Individuals are presented with a particular scenario and asked to make an initial decision. Participants are then presented with additional pieces of information, half of which support one of the initial choices and the other half support the alternative choice, and given the opportunity to access as many pieces of the additional information as they would like before making a final decision between the two choices. Test-takers are shown 8 additional pieces of information, half of which support the organic business idea and other half the diet product idea. Thus, half of the information was consistent (and half was inconsistent) with the participants’ initial decision (Fischer et al., 2011). CB is measured as the difference between the number of selected, inconsistent pieces of information and the number of selected, consistent pieces. In addition to the “snack stand” decision making task, we administered two clones that asked participants to choose between two different types of bakeries to open or exercise classes to offer.

We developed another set of tasks that also involved information search in the context of making a decision between two products—Car Comparison, Cruises, Making Music, and Working Out tasks. In these tasks, test-takers sample brief comments from an unbalanced pool of comments. After making an initial product selection, test-takers are presented with an assortment of comments from which to sample, most of which favor the initial product selection (as indicated by thumbs up and thumbs down icons displayed next to the comment headers). Mild time pressure and a fake monetary incentives are also provided in order to encourage "System 1" responding. System 1 responding involves automatic mental processes and affective reactions that occur quickly and effortlessly, often without conscious attention or even awareness (Gertner et al., 2011). It is this mode of thinking that often leads to the biases of the sort measured by the ABC-1.

Directions: Answer the question below.

What type of snack stand would you like to open? You want to open a new kind of snack stand. You have two good ideas and have to decide on one. You could either choose to sell diet products (e.g. low-fat and low-carb products) or organic products (e.g. vegetables grown without pesticides or genetic manipulation). Both the diet and organic industries seem to be very popular in your city.

Click on the type of stand you would like to open. You will get more information later on and will be able to change your decision if you wish.



organic snacks diet snacks

Submit



Figure 3: Information Search Decision Making Paradigm (“Snack Stand” Task)

Directions: Answer the question below.

You receive some additional information:

- Organic snacks are healthier than diet snacks.
- Diet snacks are less expensive than organic snacks.
- Organic snacks taste better than diet snacks.
- Diet snacks are more popular than organic snacks.
- Organic snacks are better for the environment than diet snacks.
- Diet snacks are more appealing to women shoppers.
- Organic snacks are better for the local economy.
- Diet snacks have more attractive packaging than organic snacks.

Select the type of stand you would like to open based on this additional information. You will get more information later on and will be able to change your decision if you wish.



organic snacks diet snacks

Submit



Figure 4: Information Search Decision Making Paradigm (“Snack Stand” Task)

Test 4 of 21 Test Time Elapsed 0:02:41

Organic snacks are healthier than diet snacks

Diet snacks are less expensive than organic snacks

Organic snacks taste better than diet snacks

Diet snacks are more popular than organic snacks

Dietary products are immensely popular in the United States. For instance, in one survey, approximately 42%, or an estimated 96 million, of people living in the U.S. dieted during 2008. Of those people, approximately 56 million attempted to lose weight and 40 million attempted to maintain their weight. Another recent survey of Americans consumption of low-calorie foods and beverages found that 54% of adults dieted in 2010. This was a major increase from 33% in 2004 and the highest percentage recorded since 1986. Indeed, consumption of dietary products is the most popular weight loss method, with diet soft drinks being the most popular, low-calorie, sugar-free product. According to recent surveys, 86% of dieters cut down on foods high in sugar, 73% combine calorie reduction with exercise, 13% use diet pills, 8% follow a restrictive weight loss plan, and only 8% participate in a structured weight loss program (while 7% also use an online diet plan). Taken together, the total estimated market in the U.S. for dietary products is \$65 billion and growing. Consumers are currently eating more low-cost fast food and comfort food. With a deep and prolonged recession, more people may join or continue to stay on weight loss programs. By contrast, the total market for organic products is only estimated to be less than \$15 billion. Thus, the greater popularity of dietary products will lead to greater sales at your stall if it sells diet snacks rather than organic foods.

Organic snacks are better for the environment than diet snacks

Diet snacks are more appealing to women shoppers


Organic snacks are better for the local economy

Diet snacks have more attractive packaging than organic snacks


Directions: Refer to the information on the left and answer the question below.

Now you can choose to get more detailed information by clicking on the choices on the left. You must choose at least one source of information before making a final decision.

When you are ready to make your final decision, click on the type of snack stand you would like to open.



Organic snacks



Diet snacks

organic snacks diet snacks




Figure 5: Information Search Decision Making Paradigm (“Snack Stand” Task)

Your uncle is an elderly driver who wants a vehicle that is safe and economical to run. He knows an importer of foreign cars, and he wants to buy one of a two Italian models, either a 2006 Ruggini or a 2006 Collasare.

Your uncle calls you up and says he is finding it hard to choose, and he wants your initial impression of which car is better, even if you are not sure.



What car would you choose for your uncle?

- Ruggini Collasare

Submit



Figure 6: Information Search Decision Making Paradigm (“Car Comparison” Task)

On the next screen, you will have the opportunity to review information from a website that has information about Italian cars, but you will have to pay \$1.00 for each piece of information. You will have 1 minute to read the information and make your final choice.



Continue



Figure 7: Information Search Decision Making Paradigm (“Car Comparison” Task)

Test 2 of 20 Test Time Elapsed 0:02:07

Comment from Ruggini owner Guilia (age 60) of Bologna	
Comment from Collasare owner Chad (age 64) of Pittsburgh	
Comment from Collasare owner Gianfranco (age 70) of Naples	
Comment from Ruggini owner Corey (age 67) of Dallas	
Excerpt about Rugginis from AARP newsletter	
Excerpt about Collasares from AARP newsletter	
Statement about Collasares from Italian Elderly Association	
Statement about Rugginis from Italian Elderly Association	
Comment from <i>Senior Journal</i> editor who drives Collasares	
Comment from <i>Senior Journal</i> editor who drives Rugginis	
Comment from <i>Today's Senior</i> editor who rents Rugginis in Italy	
Comment from <i>Today's Senior</i> editor who rents Collasares in the U.S.	

Money Left: \$ 12

On the left is a website that has information about Italian cars. The thumbs up and thumbs down symbols indicate whether the information is positive or negative.

Click on the choices on the left to access more information. You can make your decision between the two cars at any point, but you must access at least one source of information. Your task is to make the best decision you can, while paying as little money as possible. A total of how much time and money you have will be shown above and below.

Time Left: 00:00:45

Ruggini	Collasare







What car would you choose for your uncle?

Ruggini Collasare

Submit

Figure 8: Information Search Decision Making Paradigm (“Car Comparison” Task)

Test 2 of 20 **Test Time Elapsed** 0:02:48

<p>Comment from Ruggini owner Guilia (age 60) of Bologna The Ruggini is comfortable and easy to handle.</p>	<div style="text-align: right; background-color: #e0e0e0; padding: 5px; margin-bottom: 10px;">Money Left: \$ 0</div> <p>On the left is a website that has information about Italian cars. The thumbs up and thumbs down symbols indicate whether the information is positive or negative.</p> <p>Click on the choices on the left to access more information. You can make your decision between the two cars at any point, but you must access at least one source of information. Your task is to make the best decision you can, while paying as little money as possible. A total of how much time and money you have will be shown above and below.</p> <div style="text-align: center; background-color: #333; color: white; padding: 5px; margin-bottom: 10px;">Time Left: 00:00:03</div> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="text-align: center; padding: 10px;"></td> <td style="text-align: center; padding: 10px;"></td> </tr> <tr> <td style="text-align: center; padding: 5px;">Ruggini</td> <td style="text-align: center; padding: 5px;">Collasare</td> </tr> </table> <p style="text-align: center; margin-top: 10px;">What car would you choose for your uncle?</p> <p style="text-align: center;"> <input checked="" type="radio"/> Ruggini <input type="radio"/> Collasare </p>			Ruggini	Collasare
					
Ruggini		Collasare			
<p>Comment from Collasare owner Chad (age 64) of Pittsburgh The Collasare does not handle well in rain and snow.</p>					
<p>Comment from Collasare owner Gianfranco (age 70) of Naples I sometimes feel unsafe driving a Collasare through the winding streets of Naples.</p>					
<p>Comment from Ruggini owner Corey (age 67) of Dallas I most enjoy driving the Ruggini when I have to take long road trips.</p>					
<p>Excerpt about Rugginis from AARP newsletter The Ruggini doesn't have much interior space and some replacement parts are hard to find.</p>					
<p>Excerpt about Collasares from AARP newsletter The Collasare has good transmission and climate control.</p>					
<p>Statement about Collasares from Italian Elderly Association The Collasare lacks some desirable features and requires frequent maintenance.</p>					
<p>Statement about Rugginis from Italian Elderly Association The Ruggini has reliable power steering and provides a nice driving experience.</p>					
<p>Comment from Senior Journal editor who drives Collasares The Collasare is good for long road trips and driving in the mountains.</p>					
<p>Comment from Senior Journal editor who drives Rugginis The Ruggini sometimes stalls when climbing steep hills.</p>					
<p>Comment from Today's Senior editor who rents Rugginis in Italy Rugginis are quite popular among senior citizens who visit Italy.</p>					
<p>Comment from Today's Senior editor who rents Collasares in the U.S. Collasares are not very popular among senior citizens who visit the U.S.</p>					

Submit




Figure 9: Information Search Decision Making Paradigm (“Car Comparison” Task)

Evaluation/Weighting of Evidence/Questions

These tasks involve asking participants to select which pieces of evidence they would consider, or questions they would ask in order to evaluate a given hypothesis, from a balanced pool of confirming and disconfirming response options. One task was adapted from a study conducted by Snyder and Swann (1978) in which participants were provided in advance with a hypothesis regarding an interviewee's personality (introverted or extroverted) and were then allowed to determine what questions to ask those interviewees to evaluate the hypothesis. Participants tested their hypotheses by preferentially selecting, by means of the questions they chose to ask, behavioral evidence the presence of which would confirm their hypothesis. That is, participants assigned to evaluate whether an interviewee was introverted, asked the kinds of questions that would normally be asked of introverts (as determined in pretesting), and individuals assigned to evaluate whether an interviewee was extroverted, asked the kinds of questions that would normally be asked of extroverts (again, as determined in pretesting).

Similar to the Snyder and Swann (1978) study, individuals are placed in the role of a human resources employee assigned to make a closer assessment of a promising job applicant, where the "promise" is based on a high score on a scale from a standardized personality test that measures of attributes critical for successfully performing the job in question. Participants are asked to assemble a pool of interview/reference check items from a pool of items assembled by an expert external consultant. The items are all essentially personality items adapted for use in an interview or reference check. Half are keyed positively (confirmatory) and half are keyed negatively (disconfirmatory). A small number of additional response options measure unrelated traits. In a second task, participants are placed in the role of an intelligence analyst and asked to review information about a fictitious scenario adapted from published case studies in intelligence analysis (Beebe & Pherson, 2011) and select additional pieces of information that should be considered before deciding on a course of action. Half of the information options are considered to be confirmatory in the sense that they support the accepted hypothesis presented in the scenario and the remaining information options are considered disconfirmatory.

Directions: You are in the role of intelligence analyst. A junior analyst has gathered information, which is presented in the table below. Your job is to review this information and decide what additional information you would need before deciding on a course of action. Read the information and answer the question.

Event	Hamid Elahi is a politician from a small Middle Eastern country who has been living in exile in the United States. Recently, Elahi announced that he plans to return to his home country, an unofficial military dictatorship, in order to run for president on a pro-democracy ticket. Popular support for democracy in Elahi's home country is growing rapidly, but Elahi has a number of enemies there, including Imran Shah, the leader of the current government and an outspoken critic of Elahi's political beliefs. Many people fear Shah will attempt to have Elahi assassinated on his return home.
Key Question	Will Shah attempt to have Elahi assassinated once Elahi returns to his country?
Accepted Hypothesis	Shah is an aggressive leader who is intensely critical of the democratic movement within his country, and who has the means and motivation to resort to violence in order to achieve his goals. He WILL attempt to have Elahi killed once Elahi returns home to campaign for the presidency.

Which **FOUR** of these issues are the most important for you to investigate in attempting to answer the question?

- Shah's top political advisor has urged him to offer Elahi a power-sharing deal.
- Shah fears that growing support for democracy will make it impossible for him to maintain power even if his party wins the upcoming election.
- Corruption charges against senior members of Elahi's party could be used to discredit him.
- Shah has often stated publicly that he believes violence is a legitimate means to an end.
- Shah is concerned about losing the support of more moderate members of his party to Elahi.
- Shah has close ties to a group of religious extremists who have been known to carry out assassinations in the past.
- Shah has recently added a number of people to his personal security force.
- Shah is rumored to have become less convinced of the effectiveness of violence within the past several months.

Submit



Figure 10: Evaluation/Weighting of Evidence Paradigm (“Intelligence Analyst” Task)

You work in the HR department of a large corporation. In that role, you often interview and conduct reference checks for job candidates.

You have been asked to do an assessment for a candidate applying for a job in which **Adaptability** has been found to be critical for success. The candidate scored high on **Adaptability** in a preliminary personality test.

Prior to your assessment, you will need to select questions from a larger pool of questions that a consultant prepared for your HR department for general use in interviews and reference checks.

Definition of Adaptability

Open to new ways of completing tasks and projects; works well with different types of people by adapting his/her approach to fit the person; adjusts easily to changes at work; dislikes routine.

Below are questions provided by the consultant. Choose the FIVE questions from this pool that you believe will provide the best information about whether this candidate has Adaptability.

- Do you typically prefer variety to routine at work?
- Given a choice, would you prefer working with people who are similar to you?
- How annoyed do you get when changes are made to your work environment?
- Would past employers describe you as compassionate?
- How easily do you adapt to new situations?
- Do you generally prefer to stick with established ways of getting work done?
- How annoyed do you get when you are asked to switch from one assignment to another?
- Would past employers say you are generally open to change?
- Do you consider yourself more of a team player than most?
- How easy do you find it to work with people who are very different from you?
- Would past employers say you are usually among the last to be persuaded to try new systems and equipment?
- Are you usually among the first to try out new methods for getting work done?

Submit



Figure 11: Evaluation/Weighting of Questions Paradigm (“HR Department” Task)

2.4.5 Fundamental Attribution Error

Attitude Attribution Paradigm

In the original demonstration of the attitude attribution paradigm (e.g., Jones & Harris, 1967), participants read either a pro- or anti-Castro essay (Studies 1 & 2) or a pro- or anti-segregation essay (Study 3). Participants were told either that the essay writer had freely chosen to write the essay (choice condition) or that they were assigned to write the essay as the first part of an opening statement of a debate for a class assignment (no-choice condition). They then predicted the essay writer's true attitude on several Likert-type scales. Results indicated that participants predicted that essay writers who wrote pro-Castro/segregation essays had more positive attitudes toward Castro or segregation than those who wrote anti-Castro/segregation essays. Importantly, this was true in both the choice and no-choice conditions, indicating that participants were discounting the power of the situation (the fact that essay topic was assigned).

We developed tasks representing the *Attitude Attribution Paradigm* that involve watching a video of a person making a speech advocating a position and then making judgments about the extent to which that person believes what he or she was saying. In the ABC-1, the Attitude Attribution Paradigm is represented by three tasks that are "clones" of one another. In one task, the speech-maker advocates against keeping dolphins in captivity. In a second task, the speech-maker advocates against farm subsidies. In the third task, the speech-maker advocates for the use of violence to affect social change. In general, people conclude that the person believed what he or she was saying and would be likely to act in a manner consistent with that dispositional attribution, even if the instructions indicate that the position advocated in the speech was assigned randomly, or the person has found the position he or she is advocating to be ineffective.

Below is a video showing a man making a speech for a political science class. He was randomly assigned to give a speech arguing for the use of violence to affect political change.

You will NOT be able to view the video again after you have clicked on continue.



Continue



Figure 12: Attitude Attribution Paradigm (“Revolutionary” Task)

Directions: Answer the questions below.



How do you think this person feels about the use of violence to affect political change?

Definitely against political violence	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Definitely for political violence
Thinks political violence is a bad thing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Thinks political violence is a good thing
Definitely does not like political violence	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Definitely likes political violence
Thinks more political violence would harm society	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Thinks more political violence would help society

Submit



Figure 13: Attitude Attribution Paradigm (“Revolutionary” Task)

Directions: Answer the questions below.



Imagine that this person led a non-violent demonstration to protest a new law, but the demonstration ended soon after police officers arrested all of the protesters, and it had little impact. If he had led a violent protest, the demonstration would have attracted publicity and increased pressure on the government to repeal the new law. With that understanding, how do you think this person would feel about the use of violence to affect political change after this event?

Would be against political violence	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Would still be for political violence
Would think political violence is a bad thing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Would think political violence is a good thing
Would not like political violence	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Would like political violence
Would think more political violence would harm society	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Would think more political violence would help society

Submit



Figure 14: Attitude Attribution Paradigm (“Revolutionary” Task)

Directions: Answer the questions below.



How likely do you think this person would be to participate in the following activities?

	Not likely at all	Not likely	Somewhat likely	Likely	Very likely
Join a hunger strike with fellow political activists	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Join an armed militia to fight for his cause	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Write an op-ed for a local newspaper in which he advocates the use of non-violent tactics to affect political change	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Create a blog that advocates the use of violence to affect political change	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Write and distribute a petition to garner support for his cause	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Submit

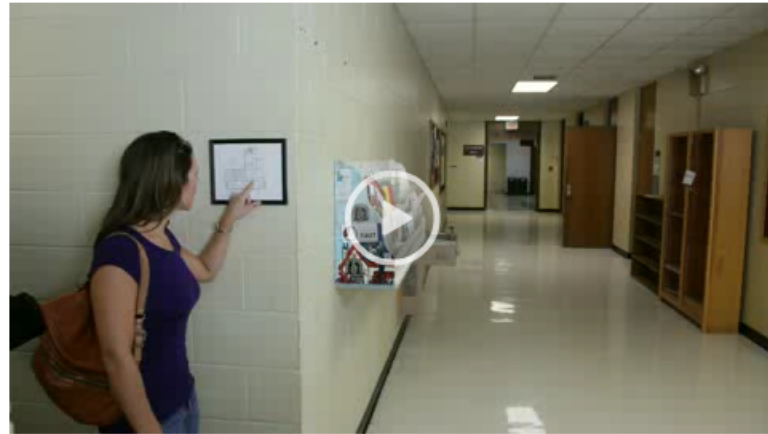


Figure 15: Attitude Attribution Paradigm (“Revolutionary” Task)

Good Samaritan Paradigm

We developed several tasks modeled off of the “Good Samaritan” study conducted by Darley and Batson (1973). In that study, people who were in a hurry to make an appointment were more likely to pass over a person slumped by the side of the road without helping than people who were not in a hurry. The fact that some were going to listen to a lecture about the parable of the Good Samaritan did not influence the results. This study speaks to the power of the situation influencing altruism.

The ABC-1 tasks representing the *Good Samaritan Paradigm* involve watching a video in which an individual has an opportunity to be a “Good Samaritan” or not to help, and chooses not to help. We developed a prototype and two Good Samaritan clones, which are differentiated on the basis of the color of the shirt worn by the non-helping individual played by a different actor in each scenario. In each case, that individual quickly walks past someone who is in obvious need of assistance. FAE is indicated by people's tendency to attribute a failure to help to personalities rather than possible situational causes.



Directions: Watch the video. You will NOT be able to view the video again after you have clicked on Continue.

Continue



Figure 16: Good Samaritan Paradigm (Good Samaritan “Woman in White” Task)



Regarding the woman in the white shirt, to what extent do you agree that this person has the following characteristics?

	Completely disagree	Disagree	Somewhat disagree	Somewhat agree	Agree	Completely agree
Helpful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Empathetic	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nice	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Generous	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Caring	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Friendly	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Neighborly	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Considerate	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Altruistic	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Self-centered	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Submit



Figure 17: Good Samaritan Paradigm (Good Samaritan “Woman in White” Task)



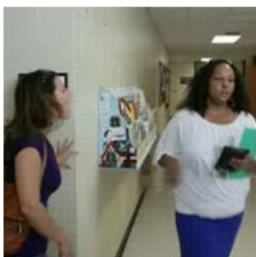
Imagine the following situations are occurring. How likely would this person be to help?

	Very unlikely to help	Unlikely to help	Somewhat unlikely to help	Somewhat likely to help	Likely to help	Very likely to help
She is fishing with friends and sees a person in the lake struggling to stay above water.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
A friend calls on the weekend and asks her to help her paint her house.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
She received a phone call asking her to donate money to the local children's hospital.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
While walking in the park, she sees a child fall and skin his knee.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
A person stops her on the sidewalk and asks her to sign a petition.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Her mother asks her to come over and mow her lawn.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Submit



Figure 18: Good Samaritan Paradigm (Good Samaritan “Woman in White” Task)



How likely are each of the following reasons for this person not helping the woman with directions?

	Very unlikely	Unlikely	Somewhat unlikely	Somewhat likely	Likely	Very likely
She is not a helpful person	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
She was distracted by her reading	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Submit



Figure 19: Good Samaritan Paradigm (Good Samaritan “Woman in White” Task)

Attributional Style Paradigm

We developed two tasks representing the *Attributional Style Paradigm* that were adapted from Riggio and Garcia (2009). In the first task, a text-based scenario is presented to test-takers. Test-takers then rate the importance of three dispositional and seven situational factors as causes of the events described in the scenario. FAE is measured as the extent to which test-takers rate dispositional explanations more highly than situational explanations of the events described in that scenario. In the second task, test-takers are presented with a series of common events that happen to a number of focal characters. Test-takers must make a series of ratings regarding the extent to which the most likely cause of each event is dispositional (FAE) versus situational (non-FAE). There are three primary types of questions for each focal character's scenario: questions indicating (1) the extent to which the main cause of the event is dispositional versus situational, (2) whether the same type of attribution will apply if the same event occurs in the future, and (3) the extent to which the dispositional versus situational explanation likely generalizes to other aspects of the focal character's life. The only type of question that is actually scored, however, is the "main cause" question for each scenario. The other two question-types are foils.



Drew had a good night's sleep and woke up feeling refreshed and ready for the day on Wednesday morning. He arrived to work on time, got busy right away, and completed several important tasks before his 10:00 AM meeting. His 10:00 AM meeting ended early, so he made it to his 11:30 AM meeting early. This meant that he had a few extra minutes to talk to his manager about some new ideas he had for improving company efficiency. During the 11:30 AM meeting, he and his coworkers formally presented the new ideas to his manager. The presentation went smoothly and the manager seemed positive about the new ideas. At 3:00 PM, when Drew was back at his desk, three clients called and extended their contracts with Drew's firm. At 4:00 PM, Drew was called to his manager's office and given a promotion.

Identify the most important causes of Drew's good day by rating each factor that contributed to the eventual outcome of Drew getting promoted.

	Extremely unimportant	Unimportant	Somewhat unimportant	Somewhat important	Important	Extremely important
Drew's personality	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The workplace environment	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Drew's coworkers	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Other staff	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Drew's skills	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The clients	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The Manager	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Drew's abilities	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The comfort of his bed	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Work technology	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Submit



Figure 20: Attributional Style Paradigm (“Drew’s Good Day” Task)

Main Character	Simon Winkler, a 56-year old playwright
Event	This year, Simon Winkler won an international drama competition for his latest play about soldiers in Afghanistan.

Directions: Answer the questions below.

What do you think is the *most likely cause* of this event? Choose one most likely (or main) cause below.

Simon is a talented writer.
 The topic of play is related to current events.

What was the main cause of Simon winning the international drama competition? Was it mainly due to him or other people or circumstances?

Totally due to other people or circumstances Totally due to Simon

In the future, when Simon wins a prize for one of his plays, will it be due to the same cause you specified above?

Will never again be due to the same cause Will always be due to the same cause

Is the cause something that just influences the winning of the international drama competition, or does it also influence other areas of Simon's life?

Influences just this particular situation Influences all situations




Figure 21: Attributional Style Paradigm (“What Causes Things?” Task)

Confession Paradigm

Another task created for the ABC-1 was modeled after a study conducted by Kassin and Sukel (1997), who were interested in examining the extent to which juries ignore coerced confessions that have been determined inadmissible by a judge. In the original demonstration, participants read a transcript of a confession to a crime. The transcript was manipulated such that some transcripts discussed a highly coerced confession (e.g. "Officer Heffling handcuffed me, took out his gun and started asking me questions about the murders..."), a confession that was not highly coerced (He was not handcuffed, verbally abused, or threatened), or no confession. Results revealed that the presence of the confession, whether it was seen as coerced or not, led to higher judgments of guilt versus a control condition. This provides evidence that the respondents were not fully taking situational pressure (i.e. coercion) into account when making their judgments.

In the ABC-1, the *Confession Paradigm* task involves listening to an audiotape of a conversation between two individuals who are roommates. Roommate A (Interrogator) is suggesting that her roommate made up an excuse not to go to a party to avoid having to go. Roommate B (Confessor) denies it, but after the Interrogator presses the point, eventually "confesses" that she did indeed make up the excuse that she was sick for the party. However, the confession is made in a tone of voice that suggests that the confession *may* not have been genuine. FAE is indicated by test-takers making attributions that the confessor was not affected by the situational pressure imposed by the persistent questioning of the Interrogator and instead has attributes consistent with feelings of guilt.

Directions: You will hear audio from two friends having a discussion. One woman is accused by her friend of lying about her reason for not attending a party. Although the woman originally denied being dishonest, she admits her guilt when accused by her friend. Listen to the audio clip. You will NOT be able to listen to the clip again after you have clicked continue.



[▶ Click to Play Audio](#)

Continue



Figure 22: Confession Paradigm (“Sick for Party” Task)

Directions: Answer the questions below.



Would you say the woman is guilty or not guilty of lying to her friend?

Guilty Not guilty

	Not at all confident	A little confident	Somewhat confident	Confident	Very confident
How confident are you that the woman is guilty?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How confident are you that the woman is not guilty?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

	Not at all	A little	Somewhat	A lot	A great deal
How much did the confession influence your decision?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Submit



Figure 23: Confession Paradigm (“Sick for Party” Task)

Directions: Answer the questions below.



	Not at all	A little	Somewhat	A lot	A great deal
How much did the pressure imposed by her friend influence the woman's confession?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How much did the woman's guilty conscience influence her confession?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Submit



Figure 24: Confession Paradigm (“Sick for Party” Task, continued)

Directions: Answer the questions below.



To what extent do you agree that the woman's personality has the following characteristics?

	Completely disagree	Disagree	Neither agree nor disagree	Agree	Completely agree
Nervous	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Afraid	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Calm	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Assured	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Submit



Figure 25: Confession Paradigm (“Sick for Party” Task, continued)

Quiz Role Paradigm

The “Quiz-Role” paradigm was first developed by Ross and colleagues (1977) and more recently studied by Gawronski (2003). In the paradigm, participants are invited to play in a “quiz game” and randomly assigned to the role of questioner or contestant. The questioner writes difficult (but not impossible) quiz questions and then asks contestants to answer them. Contestants are aware of condition assignment and the task the questioner was asked to complete. After the contestant answers the questions, participants rate themselves and their partners on “General knowledge compared to the average [Name of participants’ school] student” on a 0 to 100 scale, ranging from *much worse than average* (0) to *much better than average* (100). The typical finding is that both the questioner and the contestant rate the questioner as having more general knowledge than the contestant. In a yoked control condition, the questioner asks the same questions that were written by questioners in the experimental condition. In this case, the observed difference in general knowledge ratings is typically diminished for contestants but remains the same for questioners.

We developed tasks for the ABC-1 representing the *Quiz Role Paradigm* that involve one person posing questions to another person. The general finding is that questioners who ask respondents questions tend to be rated as more knowledgeable than respondents, even when there is no basis for reaching that conclusion. In the ABC-1, the Quiz Role paradigm is represented in the Personnel Selection task and the two Trivia Quiz tasks: one prototype and one clone. In the Personnel Selection task, test-takers read a transcript in which two management trainees are ostensibly learning to administer job interviews. As part of their training, one trainee selects questions to ask of another trainee, as if the other trainee were a job applicant. FAE is indicated by a tendency for test-takers to rate the knowledge, quickness, and aptitude of the trainee who is asking questions as being higher than the knowledge, quickness, and aptitude of the trainee who is answering the questions. In the Trivia Quiz tasks, test-takers are presented with a video clip in which a questioner asks trivia questions of another individual, designated the “answerer.” FAE is indicated by a tendency to rate the questioner as having more knowledge, skill, and aptitude than the answerer.

Below is a video showing one person asking another person a series of trivia questions. The two people in the video were assigned to the roles; one person was assigned to be the questioner and the other person was assigned to answer the questions. The person assigned to the role of the questioner wrote the questions.



Continue



Figure 26: Quiz Role Paradigm (“Trivia” Task)

Directions: Answer the questions below.



Compared to the average college student,

	Much less intelligent	Less intelligent	Neither less nor more intelligent	More intelligent	Much more intelligent
how intelligent is the questioner?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
how intelligent is the answerer?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Who would you say is more intelligent?

Definitely the answerer	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Definitely the questioner
-------------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	---------------------------

Submit



Figure 27: Quiz Role Paradigm (“Trivia” Task)

Silent Interview Paradigm

The *Silent Interview Paradigm* is a classic laboratory paradigm in which participants are told that they are about to watch two silent videos of women being interviewed (Snyder & Frankel, 1976). They are told that one interview is about sex and the other interview is about politics, and are provided with this information either before or after they watch the video. Participants are also provided with example interview questions. After watching the video, participants are presented with six hypothetical situations and asked how “apprehensive” the interviewee would be in each, compared to the average person. Participants typically think that the interviewee asked about sex would be more apprehensive across the six situations than the interviewee asked about politics, and the interviewee asked about sex is rated higher in anxiety. Thus, participants make dispositional attributions to the woman asked about sex, although they are fully aware of the situation before viewing the behavior. In other words, participants tend to state that the woman in the anxiety invoking interview will also tend to be anxious in other situations, despite the fact that they know that she is talking about something that invokes anxiety. Thus, they are discounting the situation when predicting future behavior.

We developed several tasks representing the *Silent Interview Paradigm* in which test-takers are shown a video clip of a person being interviewed without sound. Before the interview, test-takers are told that the person is being interviewed about something anxiety-provoking. In the ABC-1, this involves waiting for a job interview, interacting with a physician, and being accused of lying at work. The nonverbal behaviors, however, are not clearly consistent with an anxiety-provoking situation. FAE is indicated by ratings that the individual depicted in the video clip has a nervous disposition, as opposed to situational explanations.

Below is a video showing a woman being interviewed. She has been accused of lying to her boss. Your task is to watch the video and then answer some questions about the woman. The sound for the video is not available so you must look at her facial expressions and body language to form your opinions. You may pause and replay the video as needed.

You will NOT be able to view the video again after you have clicked on Next.



Continue



Figure 28: Silent Interview Paradigm (“Lying” Task)

Directions: Answer the questions below.



	Much less than average	Less than average	Neither less nor more than average	More than average	Much more than average
On a day-to-day basis, how anxious is this person compared to the average person?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
On a day-to-day basis, how confident is this person compared to the average person?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
On a day-to-day basis, how nervous is this person compared to the average person?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
On a day-to-day basis, how happy is this person compared to the average person?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
On a day-to-day basis, how proud is this person compared to the average person?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
On a day-to-day basis, how tense is this person compared to the average person?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Submit



Figure 29: Silent Interview Paradigm (“Lying” Task, continued)



To what extent do you agree that this person has the following characteristics?

	Completely Disagree	Disagree	Neither Agree nor Disagree	Agree	Completely Agree
Nervous	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Afraid	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Calm	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Assured	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Peaceful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Tense	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Uneasy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Distressed	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Jumpy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Shaky	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Submit



Figure 30: Silent Interview Paradigm (“Lying” Task, continued)

2.4.6 Bias Blind Spot

The ABC items measure BBS with respect to a range of social and cognitive biases to help ensure that it is measuring BBS as accurately as possible and not finding effects that may be unique to one particular bias. This measurement strategy essentially follows that used by Pronin, Lin, and Ross (2002), and adopted by other prominent researchers (e.g., West, Meserve, & Stanovich, 2012).

In pilot test research, we found that there was essentially no order effect with respect to presentation of "Average American" versus "You" questions. So, presentation of the "Average American" items, rather than "You" items, first does not systematically influence ratings or difference-score values. More importantly, we found that the bias blind spot effect, the tendency for individuals to judge other people as being more susceptible to cognitive biases than themselves, occurs even when individuals are asked to evaluate their own tendency to use “adaptive” mental heuristics in judgment and decision making scenarios relative to the “average American”. This finding may suggest the more general phenomenon of a “heuristic blind spot”—the tendency for individuals to perceive others as being more susceptible to heuristic thinking than themselves. In addition, this so-called “heuristic blind spot” effect extends to descriptions of psychological effects that are framed in more balanced or neutral terms, as well as to descriptions of fictitious, albeit seemingly plausible, psychological phenomena.

The BBS BE scale consists of eight difference score-based items on each test form with higher scores indicating less susceptibility to BBS. Test-takers first rate the susceptibility of the “Average American” to a given heuristic, bias, or mode of reasoning (hereafter collectively referred to as "Bias") on a 5-point Likert-type scale. After making these ratings, test-takers are asked to rate their own susceptibility to each effect. BBS is indicated by higher ratings of the susceptibility of the “Average American” to a given Bias than for one’s own ratings of susceptibility to each effect.

2.5 Phase 1 “Pre-Pilot” Studies

Because the constructs targeted for measurement in the ABC-1 were not well understood from an individual differences perspective, we conducted a considerable amount of pilot test research prior to the Phase 1 Field Test (referred to throughout this report as “Pre-Pilot” research/studies). Due the fast-paced nature of the project, it was necessary to conduct some of the research either in parallel or in a cascading fashion, although we conducted the research iteratively and sequentially to the extent possible so that we could build on knowledge as it was acquired.

Our pre-pilot research was intended to address a number of questions, an illustrative sample of which is as follows:

1. Conduct preliminary item/facet and scale analyses (i.e., reliability⁶, item discrimination⁷, variability, group-level distributions)

⁶ The degree to which test scores are free of random measurement error in a given group.

⁷ The extent to which an item/task discriminates between test-takers who are relatively higher and lower on an attribute targeted for measurement.

2. Vary instruction sets to use for different prototypes to identify which is most clear and efficient
3. Investigate time required to respond to items and scales
4. Conduct preliminary evaluation of scale overlap and redundancy
5. Determine the minimum number of items required to achieve adequate reliability and validity⁸
6. Investigate how use of payoff matrices and/or other motivational manipulations affect item responding
7. Determine the moderators/covariates⁹ that need to be accounted for
8. Evaluate convergent and discriminant validity¹⁰ of ABC scales
9. Investigate the effects of specifying different reference populations on bias blind spot measurement
10. Investigate whether hypothesized cognitive biases for prototypes are in fact being elicited

Because the pre-pilot test research encompassed an enormous amount of work, we focus only on key questions and results that are summarized in Tables 3, 4, 5, and 6. These tables correspond to CB, FAE, BBS, and RD, respectively. The tables are intended to stand alone, and we do not discuss them beyond the content in the tables themselves. We discuss the Field Test and Pretest Sensitization studies in detail subsequent to the summary of the pre-pilot test research.

⁸ The extent to which accumulated evidence and theory support a specific interpretation of test scores for a given use of that test, such as the ability to detect change on a construct from pretest to posttest, given an intervention designed to increase people's standing on that construct.

⁹ A moderator is a variable that affects the direction or strength of the relationship between two other variables (e.g., changes the correlation). A covariate is a variable that is not of primary interest, but may correlate with an outcome variable, or otherwise affect the relationship between two variables under study, and is therefore held constant to neutralize any confounding effect it might have.

¹⁰ Convergent evidence of validity is evidence based on the relationship between test scores (or sub scores) and other measures of the same or related construct. Discriminant evidence of validity is evidence indicating whether two tests interpreted as measures of different constructs are sufficiently independent (uncorrelated) that they do, in fact, measure two distinct constructs.

Table 3: Summary of CB Pre-Pilot Research Studies

CB Study Iteration	Brief Study Description and Key Questions	Key Results
Round 1	<p>Cognitive Laboratory studies with 22 ETS employees</p> <ul style="list-style-type: none"> • Do examinees understand the task requirements? • Are there any particular task elements or features that facilitate or hinder task performance? • What thinking strategies do examinees use to perform BE tasks? <p>Pre-pilot testing with University of Cincinnati ($n = 44$) and Washington University in St. Louis ($n = 30$) students</p> <ul style="list-style-type: none"> • Investigated psychometric properties of task prototypes and clones 	<p>Usability issues identified pertaining to:</p> <ul style="list-style-type: none"> • Instructions and scoring/feedback for Defocused Images and Face/Flag Sorting variants of the Wisconsin Card Sorting tasks • High text reading and cognitive load requirements identified for “Justify a Claim” task <p>Additional Findings:</p> <ul style="list-style-type: none"> • Some evidence for overall CB in “Shopping Malls,” “Defocused Images,” and “Hiring Manager” tasks • Individual differences observed in all tasks • Balanced information condition in the “Car Selection” information search task prototype did not elicit balanced preferences • Not enough pieces of disconfirming evidence in “Justify a Claim” task • Response times suggest test takers are generally careful and deliberative

CB Study Iteration	Brief Study Description and Key Questions	Key Results
Round 2	<p>Pre-pilot testing with University of Cincinnati ($n = 84$) and Washington University in St. Louis ($n = 38$) students</p> <ul style="list-style-type: none"> • Sought to replicate Round 1 findings • Compared “Balanced” vs. “Unbalanced” conditions in “Car Comparison” Information Search Task • Examined item performance characteristics for “Shopping Mall,” “Defocused Images,” “Hiring Manager,” “Intel Analyst,” “Face/Country Sorting,” and “Justify a Claim” task prototypes 	<ul style="list-style-type: none"> • Participants typically understand instructions for item prototypes (both revised and new prototypes) and give high usability ratings for the tasks • Replicated Round 1 findings of overall CB for “Shopping Mall,” “Hiring Manager,” and “Defocused Images” tasks • No overall evidence of CB in “Car Comparison,” “Intel Analyst,” “Face/Country Sorting,” and “Justify a Claim” tasks
Round 3	<p>Pre-pilot testing with University of Cincinnati ($n = 213$) students, Washington University in St. Louis ($n = 60$) students, and AMT workers ($n = 153$)</p> <ul style="list-style-type: none"> • Sought to replicate and extend findings from earlier rounds • Manipulated wording (causal vs. non-causal) and format (map vs. list) in Wason Selection tasks • Compared multiple versions and investigated simulated costs of information search (points vs. money) in product decision information search tasks • Compared influence of expert vs. peer on CB in Defocused Images task • Compared item performance characteristics for 6-item forced choice and Likert scale versions of “Hiring Manager” task • Examined item performance characteristics for “Snack Stand” and “HR Department” tasks 	<ul style="list-style-type: none"> • “Non-causal” wording elicited CB for Wason Selection items • “Unbalanced” simplified versions of product decision making information search tasks demonstrated evidence of CB elicitation • Original “Expert” opinion, but not the “Peer” opinion, version of Defocused Images task continued to demonstrate CB elicitation • Evidence of CB in “Snack Stand” task • Intel Analyst items demonstrated CB elicitation, but not FAE • HR Department items demonstrated CB elicitation • Forced-choice and Likert versions of Hiring Manager task failed to show CB

CB Study Iteration	Brief Study Description and Key Questions	Key Results
Round 4	<p>Small and large-scale pilot studies with ETS essay raters ($n = 1539$) and AMT workers ($n = 2965$)</p> <ul style="list-style-type: none"> • Sought to replicate findings from previous rounds • Compared extended 21-trial and shortened 14-trial versions of Defocused Images task • Examined item performance characteristics for revised and new task prototypes and clones, including video-based SJTs • Investigated cross-task correlations and correlations with other individual-difference variables and background/demographic variables 	<ul style="list-style-type: none"> • Replicated Round 3 findings of CB elicitation with Wason Selection, HR Department, and Intel Analyst tasks • Both 21-trial and 14-trial versions of Defocused Images task demonstrated CB elicitation • Decision Making Information Search tasks showed overall evidence of CB elicitation • Hiring Manager failed to show evidence of CB • Limited evidence of CB in “Circumplex” video-based SJT and “Dating Website” variants of “Urn” paradigm • Low correlations between CB paradigms • No practically-significant correlations observed between CB paradigms and BFI personality, cognitive ability, and demographic variables

Table 4: Summary of FAE Pre-Pilot Research Studies

FAE Study Iteration	Brief Study Description and Key Questions	Key Results

FAE Study Iteration	Brief Study Description and Key Questions	Key Results
Round 1	<p>Cognitive Laboratory studies with 22 ETS employees</p> <ul style="list-style-type: none"> • Do examinees understand the task requirements? • Are there any particular task elements or features that facilitate or hinder task performance? • What thinking strategies do examinees use to perform BE tasks? <p>Pre-pilot testing with University of Cincinnati ($n = 44$) and Washington University in St. Louis ($n = 30$) students</p> <ul style="list-style-type: none"> • Investigated psychometric properties of task prototypes and clones 	<ul style="list-style-type: none"> • FAE elicited in Quiz Role, Silent Interview and Attributional Style task prototypes and clones • Limited evidence for FAE elicitation in “Speed Decision Task” • Limited evidence for cross-task convergent validity
Round 2	<p>Pre-pilot testing with University of Cincinnati ($n = 84$) and Washington University in St. Louis ($n = 38$) students</p> <ul style="list-style-type: none"> • Sought to replicate Round 1 findings • Examined item performance characteristics for Attitude Attribution task prototype 	<ul style="list-style-type: none"> • Replicated Round 1 findings with Quiz Role task prototypes (“Personnel Selection” and “Trivia”) • FAE elicited in Attitude Attribution task prototype (“Revolutionary” task)

FAE Study Iteration	Brief Study Description and Key Questions	Key Results
Round 3	<p>Pre-pilot testing with University of Cincinnati ($n = 213$) students, Washington University in St. Louis ($n = 60$) students, and AMT workers ($n = 153$)</p> <ul style="list-style-type: none"> • Sought to replicate Round 2 findings • Investigated framing of cover stories in Quiz Role and Attitude Attribution paradigms (Trivia and Personnel Selection) • Examined item performance characteristics for Attributional Style paradigm task prototypes (“Ron, Alice, and Friends” and “What Causes Things”) 	<ul style="list-style-type: none"> • Gender of respondent in Quiz Role tasks did not have a significant effect on FAE elicitation • Greater FAE elicitation observed in Personnel Selection task when questioner choose own questions, but questioner writing/not-writing own questions manipulation did not have a significant effect in the Trivia task • Moderately greater FAE elicitation Attitude Attribution paradigm (Revolutionary Task) in random assignment condition • Only 1 of 5 scenarios in “Ron, Alice, and Friends” Attributional Style task elicited FAE; however, FAE reliably elicited for 14 out of 16 scenarios in “What Causes Things” task
Round 4	<p>Small and large-scale pilot studies with ETS essay raters ($n = 1539$) and AMT workers ($n = 2965$)</p> <ul style="list-style-type: none"> • Sought to replicate findings from previous rounds • Examined item performance characteristics for Good Samaritan, Confession, and Anchoring Vignettes task prototypes • Examined item performance characteristics for clones of Quiz Role, Attitude Attribution, Silent Interview, and Good Samaritan tasks • Investigated cross-task correlations and correlations with other individual-difference variables and background/demographic variables 	<ul style="list-style-type: none"> • Replicated findings from previous rounds for Quiz Role, Attitude Attribution, Attributional Style (“What Causes Things”) task paradigms • Contrary to Round 3 results, 4 out of 5 scenarios in “Ron, Alice, and Friends” task elicited FAE • Limited evidence of FAE elicitation in Speed Decision, Intel Analyst, and AV vignettes tasks • Low correlations between FAE paradigms • No practically-significant correlations observed between FAE paradigms and BFI personality, cognitive ability, and demographic variables

Table 5: Summary of BBS Pre-Pilot Research Studies

BBS Study Iteration	Brief Study Description and Key Questions	Key Results
Round 1	<ul style="list-style-type: none"> • AMT, $n = 212$ • 16 biases/ heuristics • Is there an order effect¹¹? • Is there an item framing effect¹²? 	<ul style="list-style-type: none"> • No order effect for “you” rating vs. “average American” rating presented first • BBS was elicited whether item was framed as positive heuristic or negative bias
Round 2	<ul style="list-style-type: none"> • AMT, $n = 314$ • Added 17 effects • Extended investigation of item framing • Added two “balanced” conditions: (+, -) and (-, +) • "Balanced" conditions: positive and negative aspects of an effect 	<ul style="list-style-type: none"> • Balanced conditions fell in between positive heuristic and negative bias conditions in terms of BBS elicitation • Near-zero difference between the two balanced conditions

¹¹ A question order effect occurs when responses to a prior question on a test affect responses to a subsequent one.

¹² An item framing effect occurs when item responses are affected by content that has preceded that item. For example, suppose that a survey is administered to two groups of test-takers selected from the same sample. Suppose, further, that one group of test-takers is administered a version of the survey in which all of the questions are presented in the first person (“I”) and another group of test-takers is administered a version of the survey in which all of the questions are presented in the second person (“You”). If the two groups receive different scores on the survey, and there are no other explanations of the difference, this would be an example of an item framing effect.

BBS Study Iteration	Brief Study Description and Key Questions	Key Results
Round 3	<ul style="list-style-type: none"> • Increased sample size to over 600 • Investigated test-retest reliability • Question: How many items would it require to create composite with good psychometric properties? 	<ul style="list-style-type: none"> • Balanced (+, -) condition elicited BBS about as well as the Negative Bias condition • For "Best 13" composite statistics: <ul style="list-style-type: none"> • Bias was elicited • Composite internally consistent : $\alpha^{13} \approx 0.80$'s • Composite temporally stable : $rxx' \approx 0.70$
Round 4	<ul style="list-style-type: none"> • New sample from different test-taker population (e.g., ETS essay raters) • Created 8-item unit-weighted composites for each of the three ABC forms • Investigated correlations with other individual-difference variables and background/demographic variables Investigated the psychometric properties of 8-item composites proposed for use as BBS scales 	<ul style="list-style-type: none"> • Higher cognitive workload is associated with less BBS • Modest, significant negative correlations with crystallized intelligence (Gc)¹⁴ • Inconsistent correlations with personality measures • No consistent correlations with background/ demographic variables • Each 8-item BBS composite elicited the bias (only 16-20% of test-takers failed to show BBS) • 8-item composites largely unidimensional • Alpha coefficients range from 0.71 to 0.76 ($n = 564$ to 577) • Composites show moderately good test-retest reliabilities for 8-item composites: $rxx' = 0.66$ to 0.73 ($n = 84-85$)

¹³ Coefficient alpha is one of a family of reliability metrics design to evaluate internal consistency of a composite of psychological variables. It is provides a rough estimate of the extent to which the variables comprising the composite are interrelated.

¹⁴ A facet of general cognitive ability that reflects the influences of formal learning and acculturation, including education.

Table 6: Summary of RD Pre-Pilot Research Studies

RD Study Iteration	Brief Study Description and Key Questions	Key Results
Study 1	<ul style="list-style-type: none"> • AMT, $n = 274$ • Administered 35 items covering the ABC-1 content domain • Test-takers were given a carefully developed one-page description of each bias to read prior to taking the RD test • Investigated psychometric properties of the items by computing basic descriptive statistics, conducting internal consistency reliability analyses, and conducting a principal components analysis¹⁵ 	<ul style="list-style-type: none"> • Retained 20 items covering knowledge of CB (5 items), FAE (3 items), and no bias¹⁶ (2 items) • 15 items were dropped due to excessive difficulty, lack of correlation with other items, tendency to decrease alpha, and low loading on the 1st unrotated principal component¹⁷ • Mean score for the 20-item scale was 13.87 ($SD = 4.45$) • Alpha coefficient was .86 • RD scale correlated $r = .41$ with Gc • No large correlations with demographic variables

¹⁵ Principal components analysis is a data reduction technique the purpose of which is to summarize the correlations between a set of variables using a smaller number of components. The goal is to account for as much of the variance in the correlation matrix as possible with a relatively small number of components. In factor analysis, a set of data reduction techniques closely related to principal components analysis, the dimensions are typically rotated such that the factors are more interpretable. In principal components analysis, however, the components are not to be rotated because the reduction in dimensionality is not designed to discover interpretable dimensions, but simply to summarize data.

¹⁶ A “no bias” item refers to an item the correct answer to which is either none of the targeted Study 1 biases or no bias at all. Per the Sirius BAA, the RD test was to evaluate not only ability to recognize targeted biases but also to discriminate among them. Discrimination also includes the ability to discriminate between instances of bias versus no bias. As such, “no bias” items were included in the RD test to evaluate test-takers’ ability to discriminate not only between biases but between items that described examples of targeted biases and items that did not describe examples of targeted biases.

¹⁷ In principal components analysis, if the first component accounts for a large amount of the information in the correlation matrix that the principal components analysis is intended to summarize, the magnitude of the relationships, or “loadings,” of each of the correlated variables on the first component (which is unrotated, as described in the previous footnote), are indicators of unidimensionality. That is, if all of the variables have high loadings on the first unrotated principal component, and the first unrotated principal component accounts for most of the information in the variable correlation matrix, that is evidence that one dimension can summarize the relationships between all of the variables that are correlated. Hence, unidimensionality can be inferred.

RD Study Iteration	Brief Study Description and Key Questions	Key Results
Study 2	<ul style="list-style-type: none"> • AMT, $n = 466$ • Administered 20 items retained from Study 1 • Test-takers were given a carefully developed one-page description of each bias to read prior to taking the RD test • Investigated the psychometric properties of the RD test to determine whether any revisions were indicated for the RD scale • Investigated correlations between the RD test and personality, cognitive ability, Bias Blind Spot (BBS) susceptibility, and background variable measures • Investigated whether taking the R&D Assessment increases scores on BBS bias elicitation by administering 5 BBS BE items to half of the AMT test-takers after the description of the Phase 1 biases and RD test, and the remaining test-takers completed the BBS elicitation items before the description of biases and the RD test 	<ul style="list-style-type: none"> • Mean scores were .77, .73., and .70 for items targeting BBS, FAE, and CB, respectively, and $M = .34$ for “None of the above” items • One “None of the above” item was dropped due to excessive difficulty, low corrected item-total correlation, and poor loading on the 1st unrotated principal component • Alpha coefficient for the 20-item scale was .83 • Exploratory factor analyses revealed that a three factor solution was not appropriate, and a two-factor solution did not yield interpretable factors • Items targeting CB, FAE, and BBS all loaded onto the same factors, and the two factors correlated $r = .65$, suggesting that the RD test may best be summarized by one dimension • PCA loading descriptive statistics were as follows: Mean = .48, median = .47, SD = .09, min = .35, max = .61 • RD scale correlated $r = .41$ with Gc • RD scale correlated significantly with the BFI-44 Openness to Experience ($r = .12, p < .05$) and Extraversion ($r = -.22, p < .05$) • RD had small significant correlations with cumulative GPA, father’s schooling, mother’s schooling, and – to a lesser extent – number of psychology courses taken ($r_s = .09$ to $.14$, all $p < .05$) • RD scale correlated .35 with the BBS composite. This correlation rises to .47 when correcting for unreliability in the BBS composite. Controlling for cognitive ability scores, the (uncorrected) correlation between RD and BBS dropped to .27. • Providing instruction about BBS and its manifestations did not inoculate test-takers from exhibiting the bias

2.6 ABC-1 Field Test

The purpose of the ABC-1 Field Test was to administer the entire set of tasks/items to a large and representative group of test-takers to evaluate their psychometric properties and validity. As such, we anticipated making some changes to the test forms, but also expected that, as a result of the extensive pre-pilot testing described above, the items would generally perform well. Two other critically important purposes of the field test were (1) to provide data necessary for creation of equivalent forms for use in the Phase 1 IV&V, and (2) to evaluate the sensitivity of the ABC-1 to a surrogate bias mitigation intervention in the form of the IARPA-produced instructional video (Intelligence Advanced Research Projects Activity, 2012).

2.6.1 Method

2.6.1.1 Participants

The ABC-1 Field Test was administered to a total of 2,017 test-takers, all of whom were recruited through Amazon Mechanical Turk. All test-takers were based in the U.S. and paid \$10 for completing the test. 416 (or 21%) out of the 2,017 Amazon Mechanical Turk participants either had incomplete responses on one or more of the ABC-1 scales, or they answered more than 25% of data check questions incorrectly and, therefore, their data were excluded from the analysis. Total screened sample consisted of 1,601 participants (Form 1 $n = 543$, Form 2 $n = 522$, Form 3 $n = 536$).

The sample of test-takers averaged 32 years of age ($SD = 12$ years); was approximately half male; and approximately 81% Caucasian, 7% black, 6% Asian, and 5% multi-racial. The sample was relatively high-achieving, with over 70% reporting cumulative college GPAs between 3.0 and 4.0, and 35% reporting GPAs over 3.5. The majority of the sample reported having taken one or two psychology courses, though the vast majority had not taken more than four.¹⁸

2.6.1.2 Study Design and Procedure

Table 7 lists the types of items and paradigms represented in each of the three forms administered online in the ABC-1 Field Trial. All BE items preceded RD items. The sequence of BE items varied such that the items representing each bias facet were presented in different sequences and combinations across forms. After completing the BE tests, participants read text descriptions of the Phase 1 biases prior to taking the RD test. Two attention check items were included in the BE test sequence, and two attention check items were included in the RD test sequence. Last, participants completed a demographics survey

¹⁸ Ideally, the participant population would be representative of the analyst population that will be receiving training from the Sirius video games. As a comparison, in the Sirius Phase 1 IV&V the participant demographics were as follows. *Students*: average age: 22; 45% male; 57% Caucasian, 14% Asian American, 9% Hispanic, 8% African-American, 12% other; most frequent major: Psychology. *Analysts*: average age: 38; 71% male; 62% Caucasian, 10% African-American, 12% Asian-American, 6% Hispanic, 10% other.

Table 7: Number of Items by Scale and Facets Represented in ABC-1 Field Trial Study Forms.

Scale	Facet	Form 1	Form 2	Form 3
Confirmation Bias	Wason Selection	2	2	2
	Information Search Decision Making	3	3	3
	Defocused Images	1	1	1
	Evaluation/Weighting of Evidence	3	3	3
	Evaluation/Weighting of Questions	4	4	4
Fundamental Attribution Error	Attitude Attribution	11	11	11
	Good Samaritan	18	18	18
	Quiz Role	13	13	13
	Attributional Style	22	22	22
	Confession	7	7	7
	Silent Interview	13	13	13
Bias Blind Spot		8	8	8
Recognition and Discrimination	Confirmation Bias	5	4	4
	Fundamental Attribution error	5	6	5
	Bias Blind Spot	2	3	3
	None of the above	1	0	1
Attention Check Items	BE Item Type	2	2	2
	RD Item Type	2	2	2
Demographic Items		12	12	12

2.6.1.3 ABC-1 Scale Development

As part of the Field Test, we created scales for each Phase 1 BE bias construct, together with an RD scale. In doing so, we were guided by the following goals: (1) Measure each construct as broadly as possible to ensure maximum content coverage; (2) Maximize scale reliability; (3) Create a compelling validity argument for each scale, with special emphasis on ability to detect change in scale-scores before and after bias mitigation interventions; and (4) Measure as efficiently as possible, and in no event exceed 60 minutes for any test form.

To maximize content coverage, we had developed items for each facet in the measurable content domain, described above. In developing scales, however, some facets were not equally represented because item analyses necessitated dropping a different number of items for the various facets of each scale. In determining items to retain, we conducted several statistical analyses:

- (1) We computed means, standard deviations, and other relevant statistics to identify items that were too easy, too difficult, or had anomalous frequency distributions.
- (2) We computed internal consistency reliability statistics and principal components analyses to evaluate the underlying structure of the emerging scales, and eliminate items that undermined the cohesiveness of the scales without adding to the validity argument. To this end, we reviewed corrected item-total correlations, and alpha-if-item-deleted statistics for each item, together with each item's loading on the first unrotated principal component.
- (3) We examined facet-level statistics where necessary and appropriate. Occasionally, for example, facet-level analyses revealed pockets of unidimensionality that were not obvious when analyses were conducted at the scale level. This allowed us to fine tune our item selection approach for each scale and to gain further insight into the structure underlying each scale. This, in turn, had implications for the most appropriate and interpretable reliability coefficients to use, among other things.

Five-hundred forty individuals who participated in the field test were recruited to re-take one of the three primary ABC-1 test forms (designated Forms 1-3) one month later in order to evaluate test-retest reliability. We created three additional forms, designated ABC-1 Forms 4-6, which were intended to be psychometrically equivalent to Forms 1-3. That is, Forms 4-6 differed from Forms 1-3 only superficially in that (1) they contain slightly different content designed to cloak the identity of the items without changing their measurement properties, and (2) the items are presented in a different order to further differentiate the surface characteristics of the forms. The content of the scales comprising the three primary equated forms and the three supplemental forms is listed in Table 8.

We conducted a smaller field test with an independent sample of AMT workers ($n = 280$) to verify that the supplemental forms had psychometric properties similar to the primary forms, as well as to provide data for equating Forms 4-6.

Table 8: Allocation of BE and RD Scales to ABC Test Forms

ABC Form 1	ABC Form 2	ABC Form 3
CB Scale 1	CB Scale 2	CB Scale 3
FAE Scale 1	FAE Scale 2	FAE Scale 3
BBS Scale 1	BBS Scale 2	BBS Scale 3
RD Scale 1	RD Scale 2	RD Scale 3
ABC Form 4	ABC Form 5	ABC Form 6
CB Scale 3	CB Scale 1	CB Scale 2
FAE Scale 2	FAE Scale 3	FAE Scale 1
BBS Scale 2	BBS Scale 3	BBS Scale 1
RD Scale 3	RD Scale 1	RD Scale 2

2.6.2 Results and Discussion

2.6.2.1 Descriptive Statistics

We computed raw scale-scores¹⁹ for each BE and RD scale by creating unit-weighted composites of all the item scores comprising each bias scale (see Tables 9 -12 for a listing of the items across forms for each bias scale). Higher scores for each scale indicate less bias, or in the case of RD, more knowledge of the biases. Table 13 shows descriptive statistics for each ABC-1 BE and RD scale and form and Figures 31 and 32 depict the associated frequency distributions in the form of histograms. As shown in Table 13 and Figures 31 and 32, the scale-scores for each bias have comparable statistical distributions. Note especially that the distributions are good approximations of the normal (bell-shaped) curve—one indication that they are measuring their targeted constructs well.

¹⁹ Raw scores (whether referring to computing an item score or a scale score) are based on a sum or other combination of item scores, and may be on entirely different metrics. They are distinguished from scaled scores in this technical report in that scaled scores are statistically transformed such that they are on the same metric (e.g., 0-100).

Table 9: CB Elicitation Tasks and Paradigms across Forms

Paradigm	Form 1	Form 2	Form 3
Wason Selection	Shopping Malls* Taxes	Shopping Malls* Baseball	Shopping Malls* PCT
Information Search Decision Making	Snack Stand Car Comparison* Making Music	Bakery Car Comparison* Cruises	Exercise Class Car Comparison* Working Out
Evaluation/Weighting of Evidence	Intelligence Analyst (3 items total, including 1 linking item*)	Intelligence Analyst (3 items total, including 1 linking item*)	Intelligence Analyst (3 items total, including 1 linking item*)
Evaluation/Weighting of Questions	HR Department (4 items total, including 2 linking items*)	HR Department (4 items total, including 2 linking items*)	HR Department (4 items total, including 2 linking items*)

A.1.1 *Note.* *Linking item for test equating purposes.

Table 10: FAE Behavioral Elicitation Tasks and Paradigms across Forms

Paradigm	Form 1	Form 2	Form 3
A.1.2 Attitude Attribution	A.1.3 Dolphin	A.1.4 Farming	A.1.5 Revolutionary
A.1.6 Good Samaritan	A.1.7 Good Samaritan (Purple)	A.1.8 Good Samaritan (Green)	Good Samaritan (White)
A.1.9 Quiz Role	A.1.10 Personnel Selection*	A.1.12 Personnel Selection*	A.1.14 Personnel Selection*
	A.1.11 Trivia Quiz	A.1.13 Trivia Quiz Clone (2B)	A.1.15 Trivia Quiz
A.1.16 Attributional Style	A.1.17 AS-Drew*	A.1.19 AS-Drew*	A.1.21 As-Drew*
	A.1.18 What Causes Things*	A.1.20 What Causes Things*	A.1.22 What Causes Things*
Confession	A.1.23 Sick For Party*	A.1.24 Sick For Party*	A.1.25 Sick For Party*

A.1.26	Silent Interview	A.1.27	Silent Interview – Job Interview	A.1.28	Silent Interview – Doctor	A.1.29	Silent Interview - Lying Accusation
--------	------------------	--------	----------------------------------	--------	---------------------------	--------	-------------------------------------

A.1.30 *Note.* *Linking task for test equating purposes.

Table 11: BBS Elicitation Items across Forms

BBS Difference-Score Variable	Form 1	Form 2	Form 3
Base Rate			x
Bandwagon	x	x	x
Self-Serving	x	x	x
Halo	x		
Framing	x		
In-Group		x	
Rosy Retrospection		x	
Assumed Consensus	x		
Self-Bolstering		x	
Stereotyping		x	x
Evidence Reaction			x
Adaptive Emotion	x		
Dissonance Reduction		x	
Optimism			x
Outcome			x
Generous Attribution	x		
Myside			x
Attribution		x	
Transparency	x		

Table 12: Distribution and Format of RD Items across Biases

Bias	Form 1	Form 2	Form 3
Confirmation Bias	RD3CB*	RD1CB	RD2CB
	RD5CB*	RD3CB*	RD3CB*
	RD7CB**	RD4CB	RD5CB*
	RD9CB**	RD5CB*	RD18CB**
	RD13CB**		
FAE	RD1FAE	RD3FAE	RD4FAE
	RD2FAE	RD5FAE*	RD5FAE*
	RD5FAE*	RD7FAE*	RD7FAE*
	RD6FAE	RD9FAE	RD8FAE
	RD7FAE*	RD10FAE	RD12FAE**
		RD11FAE**	
BBS	RD2BBS*	RD2BBS*	RD1BBS
	RD3BBS*	RD3BBS*	RD2BBS*
		RD6BBS**	RD3BBS*
None	RD1NONE		RD2NONE

Table 13: Descriptive Statistics for ABC-1 CB, FAE, BBS, and RD Scales by Form (Raw-scores).

Scale	Mean	SD	Skew	Kurtosis	n
CB, Form 1	41.45	8.24	.11	-.15	543
CB, Form 2	43.69	8.75	.45	.52	522
CB, Form 3	44.14	8.81	.37	.46	536
CB, Form 4	30.56	6.78	.18	.34	79
CB, Form 5	42.84	8.36	.25	.06	61
CB, Form 6	48.10	9.60	.18	.55	79
FAE, Form 1	33.05	8.89	.36	-.15	543
FAE, Form 2	27.87	8.41	.41	.24	522
FAE, Form 3	28.77	10.46	.43	.10	536
FAE, Form 4	23.38	8.20	1.58	4.95	101
FAE, Form 5	25.01	9.66	-.13	-.08	81
FAE, Form 6	25.85	10.27	-.26	-.38	98
BBS, Form 1	35.06	5.71	-.44	-.08	542
BBS, Form 2	33.72	6.30	-.40	-.40	522
BBS, Form 3	35.01	5.72	-.50	-.08	536
BBS, Form 4	34.09	6.20	-.39	-.49	95
BBS, Form 5	34.64	5.27	-.20	-.91	72
BBS, Form 6	33.54	5.92	-.66	.61	89
RD, Form 1	9.21	2.97	-.61	-.64	543
RD, Form 2	9.53	3.09	-.88	-.23	522
RD, Form 3	8.60	3.08	-.50	-.73	536
RD, Form 4	9.25	2.56	-.70	-.04	95
RD, Form 5	9.76	2.78	-.73	-.38	70
RD, Form 6	10.52	2.68	-1.47	1.75	88

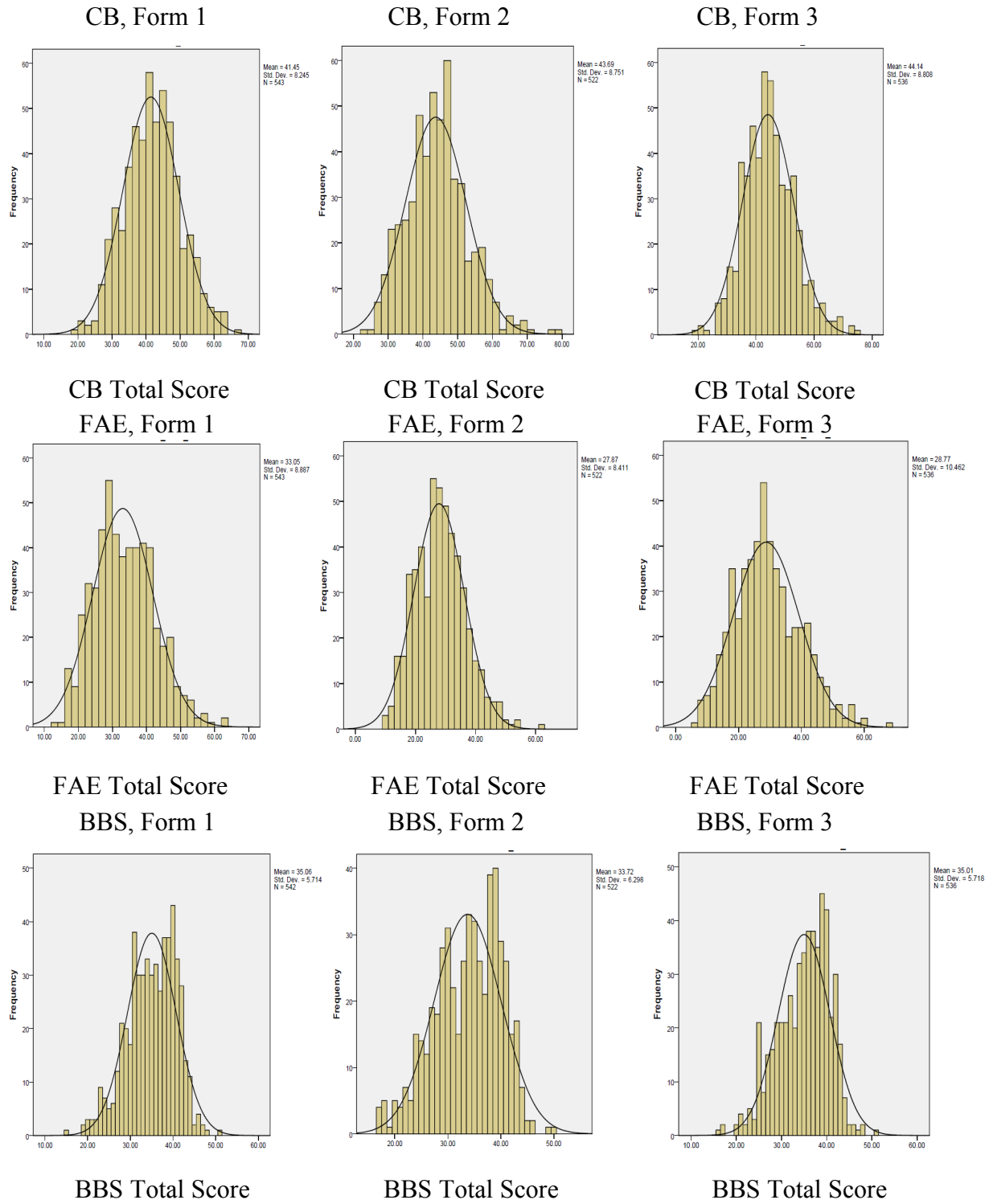


Figure 31: Histograms Depicting ABC-1 CB, FAE, and BBS Total Raw-Score Frequency Distributions by Form (1-3).

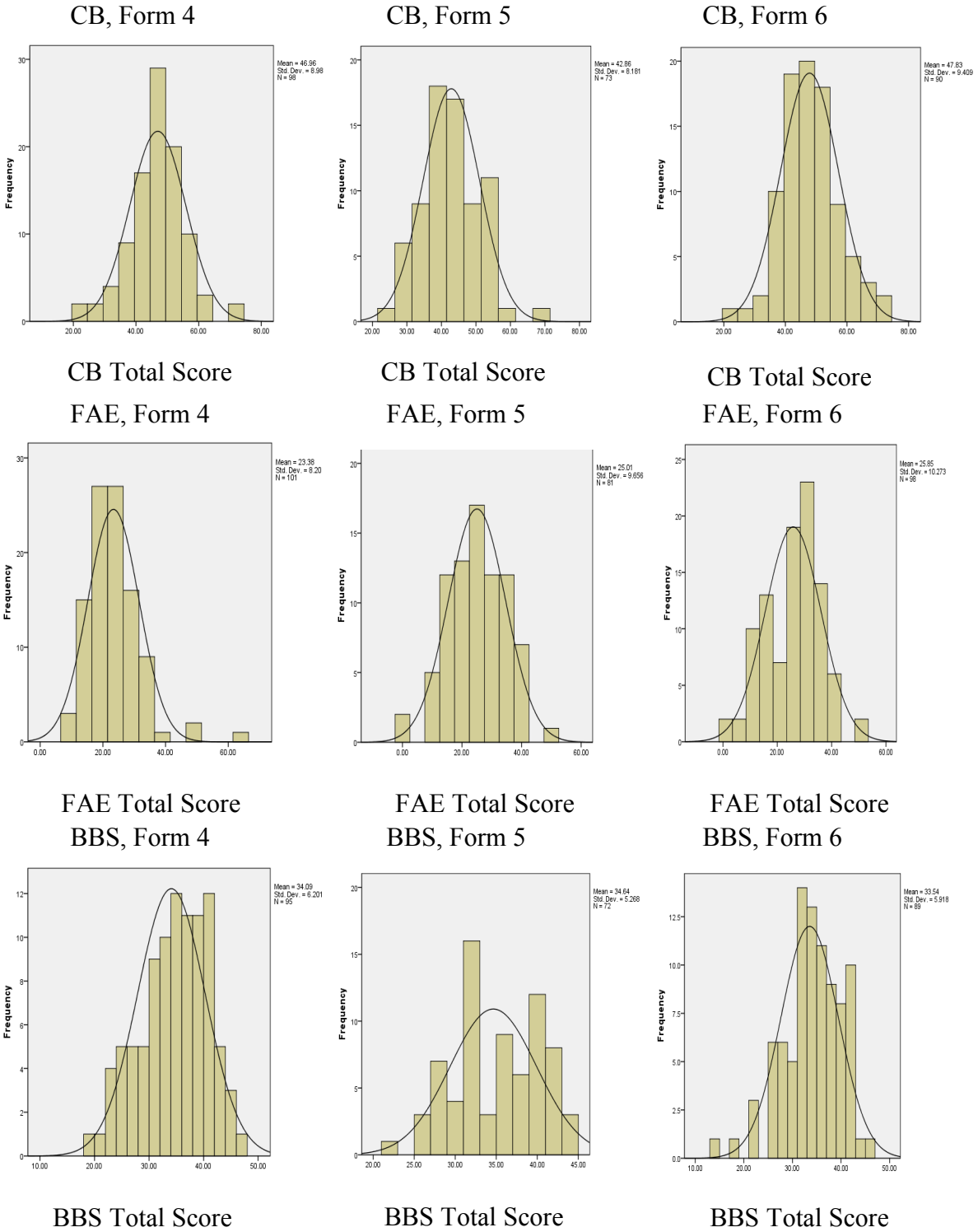


Figure 32: Histograms Depicting ABC-1 CB, FAE, and BBS Total Raw-Score Frequency Distributions by Form (4-6).

Figure 33 depicts the frequency distributions in the form of histograms of the total RD raw scores for each of the ABC-1 test forms. In general, the results reveal that the test forms are at an appropriate level of difficulty for the RD construct (mean raw score = 9.48 out of 13, or 73% correct), with a slight negative skew. That is, the test should not be too difficult if participants read and understood the text descriptions of the biases reasonably well. The histograms show the slight negative skew, and also show substantial variance across participants, indicating that the RD test differentiates across test-takers.

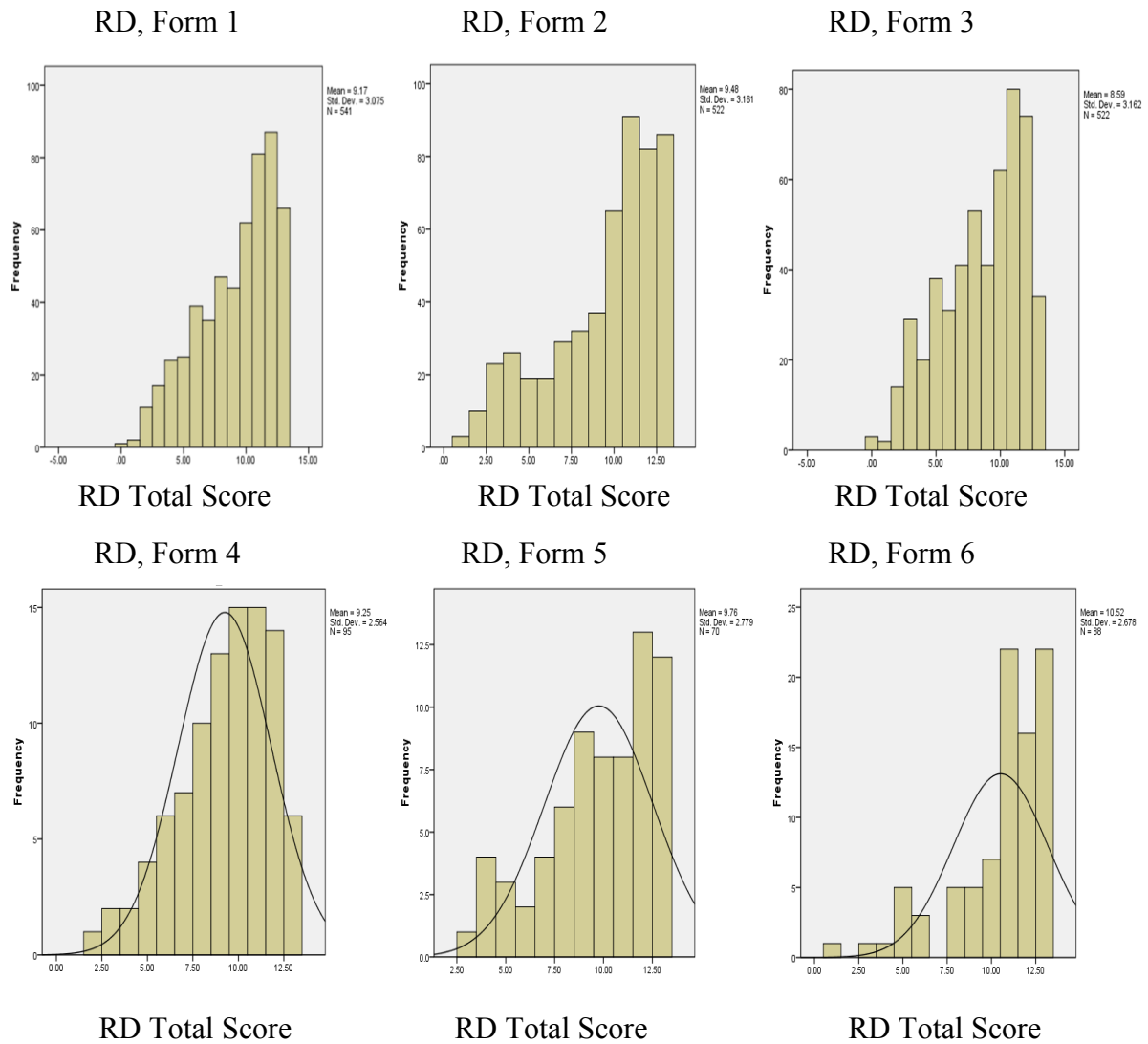


Figure 33: ABC-1 RD Total Raw-score Frequency Distributions.

2.6.2.2 Reliability Analyses and Results

Table 14 presents various reliability estimates for the ABC-1 scales based on the field test data. For each BE test form, we computed Cronbach's alpha, a measure of internal consistency reliability, and test-retest reliability, a measure of temporal stability. The retest interval was approximately 1 month.

With the exception of the CB scales, alpha coefficients were acceptably high, and their magnitude was, to a large extent, determined by the number of items in each form. The alpha coefficients associated with the CB scales were modest at best, ranging from .49 to .57. Test-retest reliabilities²⁰ indicate that the RD forms are temporally stable, but that the temporal stability of the BE forms varies considerably within bias construct and is often modest at best. With the exception of the CB scales, alpha typically exceeds test-retest reliability.

Table 14 also shows the mean loading on the first unrotated principal component for each form. This is a useful metric for evaluating unidimensionality. It is not influenced by number of items, and alpha can be high even in the presence of multidimensionality (Cortina, 1993). The mean PC loading for CB and FAE are both somewhat low, despite the fact that FAE has a high alpha coefficient and CB has a relatively low alpha coefficient. While FAE has a slightly higher alpha than CB, the main reason for the disparity in alpha is the much greater number of items in the FAE scales than in the CB scales. BBS and RD have substantially higher mean component loadings than FAE and CB. Those mean component loadings provide additional evidence of unidimensionality for the BBS and RD scale forms.

Table 14: Reliability Analysis of ABC-1 Confirmation Bias (CB), Fundamental Attribution Error (FAE), Bias Blind Spot (BBS), and Recognition and Discrimination (RD) Scales

Scale	$r_{xx'}$	n	α	Mean Loading on 1st Unrotated PC	n	N of items
CB, Form 1	.62	121	.57	.34	543	12
CB, Form 2	.46	129	.51	.33	522	12
CB, Form 3	.46	131	.49	.33	536	12
Mean	.51		.52	.33		
CB, Form 4	-	-	.47	.34	98	12
CB, Form 5	-	-	.46	.22	73	12
CB, Form 6	-	-	.50	.32	90	12
Mean			.48	.29		
FAE, Form 1	.66	125	.83	.21	543	80

²⁰ Test-retest reliability is a reliability metric that evaluates the extent to which test-takers are rank-ordered similarly on two different testing occasions separate in time. As such, it is an index of temporal stability. It is typically symbolized as $r_{xx'}$.

Scale	$r_{xx'}$	n	α	Mean Loading on 1st Unrotated PC	n	N of items
FAE, Form 2	.51	131	.85	.17	522	82
FAE, Form 3	.50	138	.84	.26	536	81
Mean	.56		.84	.21		
FAE, Form 4	-	-	.87	.27	89	82
FAE, Form 5	-	-	.81	.28	72	81
FAE, Form 6	-	-	.86	.19	89	80
Mean			.85	.25		
BBS, Form 1	.50	121	.67	.55	543	8
BBS, Form 2	.48	130	.73	.59	522	8
BBS, Form 3	.63	130	.65	.53	536	8
Mean	.54		.68	.56		
BBS, Form 4	-	-	.74	.59	95	8
BBS, Form 5	-	-	.55	.45	73	8
BBS, Form 6	-	-	.69	.56	89	8
Mean			.66	.53		
RD, Form 1	.68	109	.79	.52	543	13
RD, Form 2	.73	115	.82	.56	522	13
RD, Form 3	.77	116	.81	.54	536	13
Mean	.73		.81	.54		
RD, Form 4	-	-	.73	.48	95	13
RD, Form 5	-	-	.77	.51	71	13
RD, Form 6	-	-	.80	.54	88	13
Mean			.77	.51		

Note. $r_{xx'}$ is test-retest reliability. Test-retest data were not obtained for Forms 4-6.

2.6.2.3 ABC-1 Intercorrelations

Table 15 shows intercorrelations between each of the ABC-1 scale-scores across forms. Most notably, Table 15 does not show a positive manifold; i.e., there is no general tendency for scale scores to intercorrelate positively with one another. The lack of intercorrelation makes computation of an overall battery score inappropriate due to lack of interpretability. In other words, intercorrelation between variables indicates that they share an empirical theme, not just a rational theme. Having an empirical theme means that combining the variables into a composite (e.g., adding them up) will likely result in a better measure of whatever construct underlies that empirical theme. If there is no empirical theme, one may still combine the variables, but must do so with the understanding that they are not adding up to a measure of a common construct. Rather, they are independent measures of different constructs. This may be useful if one's intention is to predict an outcome with which each of the independent constructs is related, but

not if one's intention is to derive a measure of a common construct, such as general (overall) bias susceptibility.

Table 15: Intercorrelations between ABC-1 Scale-Scores.

	ABC-1 Scale-score	CB	FAE	BBS	RD
1	Confirmation Bias (CB)		.09	-.02	.06
2	Fundamental Attribution Error (FAE)	.06		-.01	.06
3	Bias Blind Spot (BBS)	-.01	-.01		-.36
4	Recognition and Discrimination (RD)	.04	.05	-.27	

Note. $n = 1,601$. Correlations based on observed scores are below the diagonals and correlations based on scores disattenuated for measurement error are shown in bold above the diagonal.

There are two additional points to be made about these results. First, there is a moderately high negative correlation between BBS and RD ($r = -.27, p < .01$; corrected $r = -.36$). Recalling that a high score indicates the absence of a bias, this result indicates that people who are more susceptible to BBS score higher on RD items. Given that RD correlated between the high .30s and high .40s with cognitive ability measures in various pilot tests that we conducted prior to the equating study, an intriguing interpretation of this result is that test-takers higher in cognitive ability really are less biased than the average person, which suggests that BBS may be, at least partially, an epiphenomenon. That is, the status of bias blind spot as a construct that accounts for individual differences in people's beliefs about their susceptibility to biases relative to other people may require reevaluation. However, this interpretation is not entirely satisfactory, because of the weak (near zero) correlations between RD and both CB BE and FAE BE. The magnitude of the latter two correlations suggests another important conclusion: that initial knowledge about these biases appears orthogonal to whether or not one is susceptible to these biases.

Second, there are no other consistent, practically significant intercorrelations between BE scales. Although there was a small correlation between FAE BE and CB BE ($r = .06, p < .05$; corrected $r = .09$), the intercorrelations were not statistically significant across forms.

2.6.2.4 Scoring Modifications

During the phase 1 Field test, various modifications were made to the method by which the ABC-1 BE scales were scored. Changes that were made are described in the following subsections.

2.6.2.4.1 Confirmation Bias

We examined several different scoring approaches for measuring CB within each task. Early scoring approaches resulted in low internal consistency reliability ($\alpha = .20$). One limitation of our early attempts at scoring CB was that the scoring models could not distinguish between different patterns of responding. As operationalized in the ABC-1, the scoring model used for each CB task seeks to capture different patterns and/or levels of task performance. They include the following: (1) Disconfirmatory response patterns; (2) Adaptive, balanced response patterns (indicating lack of CB or adaptive CB); (3) Exhaustive, high effort response patterns (this only

applies to information search paradigms); (4) Minimal, low effort response patterns (again, this only applied to information search paradigms); (5) Moderate CB indicated by moderately high, disproportionate selections of confirmatory responses; and (6) Extreme CB, whereby nearly all or all responses are confirmatory.

2.6.2.4.2 Fundamental Attribution Error

ABC-1 FAE items consisted of Likert-type scales whose ratings were made using a multiple-choice, selected response format. Ratings were dichotomized such that FAE was coded as 0 and non-FAE was coded as 1. The primary reason why we dichotomized the items was because Item Response Theory²¹ (IRT) analyses on the phase 1 field test data revealed that dichotomizing yielded information (in an IRT sense) that spanned more of the trait continuum. As such, measurement quality across the trait continuum was better. Each FAE task score was computed as the sum of these dichotomized ratings.

2.6.2.4.3 Bias Blind Spot

There were only very minor modifications to the scoring of the ABC-1 BBS scale forms. First, scoring was reversed such that higher scores reflected lower bias. Second, we linearly transformed the scale-scores to eliminate negative scores.

2.6.2.5 Conclusions Regarding Structure and Individual Difference Measurement of Biases

There is no support for a common bias susceptibility construct. That is, bias susceptibility appears to be formative rather than reflective. It is best conceptualized as linear combinations of bias susceptibility scales, many of which are not related to (i.e., correlated with) one another empirically. The recognition and discrimination scale is internally consistent and a good approximation of unidimensionality.

This score is best understood as a concatenation of thematically related measures of the Phase 1 biases rather than a unidimensional scale measuring elicitation of CB, FAE, and BBS. As such, “overall bias susceptibility” in this context is a label of convenience only.

The only way to achieve unidimensionality would have been to substantially reduce the ABC-1's validity; for example, by limiting a great deal of critical content (attenuating content validity) and measuring so transparently that we would be measuring RD rather than BE (attenuating construct validity by introducing a confound). We believe that the ABC-1's validity renders it a fair measure of the content domain underlying the Phase 1 Sirius biases.

2.6.3 Pretest Sensitization Study

2.6.3.1 Purpose

Following the ABC-1 Field Test study, we conducted a study designed to investigate whether the ABC-1 creates a pre-test sensitization effect. Specifically, we evaluated whether taking the ABC-

²¹ A mathematical model of the functional relationship between performance on a test item, the test item's characteristics, and the test-taker's standing on the construct being measured.

1 interacts with bias mitigating training interventions, thereby resulting in different post-intervention ABC-1 test performance than if no pretest were given.

2.6.3.2 Method

2.6.3.2.1 Participants

The Pretest Sensitization Study was administered to a total of 67 ETS employees (17-29 per condition), all of whom were located in the United States and recruited online through ETS's Performance Assessment Scoring Services (PASS). The sample of test-takers averaged 43 years of age (SD = 10 years); was 74% female; and approximately 72% Caucasian, 9% African-American, 4% Asian, and 6% multi-racial. Nine percent of the sample was Hispanic or Latino. All test-takers had graduated college, and the sample was high-achieving, with over 75% reporting GPAs over 3.5. The majority (74.6%) of the sample reported having taken at least one or two psychology courses, though the vast majority (77.8%) had not taken more than four psychology courses. Participants were compensated at a rate of \$20/hour for completing the study.

2.6.3.2.2 Experimental Design and Procedure

Study participants were randomly assigned to the following three groups:

- (1) IARPA Video Experimental Group: Participants took the ABC-1, Form 1, as a pretest, then they watched the IARPA instructional video about the Phase 1 cognitive biases before taking a posttest, which was the ABC-1, Form 2.
- (2) Control Video Group: Participants first took the ABC-1, Form 1, as a pretest; then watched an unrelated instructional video before taking the ABC-1, Form 2, as a posttest. The instructional video was a 30-minute lecture given by Dr. Steven Pinker from Harvard University about language and psychology, a topic unrelated to cognitive biases. The video was selected, because it is approximately the same length as and has audio-visual features comparable to the IARPA instructional video.
- (3) No Pretest Group: Participants did not take a pretest prior to watching the IARPA instructional video, but only took the ABC-1, Form 2, as a posttest.

Participants were emailed a link to access the assessments and instructional videos. In addition, we administered the NASA-TLX workload questionnaire immediately after both the pre- and/or posttests in order to investigate the user acceptability of the ABC test forms under conditions that were comparable to the Phase 1 IV&V video-control conditions. Last, participants completed the BFI-44 personality and demographics questionnaires. Participants were instructed to complete all study activities in a single session lasting approx. 2 – 3.5 hours.

2.6.3.3 Results and Discussion

We computed and compared pretest and posttest scale-scores for each ABC scale and group (as well as NASA-TLX ratings) in order to evaluate the following hypotheses:

- (1) Assuming Pre-test scores are equivalent across groups, Post-test scores will be higher in the IARPA Video Experimental Group and No Pre-test Group than in the Control Video Group for targeted ABC scales.

- (2) Because the IARPA instructional video most directly targets explicit/declarative knowledge of cognitive biases, improvement will be shown for the ABC Recognition and Discrimination Scale, but there may be little or no improvement in the Behavioral Elicitation scales.
- (3) ABC pre-test will neither substantially decrease engagement and effort, nor increase frustration and physical demands taking a post-test.

Table 16 reports mean pretest and posttest scale-scores for each ABC-1 scale and group. Pretest. Pre-test performance was generally comparable across groups for each scale. Consistent with Hypotheses 1 and 2 above, watching the IARPA instructional video enhanced performance on the RD post-test relative to watching the Control video [$F(1,35) = 6.47, p < .05$]. RD scores improved by 23 points from 65 to 88 in the IARPA instructional video condition (Cohen's $d = 1.22^{22}$). RD scores also improved in the Control Video condition, but to a substantially lesser extent (Cohen's $d = 0.54$). RD post-test scores were equivalent in the IARPA Video and No Pre-Test groups (Means = 88 vs. 87). By contrast, differences in pretest and posttest BE scores were small and not statistically significant. As shown in Table 17, we conducted an Analysis of Covariance (ANCOVA), which revealed that BE posttest scores were not significantly higher in the IARPA Video group as compared to the Control Video group when controlling for BE pretest scores.

²² d (sometimes referred to as “Cohen’s d ”) is an effect size quantifying the standardized difference between two independent means.

Table 16: ABC-1 Confirmation Bias (CB), Fundamental Attribution Error (FAE), Bias Blind Spot (BBS), and Recognition and Discrimination (RD) Pretest and Posttest Scale-scores.

Group	Mean CB Scores		Mean FAE Scores		Mean BBS Scores		Mean RD Scores	
	Pre-Test	Post-Test	Pre-Test	Post-Test	Pre-Test	Post-Test	Pre-Test	Post-Test
IARPA Video (N=21)	52 (11)	58 (13)	53 (13)	47(12)	69 (10)	65 (10)	65 (22)	88 (15)
Control Video (N=17)	52 (6.8)	60 (13)	48 (13)	45 (11)	68 (9.4)	62 (14)	68 (14)	78(22)
No Pre-Test (N=29)	-	64 (15)	-	46 (11)	-	61 (12)	-	87 (18)

Note. Values in parentheses are standard deviations. Each scale is standardized to a mean of 50 and SDs of 12 for the CB, FAE, and BBS BE scales and 22 for the RD scales, and ranges from 0-100.

Table 17: Summary of ANCOVA Results for ABC-1 BE and RD Measures: Did the IARPA Video Group Do Better Than the Control Group, Controlling for Pretest Scores?

Bias Scale	P-Value for Condition (IARPA Video Versus Control Video Treatment Group)	Higher Score in Experimental Group Than in Control Group?	Effect Size (Partial Eta-Squared ²³)
Confirmation Bias	.49	No	.01
Fundamental Attribution Error	.21	Yes	.05
Bias Blind Spot	.42	Yes	.02
Recognition and Discrimination	< .05	Yes	.15

We next examined pretest and posttest ratings on the NASA-TLX scales for each of the groups. (The NASA-TLX is a widely-used measure of mental workload, and its components). Tables 14 and 15 report mean ratings for each of the six NASA-TLX sub-scales following completion of the ABC pretest and posttest. As shown in Table 18, individuals reported high mental demand overall, but slightly less mental demand for the ABC post-test [Pre-test ratings ($M = 7.1, SD = 2.0$) > Post-test ratings ($M = 6.7, SD = 2.4$), $F(1,35) = 6.2, p < .05, \eta_p^2 = .15$]. By contrast, individuals reported relatively low physical demand, but there was slightly higher physical

²³ In this context, partial eta squared is the proportion of the bias mitigation effect (+ error) that is attributable to the bias mitigation intervention (i.e., the serious games).

demand reported with respect to the post-test [Post-test ratings ($M = 1.8$, $SD = 1.7$) > Pre-Test ratings ($M = 1.5$, $SD = 1.3$), $F(1,35) = 6.4$, $p < .05$]. In terms of temporal demand, whereas ratings in the IARPA video condition decreased on the post-test (4.7) relative to pre-test (5.1), ratings in the Control Video condition increased on the post-test (5.4) relative to the pre-test (4.7) [Marginally significant Test (Pre- vs. Post-) \times Condition (IARPA Video vs. Control Video) interaction, $F(1,35) = 3.5$, $p = .07$, $\eta_p^2 = .09$].

Table 18: NASA-TLX ratings following completion of the ABC pretest and posttest.

Group	Mental Demand		Physical Demand		Temporal Demand	
	Pre-Test	Post-Test	Pre-Test	Post-Test	Pre-Test	Post-Test
IARPA Video (N=21)	7.6 (1.6)	7.1 (1.9)	1.6 (1.5)	1.8 (1.8)	5.1 (2.6)	4.7 (2.6)
Control Video (N=17)	6.8 (2.3)	6.1 (2.8)	1.4 (1.1)	1.9 (1.6)	4.7 (2.9)	5.4 (2.7)
No Pre-Test (N=29)	-	6.4 (1.5)	-	1.7 (1.2)	-	4.9 (1.9)

As shown in

Table 19, participants in the IARPA Video condition showed substantially greater increase in subjective confidence in their post-test performance relative to their pre-test performance than individuals in the Control Video Condition [Test (Pre- vs. Post-) \times Condition (IARPA Video vs. Control Video) interaction, $F(1,34) = 6.5, p = .02, \eta_p^2 = .16$]. In terms of effort, there was no significant decrease in effort reported on the posttest as compared to the pretest, which suggests that participants remain engaged when taking the test a second time. Participants in the Control Video group reported somewhat greater frustration than in the IARPA Video condition following the post-test, but this effect was not statistically significant due to limited power. This latter finding is perhaps consistent with confidence in one's increased declarative knowledge having some stress-reducing effect, though a stress measure would provide more direct evidence.

Table 19: NASA-TLX ratings following completion of the ABC pretest and posttest.

Group	Performance		Effort		Frustration	
	Pre-Test	Post-Test	Pre-Test	Post-Test	Pre-Test	Post-Test
IARPA Video (N=21)	5.1 (2.3)	3.6 (2.1)	7.2 (1.8)	6.6 (1.8)	4.9 (2.4)	3.7 (1.9)
Control Video (N=17)	4.4 (1.7)	4.9 (2.0)	5.9 (2.6)	5.9 (2.6)	5.4 (2.4)	5.3 (2.7)
No Pre-Test (N=29)	-	4.3 (2.1)	-	6.2 (2.0)	-	4.9 (2.3)

Taken together, these results indicate that exposure to the ABC pretest does not appear to have a practically significant effect on BE or RD bias measures. Moreover, the pattern of changes in pretest to posttest scores suggests that participants did learn from watching the IARPA instructional video; however, they acquired explicit/declarative knowledge rather than procedural knowledge of the Phase 1 cognitive biases. The finding of a dissociation in performance between BE and RD bias measures as result of watching the IARPA instructional video relative to an unrelated control video is consistent with the hypothesis that this type of relatively mild, brief training intervention should have a stronger effect on RD measures. As such, this finding may be considered discriminant validity evidence for the ABC, providing additional support for its construct validity. However, an alternative explanation is that the instructional video featured scenarios with younger adult college students, and the study participants, being all older post-college professional, may have found the video to be informative, but not personally relevant. It's therefore possible that these findings may not generalize to college students for whom the instructional video may have a potentially significant impact on both BE and RD test performance. Last, the findings from NASA-TLX ratings indicated that the ABC requires high mental demand and effort and that taking an initial pre-test does not make the post-test much easier to complete.

2.6.4 ABC-1 Implementation

2.6.4.1 Development of Equivalent Forms

One of the requirements of this project for MITRE/ETS was to develop equivalent forms²⁴ of the ABC-1 and ABC-2. Equating involves small statistical adjustments to account for minor differences in the difficulty of alternate test forms. Subsequent to equating, these alternate forms

²⁴ Two or more versions of a test that are considered interchangeable in that they measure the same attribute in the same way, are built to the same specifications, and are administered under the same conditions using the same directions.

can be used interchangeably (Lord, 1980) despite the fact that they are comprised of different items (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 97). Equating is more precise at the group, as opposed to the individual, level (Kolen & Brennan, 2004).

A variety of methods have been developed to equate tests. For the ABC-1, we chose to use an equipercentile approach with log-linear pre-smoothing to equate the three forms. To accomplish this, we used the "Equate" package in R (Albano, 2011; R Development Core Team, 2013). *Equipercentile equating* involves executing a transformation such that standardized scale-scores on two forms correspond to the same percentile. In practice, the score distributions may be jagged; therefore, smoothing the distributions prior to matching the percentiles provides a more stable and tractable equating solution. A third-degree loglinear polynomial smoothing was used (this maintains the mean, standard deviation, and skewness of the unsmoothed distribution). This is one of the most commonly used equating methods (Kolen & Brennan, 2004). When properly implemented, equipercentile equating yields equated scores across test forms such that the percentile ranks of respective forms do not change.

As with other equating methods, equipercentile equating produces “concordance tables.” Concordance tables relate raw-scores on one equated test form to raw-scores on other equated test forms. That is, they provide a mechanism for indicating which raw-scores across equated test forms indicate the same trait level on the construct measured by the test forms. Concordance tables were created using the equipercentile approach described above for the ABC-1, and were provided to JHUAPL for use in the IV&V.

2.6.4.2 Completion Time for ABC-1

In order to meet the IV&V operational requirement that the ABC-1 scales take between 45 and 60 minutes for test-takers to complete, we analyzed the timing data from each task administered in the Field Test. We calculated the mean and median times, as well as the SDs, for each ABC-1 task, and identified additional tasks for removal, because they took too long relative to the amount of information they provided. Table 20 reports the descriptive statistics (Mean, Median, and SDs) for ABC-1 completion times for each of the three primary ABC-1 test forms in both the Field Test and Retest studies.

Table 20: Completion times for the ABC-1 (in minutes).

Test	ABC Form 1	ABC Form 2	ABC Form 3
Field Test	Mean = 57.0	Mean = 56.9	Mean = 58.7
	Median = 55.6	Median = 54.8	Median = 56.9
	Stand Dev. = 17	Stand. Dev. = 16	Stand Dev. = 16
Re-Test (1 month later)	Mean = 50.3	Mean = 50.7	Mean = 57.1
	Median = 48.3	Median = 47.2	Median = 55.5
	Stand Dev. = 14	Stand Dev. = 17.8	Stand Dev. = 17

Key Findings: Not surprisingly, the median completion times were 1-2 minutes lower than the mean completion times. This was due to the influence of extreme outlier response times. As

such, the median values are a more appropriate estimate of the average ABC-1 completion time. Second, test-takers took less time to complete the ABC-1 during the retest study conducted approximately 1 month following the Field Test.

2.6.4.3 ABC-1 User Manual and Deployment Package

To facilitate a smooth handoff of the ABC-1 to JHUAPL and any other future users of the ABC-1, we created a User Manual for use/adaptation of the ABC-1 test battery. The User Manual was not intended to serve as a definitive technical manual, but rather as a source of important information and useful guidance for implementation of the ABC-1 in the IV&V phase. The main purposes of this User Manual were to:

1. Describe the content of the ABC-1, and provide illustrative ABC-1 items/tasks;
2. Describe and explain the scoring process for the ABC-1 scales and overall battery score;
3. Describe test equating methodology used to create concordance tables that link ABC-1 scores across test forms, present and describe those concordance tables, and provide examples illustrating how to use them; and
4. Describe data processing and syntax files created to score the ABC forms.

Along with the ABC-1 User Manual, we included a deployment package to further facilitate the implementation of the ABC-1 in the IV&V phase of the project. The deployment package included:

1. Python scripts and associated files configured to process raw data files from individual test-takers and transform them into a single, master data set (.csv format).
2. SPSS syntax files, one for each ABC-1 form. Each such file has the syntax necessary to compute all the group-level total-scores for the ABC-1 scales.
3. A Microsoft Excel file that provides information about the content of the syntax files to facilitate navigation of the large number of variables that were necessary to score the ABC-1.

2.6.4.4 Identification and Resolution of Implementation Issues

Subsequent to delivery of the ABC-1 User Manual and deployment package, the IV&V team identified issues relevant to successful implementation of the ABC-1. First, concerns were raised over the content and scoring of the ABC-1 CB scale. More specifically, one of the performer teams noted that one of the CB paradigms has been the subject of controversy in the extant CB literature. Therefore, at IARPA's request, we provided scoring and equating materials for the ABC-1 CB scale both with and without inclusion of that paradigm. Second, during the process of integrating our syntax into the JHUAPL statistical software, a few minor errors were discovered in the SPSS syntax files included in our deployment package. We corrected and sent revised SPSS files to JHUAPL. Third, in the process of analyzing IV&V data for ABC-1 Forms 4-6, JHUAPL discovered non-trivial statistical differences in the descriptive statistics for analog scales across certain forms.²⁵ As a result, a decision was made to pool our field test study data

²⁵ In this context, "analog scale" means an independently developed scale intended to operationalize the same bias construct.

and create revised concordance tables. Those revised concordance tables were then transmitted to JHUAPL.

2.7 ABC-1 Integrative Summary

The ABC-1 reflects the complexity of measuring recognition and discrimination, and behavioral elicitation of confirmation bias, fundamental attribution error, and bias blind spot. Successful measurement of these constructs required extensive pilot testing (and in some cases, the development of complex scoring algorithms). As such, the road from extracting individual scores from the MITRE computer platform to group-level ABC-1 battery scores was both long and winding. Because of the novelty of the constructs, great care was taken to review the relevant literature comprehensively and extract the most promising paradigms and information to develop prototype items. These were intended to operationalize the measurable facets of each construct. Substantial research was done to identify and refine the most promising prototypes, and to “clone” those prototypes, once identified. This was an iterative process, requiring several rounds of research prior to the actual field test. In conjunction with our other research, we conducted cognitive laboratory work, which served primarily as a usability study. Items were also carefully reviewed for both quality and fairness. We developed a test administration platform specifically to support the authoring and administration of the ABC-1. The platform was designed for web-based test administration and hosted on a secure web server. The platform was also designed to facilitate the authoring, revision, and exporting of test-taker responses. In general, this test delivery software was designed to accommodate a wide variety of item/task types in the ABC-1 and to maximize usability, flexibility, and security. Last, we created a lengthy and specific ABC-1 User Manual as part of a deployment package provided to JHUAPL for the IV&V.

While the biases on the RD scale clearly form one dimension, the same is not true for the BE scales. With the possible exception of BBS, the BE scale-scores, as well as an overall battery score, are best understood as a concatenation of thematically related measures of the Phase 1 biases rather than unidimensional bias susceptibility measures. That is, they are essentially linear combinations of the items/scales of which they are comprised. Such measures are often referred to as “formative.” Indeed, even within certain biases (particularly CB), tasks that comprise their content are only modestly related. This created a trade-off between (1) maximizing capture of CB content, and (2) maximizing internal consistency such that the ABC-1’s CB BE – and, to a lesser extent, the FAE BE – scale, as well as an overall battery composite, could be interpreted as unidimensional.

Because the primary purpose of developing the ABC was to detect changes due to bias mitigation interventions, part of our pilot test research included a bias mitigation intervention study. Results of that study showed that the ABC-1 BE scales did not change on posttest, although the RD scale did. This was likely due to the fact that the BE scales required procedural knowledge, whereas the RD knowledge required only declarative knowledge.

3 Phase 2

Phase 2 of the project largely recapitulated Phase 1, except that three different biases were targeted for measurement: Anchoring Bias (ANC), Representativeness Bias (REP), and Projection Bias (PRO). In addition, as with Phase 1, Phase 2 included an RD test, except that the

RD content related to the three Phase 2 biases rather than the Phase 1 biases. In the following sections, we document test development and related activities specific to Phase 2.

3.1 Significant Changes between Phases 1 and 2

In this section, we describe several differences between Phase 1 and 2 due either to lessons learned from Phase 1 or requirements related to measurement of the Phase 2 biases.

3.1.1 Bias Instrument Coordinating Committee

One way in which Phase 2 differed from Phase 1 is that in Phase 2, a Bias Instrument Coordinating Committee (BICC) was established that consisted of representatives of the Sirius research teams, IARPA, JHUAPL, MITRE, and ETS. Their role was to create an assessment for video game pilot testing using a common set of items so that IARPA could compare “apples to apples” when trying to evaluate and compare results from the performers’ individual pilot testing. The ABC-2 represents an independent effort on the part of ETS and MITRE to develop a reliable, fair, and valid assessment of the Phase 2 biases.

3.1.2 Early and Increased Use of Online Crowdsourcing

We found use of online crowdsourcing services exceptionally useful during Phase 1 of this program. For example, Amazon Mechanical Turk (e.g., Buhrmester, Kwang, & Gosling, 2011) provided a means for recruiting and testing a large number of research study participants quickly and efficiently. As such, we increased the use of AMT in order to maximize the amount of research that could be implemented to support development of the ABC-2. As in Phase 1, we continued to use other methods of data collection as well, including use of college undergraduates and targeted adult populations (e.g., ETS employees).

3.1.3 Empanelled new TAG for Phase 2

For Phase 2, we empaneled a new technical advisory group (TAG) to advise us throughout the course of Phase 2. We invited Drs. Larry Jacoby and Steve Reise back to serve on the Phase 2 TAG, because they made critical contributions to the development of the ABC-1, and their expertise was equally relevant to Phase 2. In addition, we invited Dr. Raymond Nickerson to join the Phase 2 in order to bring added expertise in the judgment and decision-making field to the project. Dr. Nickerson (Tufts University, <http://ase.tufts.edu/psychology/peopleNickerson.htm>) is a former senior vice president of Bolt Beranek and Newman Inc. (BBN Technologies), from which he is retired. His Ph.D., in experimental psychology, is from Tufts University. He is a fellow of the American Association for the Advancement of Science, the American Psychological Association, the Association for Psychological Science, the Human Factors and Ergonomics Society and the Society of Experimental Psychologists. A past chair of the National Research Council's Committee on Human Factors (now the NRC Board on Human-Systems Integration), and a recipient of the Franklin V. Taylor Award from the American Psychological Association, he was the founding editor of *The Journal of Experimental Psychology: Applied* and of *Reviews of Human Factors and Ergonomics*, a series published by the Human Factors and Ergonomics Society. Dr. Nickerson's research interests include cognition, human factors and applied experimental psychology. His recent work at Tufts has focused primarily on probabilistic reasoning.

Three TAG meetings were held during Phase 2 in which TAG members reviewed and provided input on project documentation, such as the literature review, research plan, descriptions of item prototypes, results from pilot research, and designs for proposed studies and additional item types.

3.1.4 Updating of Literature Review

In preparation for developing a new test instrument for the Phase 2 IV&V during the summer of 2013, we revisited original Gertner et al. (2011) literature review and made updates to the sections that pertained to the Phase 2 biases. Updates of the original literature review included:

- New, real-life examples of the Phase 2 biases
- Clarification of results described in our 2011 literature review
- Additional explanations of results in bias studies
- Replications of, or failures to replicate, previous findings
- New results that carry implications for measurement of bias elicitation (e.g., dichotomous versus continuous scales)
- Identification of different possible mechanisms through which biasing stimuli exert their effects
- Additional theoretical and empirical work on cognitive processes underlying biases, including more parsimonious explanations and evaluation of boundary conditions
- New results regarding individual-difference correlates of bias susceptibility (e.g., personality, cognitive ability, knowledge, and expertise)
- New results regarding influences on the extent of bias mitigation techniques, and explanations of bias mitigation
- A substantially expanded section on focalism as it relates to anchoring
- Additional information about causes of shifts from System 1 to System 2 cognitive processing, which carries implications for bias mitigation interventions

3.2 Construct Identification

As in Phase 1, we partitioned the content domain for each of the Phase 2 biases based on our updated literature review (Gertner et al., 2013). The initial Phase 2 facets and their definitions generated substantial discussion among members of the Phase 2 TAG and IV&V team. This resulted in various modifications to the content domain. During the course of Phase 2 test development and associated research, additional changes were made. Table 21 shows the final working definitions of the Phase 2 bias constructs and their associated facets.

Table 21: Working Definitions of Phase 2 Biases: Behavioral Elicitation of Cognitive Bias Measures

Bias	Definition
Anchoring	The tendency to rely too heavily or overly restrict one’s attention to one trait or piece of information when making judgments. The information in question can be relevant or irrelevant to the target decision, as well as numerical or non-numerical. Includes focalism or the focusing illusion.
Anchoring Facet 1: Numerical Priming	The tendency to base judgments on a single decontextualized piece of numerical information from an external source (either numbers or words reflecting size or magnitude [e.g., large, small]); the information can be explicit or implicit.
Anchoring Facet 2: Selective Accessibility	The tendency to base judgments on a selective search for information that explains the difference between an externally-provided anchor and what people consider to be plausible values near that anchor. This selective search creates confirmatory hypothesis testing that biases judgments in the direction of the anchor. This facet of anchoring is most likely to be elicited in situations where context and knowledge play a significant role.
Anchoring: Facet 3: Comparative Judgment	The tendency to base judgments on an anchor generated by a comparative judgment task. The anchor generated by the comparative judgment then influences a subsequent absolute judgment. The anchoring effect is seen in that subsequent absolute judgment.
Anchoring Facet 4: Self-Generated Anchor	The tendency to alter judgments in the direction of self-generated anchors (e.g., being asked to indicate the freezing point of water). The anchoring effect results from insufficient adjustment from such self-generated anchors.
Anchoring Facet 5: Focalism/Focusing Illusion	Focalism is the tendency to base decisions on a truncated search process. As soon as a person decides that they have enough information to justify a decision, their search for information stops. As a result, only a few “focal” pieces of information are considered, which gives them undue weight. The result of this process is referred to as a “focusing illusion.”
Representativeness	The tendency for people to judge the probability or frequency of a hypothesis by considering how much the hypothesis resembles available data.
Representativeness Facet 1: Base Rate Neglect	The tendency to overweight the representativeness of a piece of evidence while ignoring how often (i.e., its base rate) the phenomenon in question occurs in the general population.
Representativeness Facet 2: Sample Size Insensitivity	The tendency to give undue weight to conclusions based on small samples when drawing conclusions about populations.
Representativeness Facet 3: Conjunction Bias	The tendency to draw incorrect conclusions by failing to apply the following rule of probability: It is always less likely that two things will happen simultaneously than that one of the two things will happen, and the other will not.

Bias	Definition
Representativeness Facet 4: Non-Random Sequence Fallacy	The tendency to assume that relatively short sequences of events are more representative of larger random sequences of events than is in fact the case. For example, they often assume that the more often/consistently a random event (such as flipping a coin) happens, the less likely it is that it will happen again.
Projection	The tendency to unconsciously assume that others share one's current emotional states, thoughts and values.
Projection Facet 1: False Consensus	The tendency to overestimate the extent to which others share one's characteristics, attitudes, and beliefs.
Projection Facet 2: Knowledge Projection	The tendency for people to discount the extent to which their knowledge differs from that of others.
Projection Facet 3: Social Projection	The tendency to expect others to be similar to oneself.

The final Phase 2 content domain consists of the paradigms described in the Phase 2 item development section below.

3.3 Cognitive Labs

We conducted two cognitive lab studies during Phase 2. As in Phase 1, the cognitive labs were conducted in two conditions, with one using concurrent think aloud and the other using retrospective think aloud protocols. The first study, which was conducted with 33 ETS employees for 12 BE item prototypes, focused on usability concerns. Do examinees understand the task requirements? Are there any particular task elements or features that facilitate or hinder task performance? We made modifications to the BE items based upon our observations from this initial study. In the second cognitive lab study, which was conducted with 33 ETS employees for 18 BE item prototypes, we continued to examine usability concerns, but in addition, we examined thinking strategies reflected in the verbal protocols.

In general, participants found the task instructions and requirements to be clear and the task designs to be appealing and engaging. We also identified task elements that participants still found to be unclear, distracting, or too demanding. For instance, we asked participants first to make an estimate about an erroneous piece of information in a self-generated anchoring task (i.e., number of hurricanes in a particular area of the world). The participants then made estimates of the price of various consumer products, with the original estimate being their self-generated anchor. However, participants said that they did not understand the question, specifically, having to use their initial estimate (i.e., number of hurricanes) as a reference number for their additional estimates. One participant did not know if that meant if they would pay “2” for a “55 inch TV” (see Figure 34). The directions did not clearly state that the reference number should be thought of in dollars. We changed the directions to ameliorate this incongruity (Figure 34 and Figure 35). We removed the hurricane story, and instead asked participants to list the number of their birth month in its place as a self-generated low-anchor condition. Additionally, we changed the term “reference number” to “ID” and incorporated the sentence, “Think of your ID as a dollar value” to make clear to participants that they were to think of their self-generated ID value as a dollar amount.

Test 11 of 12 Refresh Items ANC-BE-05-Products-Low-Anchor-v2 Select Preview Test Time Elapsed 1:40:39

Directions: Answer the question below.

The island of Kiriatsu is located in the western Pacific, a short plane journey from Australia. The climate is humid and tropical.

How many major typhoons (hurricanes) do you think Kiriatsu receives in an average year? Type the number in the box below.

We will call this number your Reference Number. Recall the six products from the previous screen. If you were on vacation, would you pay your Reference Number for each of these products?

	Yes	No
55" LED TV	<input type="radio"/>	<input type="radio"/>
48" Plasma TV	<input type="radio"/>	<input type="radio"/>
Luxury watch	<input type="radio"/>	<input type="radio"/>
Sports watch	<input type="radio"/>	<input type="radio"/>
Steakhouse dinner for 6	<input type="radio"/>	<input type="radio"/>
Florida beach weekend for 2	<input type="radio"/>	<input type="radio"/>




Figure 34: Self-generated anchoring item in one form of the cognitive labs.

Test of 84 Test Time Elapsed 29:40:45

Refresh Items ANC-BE-05-Products-Low-Anchor-v4 2 Exit Preview

Directions: Answer the questions below. Click **Submit** when you are finished.

Type the month of your birth as a number (1-12).

We will call this number your ID. Think of your ID as a dollar value. Recall the six products from the previous screen. Would you pay your ID for each of these products?

	Yes	No
55 inch LED TV	<input type="radio"/>	<input type="radio"/>
48 inch Plasma TV	<input type="radio"/>	<input type="radio"/>
Luxury watch	<input type="radio"/>	<input type="radio"/>
Sports watch	<input type="radio"/>	<input type="radio"/>
Steakhouse dinner for 6	<input type="radio"/>	<input type="radio"/>
Hawaii weekend for 2	<input type="radio"/>	<input type="radio"/>




Figure 35: Self-generated anchoring item in one form of the Field Trial Tests. The directions have been changed as a result of data from the cognitive labs, regarding the unclear directions. This item now includes “think of your ID as a dollar value,” for greater clarity.

Verbal protocols from concurrent thinking aloud as well as retrospective verbal accounts of response behaviors also indicated that conscious decision making and problem solving strategies varied considerably across tasks and participants. Both qualitative and quantitative analyses showed no indication that participants were performing the BE tasks differently when given concurrent think aloud instructions as compared to being given retrospective questions alone. Interestingly, no participants reported any specific knowledge or awareness of underlying aims of the assessments.

3.4 Item Writing and Review

The basic approach to item writing and review in Phase 2 was identical to the process used in Phase 1. Each round of item generation involved item writing based on prototypes; item review; and, for items administered in a multimedia format, videotaping, editing, and programming.

3.4.1.1 Item Generation

Following the development and evaluation of the task prototypes, we created clones of those prototypes with the objective of substantially increasing the item pool for the ABC-2. We developed a total of 1325 BE and RD items (566 anchoring bias, 229 representativeness bias, and 530 projection bias items) in two rounds.

3.4.1.2 Script Writing and Production of Video-Based SJTs

Working closely with CML, we wrote and videotaped a total of 64 scripted scenarios both for BE and RD items in two rounds of film production held in Louisville, KY, primarily with local professional actors. The same considerations as to script production and logistical constraints described in Phase 1 also applied to Phase 2.

3.4.1.3 Item Review

The Phase 2 item review process essentially recapitulated the Phase 1 item review process, and the same criteria for retention applied.

As in Phase 1, we will describe in some detail the paradigms used to generate item prototypes to operationalize the Phase 2 BE constructs and also include illustrative screenshots of items within each paradigm.

3.4.2 Anchoring Bias

Numerical Priming Paradigm

As operationalized in the ABC-2, the *Numerical Priming paradigm* involves the tendency to base judgments on a single decontextualized piece of numerical information from an external source (either numbers or words reflecting size or magnitude [e.g., large, small]). The information can be explicit or implicit. In the ABC-2, the Numerical Priming paradigm is represented by four tasks. In these tasks, the numerical prime is embedded in an image that accompanies a question regarding how much test-takers would pay for advertised products. Test-takers type in an amount. The closer the amount they are willing to pay is to the embedded numerical prime, the greater the anchoring bias effect. People differ in the extent to which they are influenced by the embedded numerical prime.

Test of 84 Test Time Elapsed 9:14:26

Refresh Items ANC-BE-01-Front-Load-Washer-HA-SR-v2 1 Exit Preview

Directions: Look at the screenshot of a web page advertisement for a washing machine on the right. Then answer the question below. Click **Submit** when you are finished.

How much would you pay (in U.S. dollars) for the washing machine in this advertisement?

dollars

Submit




Figure 36: Numerical Priming (“Front Load Washer”)

Selective Accessibility Paradigm

The *Selective Accessibility paradigm*, as operationalized in the ABC-2, involves the tendency to base judgments on a selective search for information that explains the difference between an externally-provided anchor and what people consider to be plausible values near that anchor. This selective search creates confirmatory hypothesis testing that biases judgments in the direction of the anchor. This facet of anchoring is most likely to be elicited in situations where context and knowledge play a significant role. In the ABC-2, the Selective Accessibility paradigm is represented by two isomorphic tasks, both of which are linking tasks that appear on all ABC-2 test forms: (1) a Candy Jar estimation task, and (2) a Coin Jar Estimation task. In these tasks, individuals estimate a quantity after initially being provided with an anchor derived from earlier estimates.

Directions: Please provide your response to the question below. Click **Submit** when you are finished.



At a charity function you are attending, there is a contest that involves guessing how many candies are in a jar. The jar of candy is pictured on the left. The person whose guess is closest to the correct answer wins and gets to take home the jar of candy.

The other people at your table gave the following guesses:

- 1,275
- 871
- 914
- 850
- 892

What is your best guess?

pieces of candy

Submit



Figure 37: Selective Accessibility (“Candy Jar”)

Comparative Judgment Paradigm

The *Comparative Judgment paradigm*, as operationalized in the ABC-2, involves asking test-takers first to evaluate the cost of a product relative to a high or low anchor value. They then estimate the cost of the product. Anchoring bias is measured as the difference in estimates between the anchored condition and the no-anchor condition. This is an example of a classic anchoring task that yields a robust effect in several different variants (Mochon & Frederick, 2013). A comparative judgment task provides the anchor that influences a subsequent absolute judgment.

Test of Test Time Elapsed 43:37:05

ANC_BE_03_Camera_High_Anchor_v2_F 1 Exit Preview

Directions: Please review the image on the right and then answer the questions.

Does this camera cost more than \$600?

Yes No

What is the exact cost of this camera (in US dollars)?

dollars





Figure 38: Comparative Judgment (“Camera”)

Self-Generated Anchoring Paradigm

The *Self-Generated Anchoring paradigm*, as operationalized in the ABC-2, involves the tendency to alter judgments in the direction of self-generated anchors (e.g., recalling the freezing point of water when asked to indicate the freezing point of ethanol). The anchoring effect results from insufficient adjustment from such self-generated anchors. In one class of self-generated anchoring tasks in the ABC-2, test-takers are required to rate the value of two types of products: (1) relatively cheap products (e.g., chocolates); and (2) luxury products (e.g., big screen TV). For both product types, there is a no-anchor control condition. For the cheap products, a high anchor is self-generated through self-estimation of a number in the 990s. For luxury products, a low-anchor, 1-12, is self-generated by indicating one's birth month.

In another class of self-generated anchoring tasks in the ABC-2, test-takers are required to provide numeric answers to various esoteric knowledge questions (e.g., the freezing point of rubbing alcohol). Test-takers are presumed to generate well-known reference values in attempting to answer the target questions. Anchoring bias is reflected by the tendency to insufficiently adjust one's answer from the self-generated starting point. (The assumed self-generated starting point for the question about the freezing point of rubbing alcohol would be the freezing point of water, or 32°F.) In the ABC-2, the esoteric knowledge questions involve the approximate number of days it takes different planets in the Earth's solar system to revolve around the sun. The assumed self-generated starting point for these questions is 365 days, the number of days it takes the Earth to revolve around the sun.

Test of Test Time Elapsed 43:34:41

ANC_BE_05_Earth_Days_v3_F 1 Exit Preview

Directions: Please read and respond to the questions below.

Approximately how many Earth days does it take *Mars* to revolve around the sun?
 Earth days

Approximately how many Earth days does it take *Jupiter* to revolve around the sun?
 Earth days

Approximately how many Earth days does it take *Saturn* to revolve around the sun?
 Earth days

Approximately how many Earth days does it take *Neptune* to revolve around the sun?
 Earth days


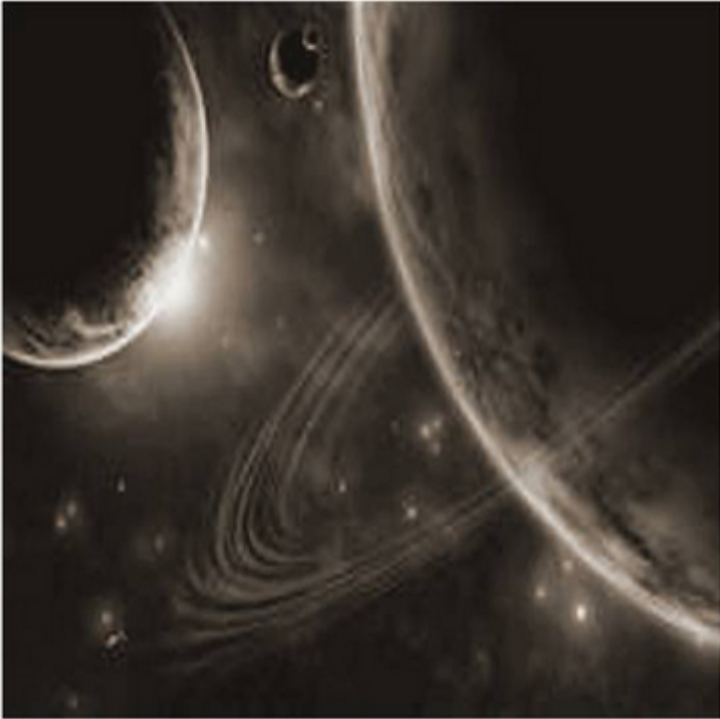


Figure 39: Self-Generated Anchor (“Earth Days”)

Focalism/Focusing Illusion Paradigm

Focalism is the tendency to base decisions on a truncated search process. As soon as a person decides that they have enough information to justify a decision, their search for information stops. As a result, only a few “focal” pieces of information are considered, which gives them undue weight. The result of this process is referred to as a “focusing illusion.” The focalism tasks in the ABC-2 are based on Del Missier, Bonini, and Ranyard’s (2007) account of how information search for consumer choices may be curtailed when there is an acceptable initial choice. They suggest two forms of focalism: (1) test-takers may simply fail to include potential alternatives in the problem representation (referred to as “representational focusing”), or (2) test-takers may represent the alternatives but fail to search for information about them (referred to as “search-related focusing”).

Each task requires the person to search for information about four possible products or purchases, on behalf of another individual. 6-7 relevant items are available per product, but each item sampled costs money, discouraging exhaustive search. There are three conditions:

Control condition: None of the four products is privileged, and each should be equally attractive. Equal sampling of items of information is expected.

Strong focalism: Instructions indicate a substantial advantage for one product.

Weak focalism: Instructions draw attention to one product, but do not provide firm evidence for its superiority.





In focalism conditions, it is expected that participants will sample relatively more information about the focal product, and less about the remainder. The greater the focalism effect, the greater the anchoring bias. In the ABC-2, participants are only presented with strong focalism versions of the information search tasks, and the focalism effect is calculated as the overall difference in mean search activity as compared to control condition estimates of search activity for each task obtained from independent groups of participants.

ANC-BE-07-Focalism-Condo-Search-Focal-Cond1 2 Exit Preview

Directions: Please read the information below and select your responses.

Marta recommends trying Boquete, which has a pleasant climate and a community of expatriate Americans. Although you have \$2,800 to spend, you want to minimize your costs. Each item of information will cost you \$100. Click on each item of information you will order from Marta.

Remaining: \$0

Condo Location	Compliance with local building codes	Estimated cost of utilities	Drinking water quality	Likelihood of flooding	Neighborhood crime levels	Neighborhood shopping amenities	Neighborhood medical amenities
 Boquete	A+ rating from International Home Safety Magazine	\$38 per month for electricity	A+ rating from Panama Canal Conservation League	Situated inland on a mountain	No violent crimes in past decade	Local markets within walking distance	Across from internationally recognized hospital
 Coronado	Recipient of Building Better Places Award	\$42 per month for electricity and garbage pickup	A+ from The Rural Water Board	Surrounded by self-closing flood barrier	Private security firm patrols community	Shopping outlets built into central hub	Situated near university hospital
 Las Tablas	LEED Certified	\$48 per month covers oil	Includes rain harvesting system	Near a flood control channel	Volunteers patrol community	Near dozens of shops and restaurants	Local hospital awarded "Hospital of the Year"
 Bocas del Toro	Highly acclaimed by Panama Development Organization	Gas bill averages \$43 per month	Recognized as a leader in sanitation	Located next to dam	Listed among top-10 safest communities	Ten minutes from shopping mall	Nearby medical facility received top ratings from Panama Accreditation Agency

Submit



Figure 40: Focalism (“Focalism Condo”)

Scoring of Anchoring Bias Tasks

For numerical priming, comparative judgment, and self-generated anchor tasks, bias scores are calculated as a combination of (1) the degree to which one's anchored estimate deviates from the "true score"²⁶; and (2) the degree to which a participant's anchored estimate deviates from the anchor value. In the selective accessibility tasks, the bias is calculated as the difference between a participant's anchored estimate in one task and his/her unanchored estimate in a paired task with similar stimuli. As operationalized in the ABC-2, participants' estimates, true score and anchor values are re-scaled using a \log_{10} transformation. Higher scores for each ANC task indicate less biased responses. The raw ANC scale score is a unit-weighted composite of the ANC task scores.

3.4.3 Representativeness Bias

Base Rate Neglect Paradigm

The *Base Rate Neglect paradigm* involves the tendency to overweight the representativeness of a piece of evidence while ignoring how often the phenomenon in question occurs in the general population (i.e., its base rate). When people ignore the base rate incidence of a phenomenon while appraising the representativeness of a piece of evidence, they are committing base rate neglect (Bar-Hillel, 1980; Goodie & Edmund Fantino, 1995; Stolarz-Fantino, Fantino, & Van Borst, 2006; Tversky & Kahneman, 1982). In the prototypical base rate task, subjects are provided with a base rate summary statistic that they are expected to reference, even when presented with counterintuitive information (Koehler, 1996).

Proper use of Bayes' theorem is essential for providing the correct answer because Bayes' theorem is based on the premise that it is necessary to consider that the hypothesis is true, given the evidence $p(H|E)$, as well as the probability that the evidence would occur regardless of the truth of the hypothesis. In other words, Bayes' theorem states that the probability of a hypothesis, given some evidence, depends on not only the probability of the evidence given the hypothesis, but also the probability that the hypothesis is true in the general population (the base rate), as well as the base rate of the particular sample where the evidence is being observed:

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}, \text{ where}$$

$P(E|H)$ is the probability that the evidence would occur given the hypothesis, $P(H)$ is the probability of the hypothesis occurring, and $P(E)$ the probability of the evidence occurring (Cosmides & Tooby, 1996).

In the ABC-2, the Base Rate Neglect paradigm is represented by two linking tasks (referred to as Tanks and Vet School) and three additional tasks (Cruise, High School Activities, and Medical School). In each of these tasks, test-takers are tempted in different ways to ignore base rate

²⁶ The "true score" is calculated based on the means of unanchored estimates elicited from independent groups of test-takers that are then pooled with unanchored estimates from test takers on similar, highly correlated tasks. In the case of items that asked participants to estimate the number of Earth days it takes a certain planet to revolve around the sun, the true score is the actual correct response.

information. Application of Bayes' theorem produces high scores; that is, responses that are scored as not susceptible to base rate neglect. The tasks are each briefly described in turn.

The Tanks task adapts a classic base rate neglect task from Heuer (1999, pp. 157-158). In this task, test-takers are provided with a scenario in which they are asked to consider two pieces of evidence relevant to determining which of two types of tanks made tracks where self-propelled artillery was fired. One piece of evidence provides base rates of the two types of tanks in the location of interest, and another provides the percent of the time that the expert has been able to correctly identify the two types of tanks given this type of evidence. Test-takers are then asked to estimate the probability that one type of tank made the tracks. This requires knowledge of, and the ability to apply, Bayes' theorem to an intelligence-themed scenario.

In the Vet School task, test-takers are shown a video in which it is made clear that a focal character, who is in school, fits the stereotype of someone interested in pursuing the work of a veterinarian very well. After the video has been shown, test-takers are provided with a table that shows the relative popularity of different graduate programs, including veterinary medicine. Veterinary medicine, however, has the lowest base rate of all the graduate programs. Test-takers are then asked to indicate which of the graduate programs the focal character is most likely studying.

In the Cruise task, test-takers are given a scenario in which a cruise company collected information about people who prefer one type of cruise versus another, and were also given the number of people who took each cruise (i.e., base rate information). They were then presented with information about an individual who recently took a cruise to one of those two places, which is more consistent with the low base rate cruise. Test-takers are asked to estimate the probability that this individual took a cruise to one place rather than another. Base rate neglect is reflected in making a selection that is more consistent with the low-base rate cruise.

The High School Activities task presents the base rates of students that choose sports or orchestra, in addition to distribution counts for the favorite activities of students representing each group. Then the information is provided about a particular student who participates in either sports or orchestra. Test-takers are asked to provide the likelihood (0-100%) that this person participates in sports.

In the Medical School task, test-takers are provided with information about recent medical school graduates who have chosen one of two medical specialties; specifically, the number of physicians in each specialty that joined one of three volunteer organizations: social justice, animal shelters, or nursing homes. Test-takers are then provided with information about a particular recent medical school graduate who works in one of those two areas and asked to indicate the likelihood that this graduate works in one area rather than another.

Refresh Items

REP-BE-01-Vet-School

2

Exit Preview

Directions: The table below shows the relative popularity of various graduate programs at Fowler University. Answer the question below.

Graduate Program	Relative Popularity
Business	48%
Education	23%
Law	12%
Veterinary Medicine	2%
Engineering	5%
Psychology	10%

Diane's son, Jack, is most likely studying:

- Law
- Business
- Veterinary Medicine
- Education
- Psychology
- Engineering



Submit



Figure 41: Base Rate Neglect (“Vet School”)

Sample Size Insensitivity Paradigm

The *Sample Size Insensitivity* paradigm refers to a line of research that has demonstrated people's tendency to give undue weight to conclusions based on small samples when drawing conclusions about populations (e.g., Hamill, Wilson, & Nisbett, 1980; Tversky & Kahneman, 1974). The literature suggests that test-takers may be especially likely to disregard sample size when information about individual cases is available (Obrecht, Chapman, & Gelman, 2009) and presented vividly (Hamill et al., 1980).

In the ABC-2, the Sample Size Insensitivity paradigm is represented by two linking tasks (labeled Astronaut and Marbles) and several additional tasks. Three of these (labeled Bungee Jumping, Brain Gain, and Drug Side Effects) are intended to be isomorphic with the Astronaut task. An additional task representing this paradigm was labeled One-On-One.

In the Astronaut task and related tasks, test-takers are asked to provide ratings of psychological symptoms resulting from different experiences (space flight training simulation, bungee jumping, playing cognitively challenging games, taking prescribed medication). Test-takers are first asked to rate the extent to which four symptoms are likely to develop, thereby establishing baselines. Test-takers are then exposed to a vivid but atypical negative or positive reaction. Bias is suggested by extent of increased (or decreased) ratings.

The Marbles and One-On-One tasks present test-takers with scenarios – one of which (One-On-One) is video-based – that are designed to evaluate whether test-takers understand that the probability of obtaining an unusual or extreme result is greater for small samples because a large sample will be more representative of the general population (e.g., Tversky & Kahneman, 1974).

Directions: Please read the information below and answer the question.

“Brain Gain” is a series of online games developed to improve various mental functions. As yet there is no scientific evidence either supporting or challenging the effectiveness of these games.

Please indicate how likely you think it is that playing the games could help improve each of the following.

	Highly Unlikely	2	3	4	5	Highly Likely
Memory	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Reaction Time	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Concentration	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Problem Solving	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Submit



Figure 42: Sample Size Insensitivity (“Brain Gain”)

Directions: Jeff (left) and Stan (right) are college classmates. For the last month, Jeff has been playing “Brain Gain” games for about an hour a day. His reaction to the games may not be typical. Watch the following video and then answer the question that follows.



Please indicate how likely you think it is that playing the games could improve each of the following.

	Highly Unlikely	2	3	4	5	Highly Likely
Memory	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Reaction Time	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Concentration	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Problem Solving	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Submit



Figure 43: Sample Size Insensitivity (“Brain Gain”)

Conjunction Bias Paradigm

The *Conjunction Bias paradigm* refers to people's tendency to draw incorrect conclusions by failing to apply the following rule of probability: It is always less likely that two things will happen simultaneously than that one of the two things will happen, and the other will not. Some of the conjunction bias tasks in the ABC-2 closely follow the classic "Linda task" (Tversky & Kahneman, 1983). These tasks present characterizations of a focal individual that are consistent with a feature, such as enjoying fishing as a hobby (which we will call Feature Y) and then ask test-takers whether it is more likely that the focal individual has Feature X or has both Feature X and Feature Y. It is more likely to have one feature than both, but Feature Y is presented so as to make it so salient that the majority of people will typically indicate that the conjunction of features X and Y is the correct response. This is true of the conjunction bias tasks labeled Fishing, Tiring Weekend, and Grammarian.

Another type of conjunction bias task in the ABC-2 can be seen in the tasks labeled Bomb Threat, IRS, Security Guard, and Airport Security, and are derived from Wedell (2011). These problems measure the ability to discriminate between the instances where the conjunction rule of probability does and does not apply. The conjunction rule applies in prediction problems, which ask participants to judge the probability that an event will occur. In these situations, it is always less likely that a conjunction of events will occur than the probability that a single event will occur. The conjunction rule does not apply in diagnosis problems, which ask participants to identify which option provides the most support for a given hypothesis. In diagnostic problems Bayes' theorem applies, and the conjunction of events increases the probability of a hypothesis. Therefore, the single event is the correct answer for prediction problems, and the conjunction is the correct answer for diagnosis problems.

REP_BE_04_Bomb_Threat_DP_v2_F

1

Exit Preview

Directions: Please read the statements below and answer the questions.

British officials are notified that a bomb is likely to be set off in London tomorrow. They suspect the threat came from someone affiliated with a known terrorist group. The officials maintain a watch list of people whom they believe have been affiliated with the terrorist group in the past. Some of the individuals on this list currently self-identify as members of the terrorist group, while some do not.

The officials have identified suspects based on their investigations. Three of the suspects are listed below, along with the relevant information that is known about them. Based on this information, which of these individuals would you consider most likely to be part of the intended attack?

- Person 1, who does not currently self-identify as a member of the known terrorist group.
- Person 2, who does not currently self-identify as a member of the known terrorist group but who is still sympathetic to the group's cause.
- Person 3, who does not currently self-identify as a member of the known terrorist group and who is no longer sympathetic to the group's cause.

Which of the following scenarios is most probable?

- Person 1, who does not currently self-identify as a member of the known terrorist group.
- Person 2, who does not currently self-identify as a member of the known terrorist group but who is still sympathetic to the group's cause.
- Person 3, who does not currently self-identify as a member of the known terrorist group and who is no longer sympathetic to the group's cause.

Submit



Figure 44: Conjunction Bias (“Bomb Threat”)

Non-Random Sequence Fallacy Paradigm

The *Non-Random Sequence Fallacy paradigm* refers to people's tendency to assume that relatively short sequences of events are more representative of larger random sequences of events than is in fact the case. For example, they often assume that the more often/consistently a random event (such as flipping a coin) happens, the less likely it is that it will happen again. One example of this is known as the "gambler's fallacy." If a person predicts that a fair coin toss will come up "tails" because the seven prior coin flips have yielded all "heads," he or she would be committing this fallacy. Another example of the non-random sequence fallacy is the so-called "hot hand" effect. In a sense, the "hot hand" effect is "the opposite side of the same coin" from the gambler's fallacy. The "hot hand" effect occurs, for instance, when a person predicts that a fair coin toss will come up "heads" because the seven prior coin flips of all yielded "heads."

In the ABC-2, the Non-Random Sequence Fallacy paradigm is operationalized by tasks labeled Girls BBall, Jury Duty, Dice, and Retrospective Colored Marbles.

The gambler's fallacy is operationalized in the form of a "retrospective gambler's fallacy" (Oppenheimer & Monin, 2009). All of the BE gambler's fallacy items in the ABC-2 are based on this work. These types of items are designed to avoid overly-transparent gambler's fallacy items based, for example, on flipping coins and predicting what will happen on the next flip. Retrospective gambler's fallacy has the virtue of evaluating susceptibility to this fallacy somewhat more obliquely.

Oppenheimer and Monin (2009) investigated whether the rarity of an independent, chance observation influenced beliefs about what occurred before that event. Participants imagined that they saw a man rolling dice in a casino. In one condition, participants imagined witnessing three dice being rolled and all came up 6's. In a second condition two came up 6's and one came up 3. In a third condition, two dice were rolled and both came up 6's. All participants then estimated, in an open-ended format, how many times the man had rolled the dice before they entered the room to watch him. Participants estimated that the man rolled dice more times when they had seen him roll three 6's than when they had seen him roll two 6's or two 6's and a 3.

The Dice task was closely modeled after the Oppenheimer and Monin paradigm. The logical fallacy in these retrospective items is that a longer history of trials precedes an apparently unlikely event. Jury Duty and Retrospective Colored Marbles represent variations on this theme of retrospective gambler's fallacy.

The "hot hand" effect is represented in the ABC-2 by the Girls BBall task. More specifically, the Girls BBall task is based on the "hot hand" in basketball (Gilovich, Vallone, & Tversky, 1985). Gilovich et al. (1985) surveyed participants' beliefs about the likelihood of a basketball player making a shot based on the success of his previous shots, and found that 68% of participants thought a player was more likely to make a shot if he made his last two or three shots. The Girls BBall task is directly based on this aspect of the Non-Random Sequence Fallacy paradigm.

Test of 84 Test Time Elapsed 9:33:37

Refresh Items REP-BE-05-Retrospective-Colored-Marbles-v3 1 Exit Preview

Directions: Read the description below and then answer the question. Click **Submit** when you are finished.

- Walking around with your young niece at her school fair, you pass a booth where people are playing a game that involves drawing four marbles out of a large jar while their eyes are closed.
- There are four jars from which players may pick, and the marbles in each jar are evenly divided between six colors: blue, red, yellow, green, orange, and purple. Each person playing the game is allowed multiple turns, and each turn costs one ticket. After taking a turn, players return their marbles to the jar. The object of the game is to pick four marbles of the same color; players who succeed in doing so win a prize.
- Three children are playing the game when you arrive at the booth. As you watch, you see a girl pick four orange marbles out of her jar. Then you watch another girl pick two green marbles plus one red and one orange one out of her jar. Finally, you see a boy pick three red marbles and one purple marble out of the third jar.

Your niece would like to play, but she only has 12 tickets left and she's not sure she wants to use them all on one game. She's wondering how many turns it will take her to pick four marbles of the same color and asks you how many turns you think the three children you've been watching have already taken.

What is your best guess of how many turns each of the three children took before this turn?

	2	4	6	8	10	12
1st Girl (4 orange marbles)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2nd Girl (2 green, 1 red, 1 orange marble)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Boy (3 red, 1 purple marble)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>




Figure 45: Non-Random Sequence Fallacy (“Retrospective Colored Marbles”)

Scoring REP

For base-rate neglect tasks, bias is scored as the tendency to provide likelihood estimates that deviate from the base rate, and in the case of the “Tanks” task, the deviation from the correct Bayesian estimate. Every scored rating is made on a Likert-type scale, with varying numbers of response options and appropriate anchors that have been thoroughly pilot tested. Unit-weighted composites of these ratings or, in some cases, of difference-score ratings, are used to operationalize REP. All scored ratings are made using a multiple-choice, selected-response format. These ratings are dichotomized such that REP is coded as 0 and non-REP is coded as 1. Each REP task score is computed as the sum of these dichotomized ratings. The raw REP BE scale score is a combination of the REP task ratings.

3.4.4 Projection Bias

False Consensus Paradigm

The *False Consensus paradigm* refers to people's tendency to overestimate the degree to which other people share their characteristics, beliefs, or attitudes (Krueger & Clement, 1994). In a series of studies involving college students, Krueger and Clement found good support for the false consensus effect. In one such study, for example, participants indicated agreement or disagreement with each of 40 items from the Minnesota Multiphasic Personality Inventory (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), rated on a scale of 1-9 how socially desirable it is to agree with each item, and later estimated the percentages between 0 and 100 that best reflected their belief about the proportion of people in the population who would agree with each statement (referred to as "consensus estimates"). Evidence of projection was found both in (a) differences in mean consensus estimates computed across items, between individuals who did and did not endorse those items; and (b) the correlation between consensus estimates and endorsement.

In the ABC-2, test-takers were provided with information about a proposal that requires a vote and arguments for and against the proposal. Test-takers are asked to: (1) estimate the percentage of a sample that would be for the proposal, (2) estimate the percentage of the sample that would be against the proposal, and (3) choose an option themselves. False consensus is demonstrated when test-takers who select a particular choice estimate that choice to be more prevalent than the alternative among typical adult peers.²⁷

²⁷ There is the possibility that people who estimate that others are more likely to agree with their choice may be forming their own opinion based on what they believe to be popular rather than assuming that their choice is the most popular. We did not see any indication of this from the cognitive lab verbal protocols and did not investigate this possibility further in pre-pilot research studies.

Test of 84 Test Time Elapsed 9:40:05

Refresh Items PRO-BE-01-Tipping-BP-PD-v2 1 Exit Preview

Directions: Read the information and answer the questions below. Click **Submit** when you are finished.

Currently, most restaurants utilize a voluntary tipping system in which service staff earn an hourly base wage in addition to gratuities received from customers. A proposal has been made to abolish the practice of voluntary tipping and replace it with a mandatory service fee on all restaurant bills.

- Proponents of this movement argue that replacing the tipping system with service fees will help owners provide their employees with better wages and mitigate wage disparities between servers and kitchen staff.
- Opponents of this movement argue that gratuities motivate workers to provide high-quality service and supplement their small paychecks; replacing the tipping system would lead to an increase in menu prices.

Which option would you choose?

Support the mandatory service fee
 Oppose the mandatory service fee

What percentage of adult Americans (0%-100%) would:

Support the mandatory service fee?	<input type="text"/> %
Oppose the mandatory service fee?	<input type="text"/> %




Figure 46: False Consensus (“Tipping”)

Knowledge Projection Paradigm

The *Knowledge Projection paradigm* refers to people's tendency to discount the extent to which their knowledge differs from that of others. "They base assumptions about what others know on what they themselves know, or think they know" (Nickerson, 1999, p. 747). In the ABC-2, knowledge projection items ask test-takers to answer a general knowledge trivia question, and then estimate their level of confidence in their answers. Test-takers are then asked to estimate the percentage of American adults that might know the answer.

Test of 84 Test Time Elapsed 9:11:52

Refresh Items PRO-BE-02-KP-Himalayas 1 Exit Preview

Directions: Answer the questions below. Click **Submit** when you are finished.

What is the name of the mountain range in which Mount Everest is located?

Andes Rocky Alps Pyrenees Himalayas

How confident are you in your answer? %

What percentage of American adults would know this answer? %

Sarah is one of the people playing trivia at the restaurant. Sarah is a 39-year-old major in the military. Sarah spends her free time at the gym and volunteering for a disaster relief organization. What is the probability that Sarah would beat the average American adult on this question?

%




Figure 47: Knowledge Projection (“Himalayas”)

Social Projection Paradigm

The *Social Projection paradigm* refers to people's tendency to expect others to be similar to themselves. We divided this paradigm into two facets: (1) affective projection, and (2) personality projection.

Affective projection involves projecting one's emotions onto others such that there is an expectation that others feel the same way, whether they do or not. *Personality projection* involves projecting one's personality traits onto others such that there is an expectation that others will think, feel, and behave in a manner consistent with one's own personality traits.

Affective projection is measured by adapting two tests designed to measure emotional intelligence developed by MacCann and Roberts (2008): the Situational Test of Emotional Understanding (STEU) and the Situational Test of Emotional Management (STEM). We also developed video-based situational judgment tests (SJTs) following a similar paradigm. We adapted the STEM and the STEU because, together, they provide a set of situations rich in affective implications. These situations were adapted to measure affective projection by first asking people what a focal person would do in a given situation and then asking them what they would do if they were in that person's position. The items' response scales cover a comprehensive theoretically-based and empirically-supported array of emotional responses using the bipolarity of affect model established by Barrett and Russell (1998).

In addition to the situations provided by the STEM and the STEU, we developed situations, captured them on video, and asked people to put themselves directly into these situations. Then, similar to the other affective projection items, but with a twist, we asked them how they would feel in the situation given a specific role, and then asked them how they think the average American adult would feel after being put in the same position. The video-based items are labeled Eye Doctor, Keys Female, and Thesis.

Directions: Please read and respond to the question below.



How do you think Sam (right) feels after this interaction with his professor?

	←————→						
Alert	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Fatigued
Relaxed	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Nervous
Calm	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Tense
Elated	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Depressed
Contented	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Upset
Serene	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Stressed
Happy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Sad
Excited	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Bored

[Submit](#)




Figure 48: Social Projection (“Thesis”)

Test of Test Time Elapsed 43:44:40

PRO_BE_06_Thesis_F 3 Exit Preview

Directions: Please read and respond to the question below.



How do you think you would feel if you were in Sam's position?

	←————→						
Alert	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Fatigued
Relaxed	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Nervous
Calm	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Tense
Elated	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Depressed
Contented	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Upset
Serene	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Stressed
Happy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Sad
Excited	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Bored

Submit




Figure 49: Social Projection (“Thesis”)

Personality Attribute Projection

Personality Attribute Projection was measured with four different tasks that appear on each form. These are labeled High Tech, Photojournalism, Design, and Grant Writing. Each of these tasks are repurposed BFI-44 items. The BFI-44 is an abbreviated, 44 item version of the Big Five Inventory, a self-report assessment that has subjects indicate on a scale of 1 to 5 how strongly a statement characterizes their personality (John, Donahue, & Kentle, 1991). The Big Five Model is a personality taxonomy that breaks personality down into five bipolar dimensions that account for the majority of individual differences in personality.

In this Personality Attribute Projection task, test-takers are administered text-based SJTs in which they are asked to select which 4 out of a pool of 11 candidates they would select as co-workers and which candidates the average adult American would select based upon descriptions of behaviors linked to each of the Big-Five personality traits.

Test-takers demonstrate projection bias by the (1) degree to which their predictions of which candidates the average American would select deviate from the “true score” (i.e., proportion of individual candidate selections from the study sample); and (2) the degree to which an individual projects his/her candidate selections onto others (i.e., degree of correspondence between “self” and “other” candidate selections).

Refresh Items

PRO-BE-06-Design_BFI_Pt4

1

Exit Preview

Directions: Please read the information below. Then, select your answers and click **Submit** when you are finished.

Your architectural firm has won a contract to design a new riverfront for a U.S. city. You are tasked to recruit four people who will work with you to create the design plan. After reviewing the applications of potential candidates, you have narrowed the field to the following 11 applicants, each of whom is highly qualified. Each of the candidates also took a personality test. Statements taken from the test were assigned to the candidates and are listed below.

Which four people would you choose to hire?

- Ramiro has few artistic interests
- Jamie is ingenious, a deep thinker
- Janet worries a lot
- Greg can be cold and aloof
- Jean does a thorough job
- Jerome generates a lot of enthusiasm
- Crystal is easily distracted
- Evita is outgoing, sociable
- Anisha can be somewhat careless
- Effie is considerate and kind to almost everyone
- Shantell is emotionally stable, not easily upset



Figure 50: Social Projection (“Design”)

Test of 84 Test Time Elapsed 10:03:52

Refresh Items PRO-BE-06-Design_BFI_Pt4 2 Exit Preview

Directions: Please read the information below. Then, select your answers and click **Submit** when you are finished.

Your architectural firm has won a contract to design a new riverfront for a U.S. city. You are tasked to recruit four people who will work with you to create the design plan. After reviewing the applications of potential candidates, you have narrowed the field to the following 11 applicants, each of whom is highly qualified. Each of the candidates also took a personality test. Statements taken from the test were assigned to the candidates and are listed below.

Which four people do you think the average American would choose to hire?

- Ramiro has few artistic interests
- Jamie is ingenious, a deep thinker
- Janet worries a lot
- Greg can be cold and aloof
- Jean does a thorough job
- Jerome generates a lot of enthusiasm
- Crystal is easily distracted
- Evita is outgoing, sociable
- Anisha can be somewhat careless
- Effie is considerate and kind to almost everyone
- Shantell is emotionally stable, not easily upset




Figure 51: Social Projection (“Design”)

3.5 Phase 2 Pre-Pilot Studies

As with the ABC-1, the constructs targeted for measurement in the ABC-2 were not well understood from an individual differences perspective. As such, we again conducted a considerable amount of pilot test research prior to the Phase 2 Field Test. Because the pace of Phase 2 was even faster than the pace of Phase 1, it was necessary to conduct the research in parallel or in a cascading fashion to an even greater extent than in Phase 1, although we conducted the research iteratively and sequentially to the extent possible so that we could build on knowledge as it was acquired.

The Phase 2 pre-pilot research explored many of the same questions as the Phase 1 pre-pilot research. However, there were also some differences. For example, some questions investigated in Phase 1 did not require further investigation in Phase 2. There were, however, various questions that we explored in the Phase 2 pre-pilot research that had not been investigated in Phase 1 because items specific to the Phase 2 biases used measurement methods specific to those biases. For example, Phase 2 pre-pilot research investigated: (1) the effects of manipulating experimental design (within- vs. between-subjects²⁸) on the presence and magnitude of anchoring bias; (2) the effects of manipulating response format (percent vs. frequency) on likelihood estimates in base rate neglect problems; and (3) the psychometric properties of alternate scoring approaches for knowledge and social projection.

As in Phase 1, the Phase 2 pre-pilot test research encompassed an enormous amount of work. Therefore, we again focus only on key questions and results. These are summarized in Tables 22 - 25. These tables correspond to ANC, REP, PRO, and RD, respectively. The tables are intended to stand alone, and we do not discuss them beyond the content in the tables themselves. We discuss the Field Test and Pretest Sensitization studies in detail subsequent to the summary of the pre-pilot test research.

²⁸ Within-subjects and between-subjects refer to distinct experimental designs. A within-subject design focuses on changes in individuals on an item, test, or composite; or in their performance on an experimental task across multiple trials. A between-subjects experimental design focuses on change at the group level on the same types of variables.

Table 22: Summary of Anchoring Bias (ANC) Pre-Pilot Research Studies

ANC Study Iteration	Brief Study Description and Key Questions	Key Results
Round 1	<p>Cognitive Laboratory studies with 33 ETS employees</p> <ul style="list-style-type: none"> • Do examinees understand the task requirements? • Are there any particular task elements or features that facilitate or hinder task performance? • What thinking strategies do examinees use to perform BE tasks? <p>Pre-pilot testing with AMT workers ($n = 766$) in which we examined the following task variable manipulations:</p> <ul style="list-style-type: none"> • High, Low, and No Anchor task formats • Item and anchor presentation order counterbalanced between subjects • Within- and between-subjects design used to examine item and anchor type variations for paired items 	<ul style="list-style-type: none"> • With the possible exception of the “Washing Machine” cost estimation task, there was no indication of ANC due to irrelevant, experimenter-provided anchors. • Participants in cognitive laboratory studies indicated that they noticed the irrelevant anchor value • Moderate to large effects of high and low anchors in target estimates for Selective Accessibility paradigm tasks relative to estimates made with no anchors • Pattern of results similar for between- and within-subjects designs • Anchoring effects differed with respect to the magnitude of the anchor (high vs. low) and item-pair combination • Higher confidence ratings observed in the anchoring conditions relative to the no-anchor conditions for the candy/coins estimation task (possibly another index of anchoring bias?) • Small to moderate effects of high anchors in target estimates relative to estimates made with no anchors in Comparative Judgment and Scale Distortion tasks • Pattern of results similar for between- and within-subjects designs • Anchoring effects differed with respect to the magnitude of the anchor (high vs. low) and item-pair combination • Self-Generated Anchoring temperature estimation questions demonstrated predicted anchoring effects, but effect sizes varied substantially across items • Presenting reference questions before or after the target questions did not matter • Sales Forecasting task showed limited evidence of ANC

ANC Study Iteration	Brief Study Description and Key Questions	Key Results
Round 2	<p>Pre-pilot testing with AMT workers ($n = 755$) in which we examined the following task variable manipulations:</p> <ul style="list-style-type: none"> • High, Low, and No Anchor task formats • Varied semantic relevance of numerical primes in Washing Machine cost estimation task • Manipulated credibility (High vs. Low) and plausibility (Extreme vs. Moderate distance from unanchored mean estimate) of anchors in paired “selective accessibility” items • Item and anchor presentation order counterbalanced between subjects • Also investigated cross-task correlations and correlations with other individual-difference variables and background/demographic variables 	<ul style="list-style-type: none"> • Small ANC effects observed in Washing Machine task using semantically unrelated anchors • Large ANC effects observed in Washing Machine Task using semantically-relevant anchors • Despite modifications to increase the perceptual salience of irrelevant anchors, most numerical priming tasks showed small or negligible effects • Moderate to large effects of high and low anchors in Selective Accessibility task estimates relative to estimates made with no anchors • Anchoring effects differed with respect to the magnitude of the anchor (high vs. low) and item-pair combination • Extremity (plausibility) of the anchor value had the largest effect on target estimates, while credibility manipulations had little influence • Replicated Round 1 findings for Comparative Judgment and Self-Generated Anchoring tasks

ANC Study Iteration	Brief Study Description and Key Questions	Key Results
Round 3	<p>Pre-pilot testing with AMT workers ($n = 459$) in which we examined the following task variable manipulations:</p> <ul style="list-style-type: none"> • High, Low, and No Anchor task formats • Varied semantic relevance of numerical primes in Washing Machine cost estimation task • Manipulated credibility (High vs. Low) and plausibility (Extreme vs. Moderate distance from unanchored mean estimate) of anchors in paired “selective accessibility” items • Item and anchor presentation order counterbalanced between subjects • Manipulated focal target (Self vs. Other) between-subjects in focalism items • Also investigated cross-task correlations and correlations with other individual-difference variables and background/demographic variables 	<ul style="list-style-type: none"> • In focalism tasks, strongest focalism effects observed for correlations between focal character estimates and the % of average Americans who would know the answer • Most individuals are showing focalism to some degree • Low correlations between ANC paradigms • No practically-significant correlations observed between ANC paradigms and BFI personality, cognitive ability, and demographic variables

ANC Study Iteration	Brief Study Description and Key Questions	Key Results
Round 4	<p>Pre-pilot testing with University of Central Florida ($n = 158$) students, ETS essay raters ($n = 84$), and AMT workers ($n = 481$)</p> <ul style="list-style-type: none"> • Sought to replicate findings from previous rounds • Manipulated webpage screenshot format in Numerical Priming tasks • Investigated ANC elicitation in Selective Accessibility tasks using conditional anchoring paradigm • Manipulated number of anchors, size and variability of multiplier used to generate anchor values • Examined item performance characteristics for information search focalism tasks • Evaluated ratio vs. difference score approaches to measuring ANC • Evaluated sensitivity to bias mitigation training using an instructional video 	<ul style="list-style-type: none"> • Small to moderate ANC effects replicated for semantically relevant anchors in Washing Machine task • Failed to replicate ANC effects in Washing Machine task using irrelevant anchors and anchors presented in the sidebar of a webpage • Conditional anchoring versions of Selective Accessibility tasks demonstrated robust ANC effects • Larger ANC effects observed with multiple anchors and larger factor multipliers • Information Search Focalism items demonstrated moderate to large focalism effects (i.e., restricted information search in conditions where a product preference is strongly implied in the text scenario descriptions) • Limited evidence of ANC mitigation after viewing instructional video about the Phase 2 biases

Table 23: Summary of Representativeness Bias (REP) Pre-Pilot Research Studies

REP Study Iteration	Brief Study Description and Key Questions	Key Results
Round 1	<p>Cognitive Laboratory studies with 33 ETS employees</p> <ul style="list-style-type: none"> • Do examinees understand the task requirements? • Are there any particular task elements or features that facilitate or hinder task performance? • What thinking strategies do examinees use to perform BE tasks? <p>Pre-pilot testing with AMT workers ($n = 403$) in which we examined the following task variable manipulations:</p> <ul style="list-style-type: none"> • Base Rate Information (direct, indirect, self-generated) • Response format (probability, frequency) • Task presentation order 	<ul style="list-style-type: none"> • Base rate neglect (BRN) tasks consistently demonstrate base rate neglect • BRN tasks also showed evidence that many participants responded with the test accuracy, but few participants appeared to combine base rate and test accuracy to provide the correct “Bayesian” answer <ul style="list-style-type: none"> • Similar results observed using frequency and percentage-based response formats • BRN tasks with non-diagnostic cases showed overall BRN • “DNA Test” and “House Party” tasks did not show evidence of BRN, however • “Astronaut Simulation” and “traditional” Kahneman & Tversky-type sample size insensitivity problems showed evidence of insensitivity to sample consistent with prior research • Regression to the mean problems did not show convincing evidence of the bias • Majority of participants committed the conjunction fallacy for most conjunction fallacy items • Gambler’s Fallacy video-based SJTs did not show bias elicitation—the vast majority of respondents provided the correct answers

REP Study Iteration	Brief Study Description and Key Questions	Key Results
Round 2	<p>Pre-pilot testing with AMT workers ($n = 524$) in which we examined the following task variable manipulations:</p> <ul style="list-style-type: none"> • Base-Rate and Test Accuracy information • Stereotypical descriptions of focal characters in problem scenarios • Predictive vs. Diagnostic framing of conjunction bias problems 	<ul style="list-style-type: none"> • Replicated Round 1 findings of BRN as well as the tendency to focus on test accuracy • 3 out of 4 BRN problem with non-diagnostic cases elicited BRN • 2 out of 4 BRN video-based SJTs showed substantial BRN • Replicated Round 1 findings for Sample-Size insensitivity and Conjunction Fallacy tasks • Participants preferred the conjunction response for both prediction and diagnosis items, as expected
Round 3	<p>Pre-pilot testing with University of Central Florida ($n = 158$) students, ETS essay raters ($n = 84$), and AMT workers ($n = 481$)</p> <ul style="list-style-type: none"> • Sought to replicate findings from previous rounds • Investigated item performance characteristics for task prototypes and clones • Investigated the “hot hand” and “gambler’s fallacy” using new text and video-based SJTs • Investigated “Retrospective Gambler’s Fallacy” paradigm • Also investigated cross-task correlations and correlations with other individual-difference variables and background/demographic variables • Evaluated sensitivity to bias mitigation training using an instructional video 	<ul style="list-style-type: none"> • Replicated Round 2 findings for BRN, Sample-size insensitivity, and conjunction fallacy problems • Regression to the mean items performed as intended • Most of the “hot hand” and “gambler’s fallacy” items demonstrated adequate difficulty with individual differences • Retrospective Gambler’s fallacy tasks showed evidence of bias elicitation • Limited evidence of REP mitigation after viewing instructional video about the Phase 2 biases • Low correlations across REP items and paradigms • Small to moderate correlations between REP items and cognitive ability measures

Table 24: Summary of Projection Bias (PRO) Pre-Pilot Research Studies

PRO Study Iteration	Brief Study Description and Key Questions	Key Results
Round 1	<p>Pre-pilot testing with AMT workers ($n = 402$):</p> <ul style="list-style-type: none"> • Examined item performance characteristics for false consensus effect, knowledge projection, and social projection task prototypes • Investigated correlations between BFI-44 factor scores and responses to social projection “Co-worker Selection” items • Investigated correlations between self and other ratings as measures of social projection 	<ul style="list-style-type: none"> • Evidence of false consensus effect in group-level item scores • Knowledge projection items showed poor calibration in individuals’ estimates of others’ knowledge • Observed “Like Me” response pattern in “Co-worker Selection” social projection tasks, although characters with negative personality traits were seldom selected • Substantial variability in correlations between self and other ratings in social projection items
Round 2	<p>Cognitive Laboratory studies with 33 ETS employees</p> <ul style="list-style-type: none"> • Do examinees understand the task requirements? • Are there any particular task elements or features that facilitate or hinder task performance? • What thinking strategies do examinees use to perform BE tasks? <p>Pre-pilot testing with AMT workers ($n = 335$):</p> <ul style="list-style-type: none"> • Manipulated order of presentation for self and other ratings in false consensus effect and social projection items • Varied difficulty of knowledge questions used to assess knowledge projection • Examined item performance characteristics for revised false consensus effect, knowledge projection, and social projection task prototypes and clones • Compared Likert vs. Forced-choice response format for social projection items 	<ul style="list-style-type: none"> • Verbal protocols from cognitive laboratory studies revealed perceptions that test takers had of the “Average American” who was the subject of many PRO items • Verbal protocols also revealed evidence in some cases of knowledge and social projection • Knowledge projection items varied in difficulty, but PRO indices showed good internal consistency reliability (Alphas in the low .80s) • Replicated “Like Me” response pattern observed in Round 1 for social projection tasks

PRO Study Iteration	Brief Study Description and Key Questions	Key Results
Round 3	<p>Pre-pilot testing with University of Central Florida ($n = 158$) students, ETS essay raters ($n = 84$), and AMT workers ($n = 931$)</p> <ul style="list-style-type: none"> • Varied order of Self vs. Other ratings in false consensus effect and emotion projection tasks • Manipulated format and number of response options in false consensus effect and personality attribute projection tasks • Varied difficulty, domain, and conceptual veracity of knowledge questions used to assess knowledge projection • Compared ratio vs. difference score approaches to measuring knowledge and social projection • Evaluated sensitivity to bias mitigation training using an instructional video 	<ul style="list-style-type: none"> • Most False Consensus Effect (FCE) tasks elicited the bias in group-level scores and showed substantial individual differences using individual-level scoring approaches • FCE extended to third “Undecided” response option—that is, people who were “Undecided” on a given issue tended to estimate that others would also be “Undecided” • Difference score approaches for measure knowledge and social projection yielded higher alpha coefficients • Replicated findings of “Like Me” response patterns from previous rounds • Limited evidence of PRO mitigation after viewing instructional video about the Phase 2 biases • Low to moderately high correlations within PRO paradigms, but low correlations across PRO paradigms

Table 25: Summary of Recognition and Discrimination (RD) Pre-Pilot Research Studies

RD Study Iteration	Brief Study Description and Key Questions	Key Results
Study 1	<ul style="list-style-type: none"> • AMT, $n = 295$ • Administered 43 items covering the ABC-2 content domain • Test-takers were given a carefully developed one-page description of each bias to read prior to taking the RD test • Investigated psychometric properties of the items by computing basic descriptive statistics, conducting internal consistency reliability analyses, and conducting a principal components analysis 	<ul style="list-style-type: none"> • Retained 27 items covering knowledge of ANC (9 items), REP (10 items), and PRO (8 items) • 16 items were dropped due to excessive difficulty, lack of correlation with other items, tendency to decrease alpha, and low loading on the 1st unrotated principal component • Alpha coefficient was .83 • No large correlations with demographic variables
Study 2	<ul style="list-style-type: none"> • AMT, $n = 499$ • Administered 31 new items in addition to the 27 items retained from Study 1 • The 58 items were distributed across 4 forms (14 – 16 items per form) and subjected to the same data-analytic criteria as were used in Study 1 • Test-takers were also given a carefully developed one-page description of each bias to read prior to taking the RD test 	<ul style="list-style-type: none"> • RD item pool was expanded to 36 items • Results yielded preliminary scales ranging from 7 to 11 items, with $\alpha = .69$ to $.76$ • Principal components analysis revealed that each of the scales was relatively unidimensional

3.6 ABC-2 Field Test

As with Phase 1, the purpose of the ABC-2 Field Test was to administer the entire set of tasks/items to a large and representative group of test-takers to evaluate their psychometric properties and validity. As such, we anticipated making some changes to the test forms, but also expected that, as a result of the extensive pre-pilot testing described above, the items would generally perform well. Two other critically important purposes of the field test were (1) to provide data necessary for creation of equivalent forms for use in the Phase 2 IV&V, and (2) to evaluate the sensitivity of the ABC-2 to a surrogate bias mitigation intervention provided by IARPA (Intelligence Advanced Research Projects Activity, 2013).

3.6.1 Method

3.6.1.1 Participants

The ABC-2 Field Test was administered to a total of 2,299 test-takers (766 – 767 per form), all of whom were recruited through Amazon Mechanical Turk. All test-takers were based in the U.S. and paid \$12 for completing the test. Total screened sample consisted of 2,012 participants. The sample of test-takers averaged 34 years of age ($SD = 11$ years); was 46% male; and approximately 81% Caucasian, 8% African-American, 5% Asian, and 5% multi-racial. Eight percent of the sample was Hispanic or Latino. Approximately 85% of the sample attended college, and the sample was relatively high-achieving, with over 70% reporting cumulative college GPAs between 3.0 and 4.0, and 35% reporting GPAs over 3.5. The majority of the sample reported having taken at least one or two psychology courses, though the vast majority (79%) had not taken more than four psychology courses. About two thirds of the sample reported being currently employed at the time of testing.²⁹

3.6.1.2 Study Design and Procedure

Table 26 lists the types of items and paradigms represented in each of the three forms administered online in the ABC-2 Field Trial. All BE items preceded RD items. The sequence of BE items varied such that the items representing each bias facet were presented in different sequences and combinations across forms. After completing the BE tests, participants read text descriptions of the Phase 2 biases prior to taking the RD test. Two attention check items were included in the BE test sequence, and two attention check items were included in the RD test sequence. Last, participants completed the BFI-44 and a demographics survey. The total median time for completing all parts of the field test was 69.1, 70.1, and 70.6 minutes for Forms 1, 2, and 3, respectively.

²⁹ Ideally, the participant population would be representative of the analyst population that will be receiving training from the Sirius video games. As a comparison, in the Sirius Phase 2 IV&V the participant demographics were as follows. Students: average age: 20; 48% male; 48% Caucasian, 16% Hispanic, 11% African-American, 11% Asian American, 14% other; most frequent major: Health Science. Analysts: average age: 32; 67% male; 81% Caucasian, 8% Hispanic, 5% African-American, 3% Asian-American, 3% other.

Table 26: Number of Items and Paradigms Represented in ABC-2 Field Trial Study Forms.

Scale	Facet	Form 1	Form 2	Form 3
Anchoring Bias	Numerical Priming	4	4	4
	Selective Accessibility	3	3	3
	Comparative Judgment	7	7	7
	Scale Distortion	2	2	2
	Self-Generated Anchor	6	6	4
	Focalism	2	2	2
Representativeness Bias	Base Rate Neglect	4	4	4
	Sample Size Insensitivity	10	10	10
	Regression to the Mean	1	1	1
	Conjunction Bias	4	4	4
	Non-Random Sequence Fallacy	3	3	3
Projection Bias	False Consensus Effect	4	4	4
	Knowledge Projection	7	7	7
	Social Projection – Affective Attribution	8	8	8
	Social Projection – Personality Attribution	4	4	4
Recognition and Discrimination	Anchoring Bias	4	4	4
	Representativeness Bias	5	5	5
	Projection Bias	3	3	3
Attention Check Items	BE Item Type	2	2	2
	RD Item Type	2	2	2
BFI-44		44	44	44
Demographics		12	12	12

3.6.1.3 ABC-2 Scale Development

As part of the Field Test, we created scales for each Phase 2 BE bias construct, together with an RD scale. In doing so, we were guided by the following goals: (1) measure each construct as broadly as possible to ensure maximum content coverage; (2) maximize scale reliability; (3) create a compelling validity argument for each scale, with special emphasis on ability to detect change in scale-scores before and after bias mitigation interventions; and (4) measure as efficiently as possible, and in no event exceed 60 minutes for any test form.

To maximize content coverage, we had developed items for each facet in the measurable content domain. In developing scales, however, some facets were not equally represented because item analyses necessitated dropping a different number of items for the various facets of each scale. In determining items to retain, we conducted several statistical analyses:

- (1) We computed means, standard deviations, and other relevant statistics to identify items that were too easy, too difficult, or had anomalous frequency distributions.
- (2) We computed internal consistency reliability statistics and principal components analyses to evaluate the underlying structure of the emerging scales, and eliminate items that undermined the cohesiveness of the scales without adding to the validity argument. To this end, we reviewed corrected item-total correlations, and alpha-if-item-deleted statistics for each item, together with each item's loading on the first unrotated principal component.
- (3) We examined facet-level statistics where necessary and appropriate. Occasionally, for example, facet-level analyses revealed pockets of unidimensionality that were not obvious when analyses were conducted at the scale level. This allowed us to fine tune our item selection approach for each scale and to gain further insight into the structure underlying each scale. This, in turn, had implications for the most appropriate and interpretable reliability coefficients to use, among other things.

As depicted in Figure 52, the individuals who were administered the ABC-2 Field Test were also administered other measures, or alternate equivalent forms of the ABC-2, to provide data for other studies described more fully below. The number of test-takers providing data for each of these studies is shown in Figure 52. The studies included: (1) an evaluation of test-retest reliability; (2) an evaluation of relationships with measures representing other individual difference domains and measures of bias, including the BICC developed by the Sirius performer research teams; and (3) an evaluation of the relationship between the ABC-1 and the ABC-2.

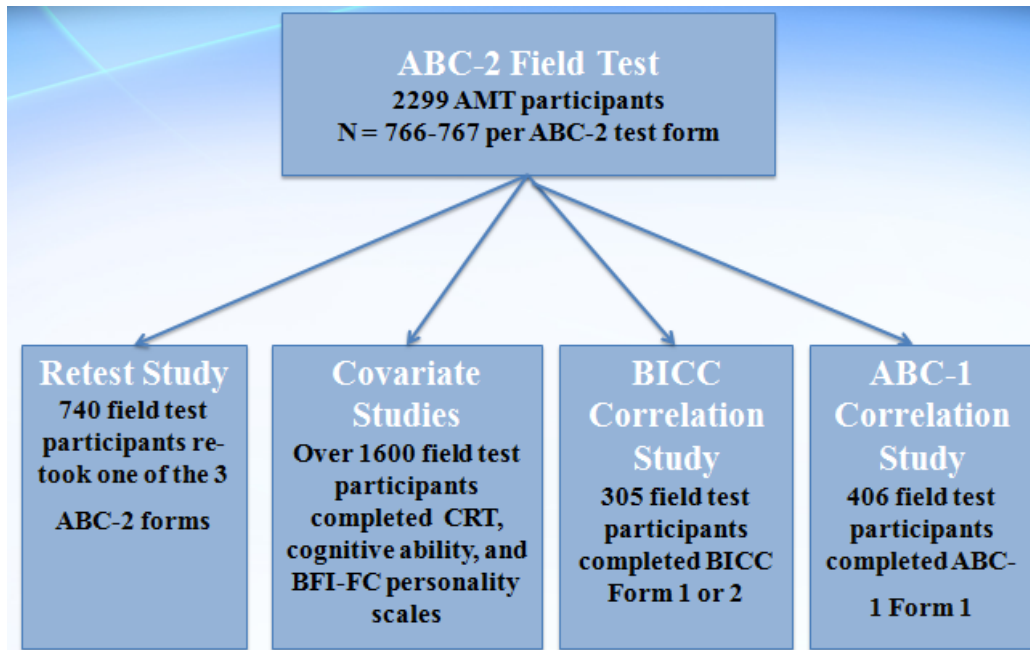


Figure 52: ABC-2 Field Trial Study Design.

3.6.2 Results and Discussion

3.6.2.1 Descriptive Statistics

Table 27 shows descriptive statistics for each ABC-2 BE and RD scale. We computed raw scale-scores for each BE and RD by creating unit-weighted composites of all the item scores comprising each bias scale (see Tables 28 - 31 for a listing of the items across forms for each bias scale). In addition, we differentially weighted the item scores in each of the BE scales in order to ensure that each facet contributed equally to the total scale-score. Higher scores for each item and scale indicate less biased responses. As shown in Figures 53 and 54, histograms of total raw scale-scores for each of the bias scales and test forms reveal approximately normal (bell-shaped) distributions, with the exception of the RD scales, which are not normal because they are skewed.

Table 27: Descriptive Statistics for ABC-2 ANC, REP, PRO, and RD Scales by Form (Raw-scores).

Scale	Mean	SD	Skew	Kurtosis	n
ANC, Form 1	-.07	6.16	-.33	.29	629
ANC, Form 2	.04	5.75	-.30	.20	647
ANC, Form 3	-.01	5.63	.05	.45	657
REP, Form 1	.01	6.98	.70	.70	650
REP, Form 2	.00	6.85	.63	.46	692
REP, Form 3	.00	6.90	.63	.57	668
PRO, Form 1	.01	7.47	-.02	.20	650
PRO, Form 2	.01	7.17	.05	.12	691
PRO, Form 3	.00	7.79	.24	.27	666
RD, Form 1	5.65	2.65	-.46	-.92	651
RD, Form 2	5.76	2.45	-.47	-.82	692
RD, Form 3	5.75	2.35	-.46	-.75	668

Table 28: ANC Behavioral Elicitation Items and Paradigms across Forms

Facet	Form 1	Form 2	Form 3
Numerical Priming	ANC_BE_01_Front_Load_Washer_HA_LINK	ANC_BE_01_Front_Load_Washer_HA_LINK	ANC_BE_01_Front_Load_Washer_HA_LINK
	ANC_BE_03_Portable_Dishwasher_HA	ANC_BE_01_Steel_Gas_Range_HA	ANC_BE_01_Steel_Gas_Range_HA
Selective Accessibility	ANC_BE_02_Candy_Jar_Guess_LINK	ANC_BE_02_Candy_Jar_Guess_LINK	ANC_BE_02_Candy_Jar_Guess_LINK
	ANC_BE_02_Coin_Jar_Guess_LINK	ANC_BE_02_Coin_Jar_Guess_LINK	ANC_BE_02_Coin_Jar_Guess_LINK
Comparative Judgment	ANC_BE_03_GPS	ANC_BE_03_GPS	ANC_BE_03_Laptop_LINK
	ANC_BE_03_Laptop_LINK	ANC_BE_03_Laptop_LINK	ANC_BE_03_Bicycle
	ANC_BE_03_Bicycle	ANC_BE_03_Couch	ANC_BE_03_Camera
	ANC_BE_03_Camera	ANC_BE_03_Microwave	ANC_BE_03_Couch
Self-Generated Anchor	ANC_BE_03_Microwave	ANC_BE_03_TV	ANC_BE_03_TV
	ANC_BE_05_Products_Rare_Cheese_HA	ANC_BE_05_Products_LA_LED_TV	ANC_BE_05_Earth_Days_Mars
	ANC_BE_05_Products_Avg_Cheese_HA	ANC_BE_05_Products_LA_Plasma_TV	ANC_BE_05_Earth_Days_Jupiter
	ANC_BE_05_Products_Choc_Truffles_HA	ANC_BE_05_Products_LA_Luxury_Watch	ANC_BE_05_Earth_Days_Saturn
	ANC_BE_05_Products_Belgian_Chocolates_HA	ANC_BE_05_Products_LA_Sports_Watch	ANC_BE_05_Earth_Days_Neptune
	ANC_BE_05_Products_World_Atlas_HA	ANC_BE_05_Products_LA_Steak_Dinner_6	
	ANC_BE_05_Products_Headphones_HA	ANC_BE_05_Products_LA_Hawaii_Wknd_2	
Focalism	Focalism_Car	Focalism_Dog_Breeders	Focalism_Italian_Ovens
	Focalism_Condo	Focalism_French_Bicycle	Focalism_Stereos

Note. *Variables that end in “_LINK” are linking items for test equating purposes.

Table 29: REP Behavioral Elicitation Items and Paradigms across Forms

Paradigm	Form 1	Form 2	Form 3
Base Rate Neglect	REP_BE_01_Tanks_LINK	REP_BE_01_Tanks_LINK	REP_BE_01_Tanks_LINK
	REP_BE_01_Vet_School_LINK	REP_BE_01_Vet_School_LINK	REP_BE_01_Vet_School_LINK
	REP_BE_01_Cruise	REP_BE_01_HS_Activities	REP_BE_01_Medical_School
Sample Size Insensitivity	REP_BE_02_Astronaut_LINK	REP_BE_02_Astronaut_LINK	REP_BE_02_Astronaut_LINK
	REP_BE_02_Bungee_Jumping	REP_BE_02_Brain_Gain	REP_BE_02_Drug_Side_Effects
	REP_BE_02_One_on_One_v1	REP_BE_02_One_on_One_v2	REP_BE_02_One_on_One_v2
	REP_BE_02_Marbles_LINK	REP_BE_02_Marbles_LINK	REP_BE_02_Marbles_LINK
Conjunction Fallacy	REP_BE_04_Bomb_Threat_DP_LINK	REP_BE_04_Bomb_Threat_DP_LINK	REP_BE_04_Bomb_Threat_DP_LINK
	REP_BE_04_IRS_DP	REP_BE_04_Security_Guard_DP	REP_BE_04_Airport_Security_DP
	REP_BE_04_Fishing	REP_BE_04_Tiring_Weekend	REP_BE_04_Grammarian
Non-Random Sequence Fallacy	REP_BE_05_Retrospective_Colored_Marbles	REP_BE_05_Jury_Duty	REP_BE_05_Retrospective_Colored_Marbles
	REP_BE_05_Girls_BBall	REP_BE_05_Dice	REP_BE_05_Jury_Duty
	REP_BE_05_Jury_Duty	REP_BE_05_Girls_BBall	REP_BE_05_Girls_BBall

Note. *Variables that end in “_LINK” are linking items for test equating purposes.

Table 30: PRO Behavioral Elicitation Items and Paradigms across Forms

PRO Paradigm	Form 1	Form 2	Form 3
False Consensus	PRO_BE_01_Anonymous_Posting	PRO_BE_01_GMOs	PRO_BE_01_Fat_Tax
	PRO_BE_01_Tipping_LINK	PRO_BE_01_Tipping_LINK	PRO_BE_01_Tipping_LINK
	PRO_BE_01_Video_Games	PRO_BE_01_Vaccination	PRO_BE_01_Criminal_Justice
Knowledge Projection	PRO_BE_02_KP_Burns_LINK	PRO_BE_02_KP_Burns_LINK	PRO_BE_02_KP_Burns_LINK
	PRO_BE_02_KP_Himalayas_LINK	PRO_BE_02_KP_Himalayas_LINK	PRO_BE_02_KP_Himalayas_LINK
	PRO_BE_02_KP_Salk_LINK	PRO_BE_02_KP_Salk_LINK	PRO_BE_02_KP_Salk_LINK
	PRO_BE_02_KP_Pegasus	PRO_BE_02_KP_Pancreas	PRO_BE_02_KP_Beagle
	PRO_BE_02_KP_Canberra	PRO_BE_02_KP_Morse	PRO_BE_02_KP_Barton
	PRO_BE_02_KP_Martin_Luther	PRO_BE_02_KP_US	PRO_BE_02_KP_Yangtze
	PRO_BE_02_KP_Jenner	PRO_BE_02_KP_Trevi	PRO_BE_02_KP_Guest
Social Projection	PRO_BE_06_High_Tech_BFI_LINK	PRO_BE_06_High_Tech_BFI_LINK	PRO_BE_06_High_Tech_BFI_LINK
	PRO_BE_06_Photojournalism_BFI_Pt3	PRO_BE_06_Grantwriting_BFI_Pt3	PRO_BE_06_Grantwriting_BFI_Pt4
	PRO_BE_06_Design_BFI_Pt4	PRO_BE_06_Photojournalism_BFI_Pt4	PRO_BE_06_Photojournalism_BFI_Pt2
	PRO_BE_06_Grantwriting_BFI_Pt2	PRO_BE_06_Design_BFI_Pt2	PRO_BE_06_Design_BFI_Pt3
	PRO_BE_04_STEU36_OY_LINK	PRO_BE_04_STEU36_OY_LINK	PRO_BE_04_STEU36_OY_LINK
	PRO_BE_03_STEM18_OY_LINK	PRO_BE_03_STEM18_OY_LINK	PRO_BE_03_STEM18_OY_LINK
	PRO_BE_03_STEM30_OY	PRO_BE_03_STEM27_OY	PRO_BE_04_STEU35_OY
	PRO_BE_04_STEU4_OY	PRO_BE_04_STEU7_OY	PRO_BE_04_STEU27_OY
	PRO_BE_04_STEU15_OY	PRO_BE_04_STEU14_OY	PRO_BE_04_STEU25_OY
	PRO_BE_04_STEU6_OY	PRO_BE_04_STEU3_OY	PRO_BE_04_STEU2_OY
	PRO_BE_06_Eye_Doctor	PRO_BE_04_Keys_Female	PRO_BE_06_Thesis

Note. *Variables that end in “_LINK” are linking items for test equating purposes.

Table 31: Distribution of RD Items across Biases and Forms

Bias	Form 1	Form 2	Form 3
Anchoring Bias	ANC_RD_01_English_Novel	ANC_RD_01_Biodiversity	ANC_RD_01_Art_Gallery
	ANC_RD_01_Open_House_LINK	ANC_RD_01_Open_House_LINK	ANC_RD_01_Open_House_LINK
	ANC_RD_01_Senator	ANC_RD_01_RSVP	ANC_RD_01_Cake
Representativeness Bias	REP_RD_01_Go_Fish_LINK	REP_RD_01_Go_Fish_LINK	REP_RD_01_Go_Fish_LINK
	REP_RD_01_Tim_Commute	REP_RD_01_Pet_Turtle	REP_RD_01_Classroom_Computer
	REP_RD_01_Weekend_Weather	REP_RD_01_Volleyball	REP_RD_02_1
Projection Bias	PRO_RD_01_Tanzania_LINK	PRO_RD_01_Tanzania_LINK	PRO_RD_01_Tanzania_LINK
	PRO_RD_01_Pressure	PRO_RD_00_Orientation	PRO_RD_01_Escargot
	PRO_RD_01_Pirates	PRO_RD_01_Wrong_Paperwork	PRO_RD_01_Almentania

Note. *Variables that end in “_LINK” are linking items for test equating purposes.

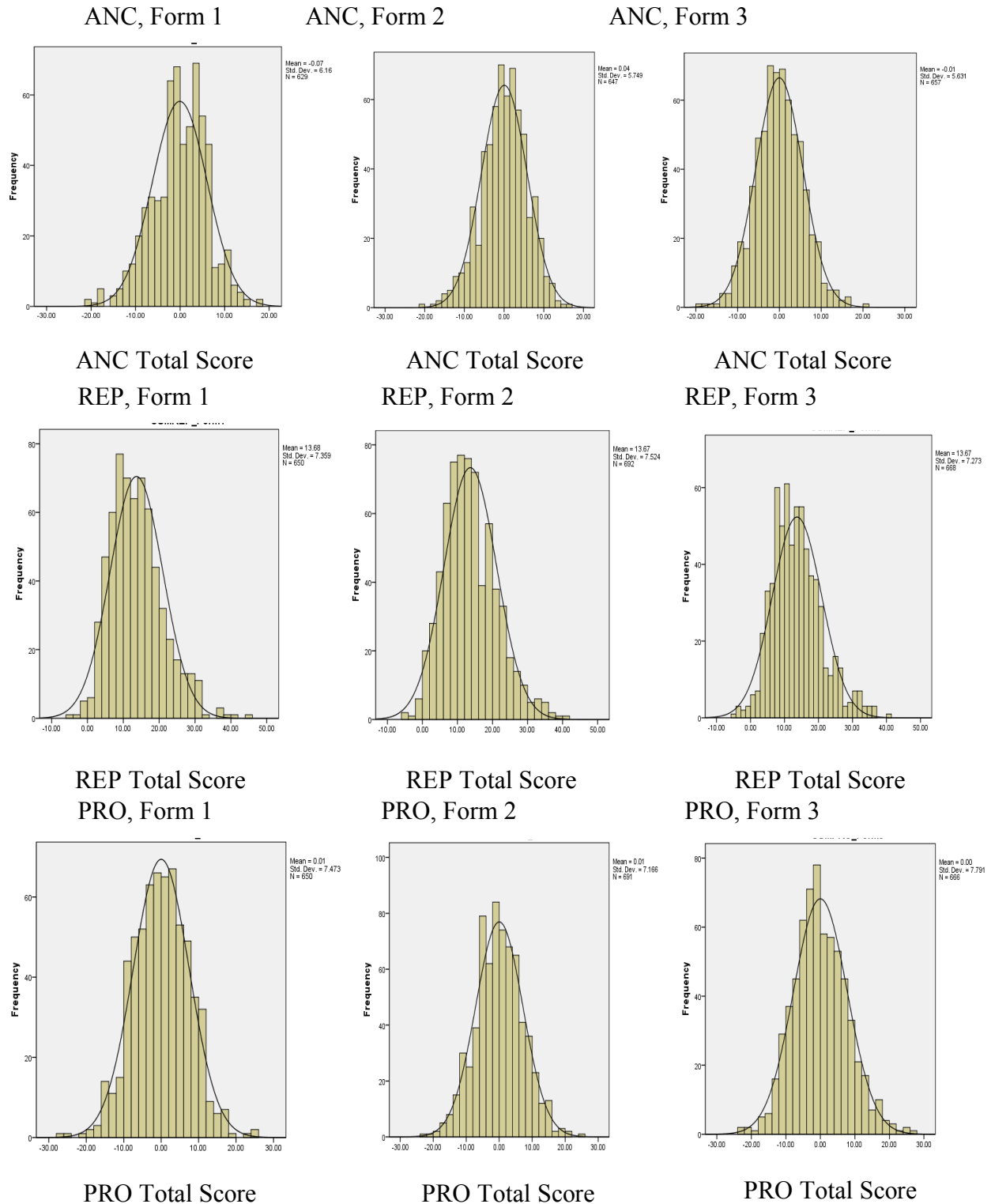
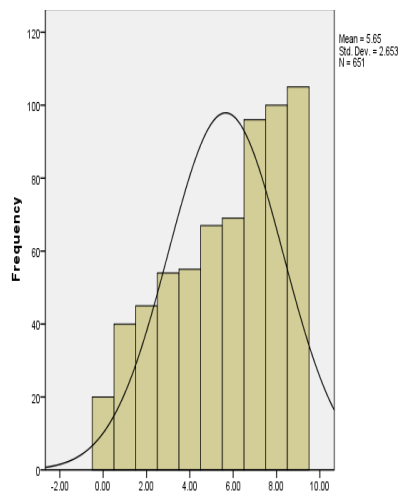


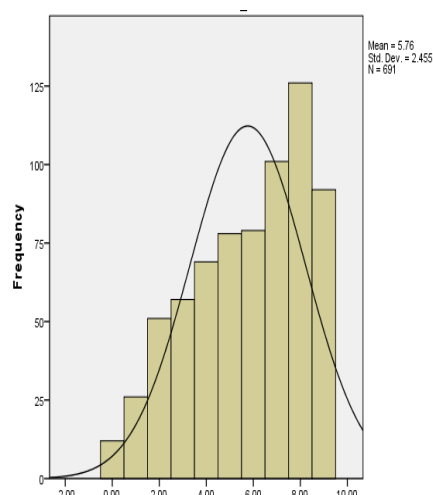
Figure 53: ABC-2 Anchoring Bias (ANC), Representativeness Bias (REP), and Projection Bias (PRO) Raw-Score Frequency Distributions.

RD, Form 1



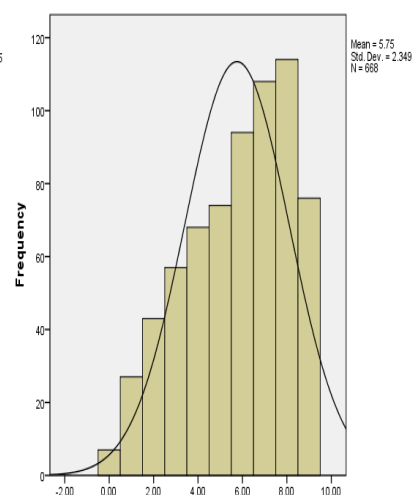
RD Total Score

RD, Form 2



RD Total Score

RD, Form 3



RD Total Score

Figure 54: ABC-2 RD Total Raw-Score Frequency Distributions.

3.6.2.2 Reliability Analyses and Results

Table 32 presents various reliability estimates for the ABC-2 scales based on the field test data. For each BE test form, several reliability metrics were computed: Cronbach's alpha, a measure of internal consistency reliability, and test-retest reliability, a measure of temporal stability. The retest interval ranged from 4 to 5 weeks. In addition, for the BE scales, stratified alpha --an alternate measure of internal consistency reliability appropriate for multidimensional data --was computed. According to Osburn (2000, p. 347), "When the components of a composite can be grouped into subsets on the basis of content or difficulty, stratified alpha may be a better estimate of the true reliability than coefficient alpha computed on the same composites."

Given the magnitude of the BE scale alpha coefficients, especially for anchoring, item analyses suggested that scales are not satisfactorily cohesive for alpha to serve as an appropriate/ideal reliability metric. As such, both alpha and stratified alpha are presented. Stratified alpha is not presented for the RD scales because item analyses revealed that, consistent with our Phase 1/ABC-1 results, RD scales are relatively unidimensional.

Consistent with the latter observation, the alpha coefficients for the three RD scales (i.e., for Forms 1-3) are higher than for any of the BE scales, with a mean alpha of .76. Alpha was lowest for the BE Anchoring scales indicating that consistent with item-analysis results, the Anchoring scales were the least cohesive of the three Phase 2 biases. It is also noteworthy that Anchoring is the only construct for which stratified alpha exceeded alpha, with the exception of Form 2, where alpha and stratified alpha were equal. For every other scale, alpha exceeded stratified alpha. This does not necessarily mean that Representativeness and Projection BE scales should be considered cohesive, but it does suggest that they are more cohesive than the Anchoring BE scales. Given the number of items in the Representativeness and Projection BE scales, the alpha coefficients are modest, but not low.

Test-retest reliability results for the BE scales are comparable to one another and to the relevant internal consistency reliability metric corresponding to those scales. That is, for anchoring, the mean test-retest reliability is exactly equal to stratified alpha, and for representativeness and

projection, test-retest reliability is comparable to alpha. The convergence of internal consistency and test-retest reliability metrics, together with the large sample sizes, suggest that for the BE scales reliability falls between .60 and .70.

Table 32: Reliability Analysis of ABC-2 Anchoring Bias (ANC), Representativeness Bias (REP), Projection Bias (PRO), and Recognition and Discrimination (RD) Scales

Scale	Stability		Internal Consistency				# of Items
	$r_{xx'}$	n	α	Stratified α	Mean Loading on 1st Unrotated PC	n	
ANC, Form 1	.62	129	.59	.62	.30	629	17
ANC, Form 2	.67	78	.59	.59	.27	647	17
ANC, Form 3	.64	119	.54	.70	.25	657	15
Mean	.64		.57	.64	.27		
REP, Form 1	.60	136	.66	.55	.32	650	19
REP, Form 2	.70	90	.55	.54	.20	692	19
REP, Form 3	.65	121	.65	.54	.32	668	19
Mean	.65		.62	.54	.28		
PRO, Form 1	.55	136	.64	.58	.33	650	21
PRO, Form 2	.69	90	.63	.54	.29	691	21
PRO, Form 3	.62	121	.70	.61	.37	666	21
Mean	.62		.66	.58	.33		
RD, Form 1	.72	136	.80		.62	651	9
RD, Form 2	.68	90	.76		.58	692	9
RD, Form 3	.61	122	.72		.55	668	9
Mean	.67		.76		.58		

Note. $r_{xx'}$ is test-retest reliability. PC is principal component. Gray highlighting indicates cells in which the reliability metric is not appropriate.

3.6.2.3 Covariate Study

We included several measures of established individual-difference domains as well as a bias susceptibility measure along with the ABC-2 in the field study. This was done primarily to evaluate the convergent and discriminant validity of the ABC-2 scales and embed the bias constructs measured by the ABC-2 within a nomological network (Cronbach & Meehl, 1955). Another purpose was to determine whether measures of certain constructs should be treated as covariates in various correlational analyses done as part of this project.

Table 33: Correlations between ABC-2 Scaled Scores and Personality and Cognitive Ability Scales.

ABC-2 Scaled Score	Forced-Choice Big Five Personality Inventory					Omnibus Cognitive Ability Test		
	ES	E	O	A	C	g	QA	VA
Anchoring, Form 1	.06	-.04	.03	-.13**	.14**	.13**	.15**	.04
Anchoring, Form 2	.14**	.06	.04	-.08	-.02	.12*	.10*	.11*
Anchoring, Form 3	.02	.05	.07	-.03	.08	.07	.05	.08
Means	.07	.02	.04	-.08	.07	.10	.10	.07
Representativeness, Form 1	.04	-.16**	.06	-.14**	.03	.23**	.24**	.14**
Representativeness, Form 2	-.01	-.03	.13**	-.06	-.10*	.21**	.22**	.12**
Representativeness, Form 3	-.05	-.12*	.03	-.11*	-.06	.30**	.31**	.18**
Means	.00	-.10	.07	-.10	-.04	.25	.25	.15
Projection, Form 1	-.07	-.06	.12*	-.08	-.14**	.24**	.16**	.32**
Projection, Form 2	-.06	-.09*	.11*	-.11*	-.12*	.21**	.17**	.20**
Projection, Form 3	-.09	-.09*	.11*	-.09	-.12**	.21**	.18**	.19**
Means	-.07	-.08	.11	-.09	-.13	.22	.17	.24
Recognition and Discrimination, Form 1	.05	-.14**	.14**	.00	-.07	.58**	.49**	.55**
Recognition and Discrimination, Form 2	-.01	-.11*	.08	-.04	-.13**	.48**	.41**	.43**
Recognition and Discrimination, Form 3	-.02	-.12*	.15**	-.03	-.09	.45**	.37**	.44**
Means	.01	-.12	.12	-.02	-.09	.50	.42	.47

Note. $n = 410-465$. ** Correlation is significant at the .01 level (one-tailed). * Correlation is significant at the .05 level (one-tailed).

Demographic Variables. We administered a demographics survey as part of the field test, and computed correlations between demographic variables and the ABC-2 scales. Although several correlations were statistically significant due to the large sample sizes, none of the correlations reached practical significance.

Personality. The personality domain was operationalized using two Big-Five measures: (1) the Big-Five inventory (BFI-44; John, Donahue, & Kentle, 1991); and (2) the Big-Five inventory (forced-choice version, BFI-FC; Brown & Maydeu-Olivares, 2011). The BFI-44 and BFI forced choice (BFI-FC) are both measures of the five factor model of personality. The BFI-44 and BFI-FC measures showed substantial convergent validity. The correlations between their analog scales ranged from .56 for agreeableness to .85 for emotional stability ($n = 397$). Because of certain anomalous correlations between the BFI-44 and other measures in the Phase 2 field tests, and the ubiquitous faking problem associated with self-report true/false and likert-type scales, we chose to switch to the BFI-FC measure as our primary operationalization of the personality domain in Phase 2.

In general, the correlations between the ABC-2 scales, both BE and RD, are largely independent of the personality domain. The highest correlation in Table 33 is $r = -.16^{30}$ ($p < .01$) between Extraversion and REP, Form 1. Note, however, that the correlation between Extraversion and REP, Form 2 is only $r = -.03$ (*n.s.*). Similarly, Openness correlates $r = .14$ and $.15$, respectively, with RD, Forms 1 and 3 ($ps < .05$), though only $r = .08$ (*n.s.*) with RD, Form 2. A more consistent finding is that Openness correlates positively with Projection ($rs = .11 - .12$, $ps < .05$). Another consistent finding is that Conscientiousness correlates negatively with Projection ($rs = -.12$ to $-.14$, $ps < .05$). In the latter case, this means that more conscientious test-takers are slightly more susceptible to the Projection Bias, but not to other biases. Moreover, conscientious people, who presumably took greater pains to read the information about the ABC-2 biases prior to responding to the RD items, nevertheless did not score higher on RD, and in fact, had a slight tendency to score lower.

Although these post-hoc interpretations are interesting, it should be emphasized that no uncorrected correlation exceeded $|.16|$. So, a key result is that the personality domain and the BE and RD bias domain appear to be independent. As such, we have evidence of discriminant validity of the ABC-2 with respect to the personality domain.

Cognitive Ability. The cognitive ability domain was operationalized using an omnibus battery of verbal and quantitative measures drawn from the ETS factor-reference kit (Ekstrom, French, & Harman, 1979). The battery included items corresponding to the following verbal and quantitative ability sub-scales:

- (1) Vocabulary: Participant is prompted to suggest a synonym for a target word. Each target word is presented with 4 radial options in a single-response format. Participants have 4 minutes to solve a set of ten problems.
- (2) Analogies: Participants are given a pair of words and then prompted to select a pair of words that reflect an analogous relationship from five word pairs. Each word pair is listed as a radial option.
- (3) Sentence Completion: Participants read a sentence with an underlined word or phrase. Participants are then given the option to either keep the sentence as is or to replace the

³⁰ Positive correlation indicates that test-takers scoring higher on personality traits are less susceptible to a biases and score higher on RD. A negative correlation indicates that test takers scoring higher in personality traits are more susceptible to bias and score lower on RD.

underlined word or phrase with one of three alternatives. Participants have 4 minutes to complete 5 questions.

- (4) Math Word Problem: Participants are given a word problem in math with five radial responses to select from. Questions are presented in a set of ten.
- (5) Math Operation Problems: Participants read a math word problem and then indicate which operation(s) they would use to solve the problem. Operations are presented in five radial options. Questions are presented in a set of ten.

Factor analytic results repeatedly showed that, in addition to the general ability factor (g), quantitative and verbal factors emerged. As such, we computed factor scores for each. The results for cognitive ability are more consistent than those for personality, more specifically, all three ability factors showed positive correlations with ABC-2 scales at both statistically and practically significant levels. The correlations between ability and Anchoring are exceptions to this. While positive, they are substantially smaller.

Predictably, the largest correlations are between cognitive ability and RD. These range from $r_s = .43 - .65$, uncorrected. Also, predictably, the verbal factor had higher correlations than the quantitative factor for RD. Given the reading load necessary to complete the RD items, both the instructions and the text-based items, and also the need to adapt the information in the descriptions/illustrations of the biases provided to test-takers, this pattern of correlations with cognitive ability is entirely sensible. While we still observe positive correlations between cognitive ability and the Representativeness and Projection Bias scales, the pattern of correlations for the Quantitative and Verbal Ability factors differs for Representativeness and Projection. Specifically, verbal ability correlates more highly than quantitative ability for Projection, whereas, quantitative correlates more highly than verbal ability for Representativeness. These results are also sensible, given the fact that the ABC-2 Representativeness items/tasks are largely quantitative in nature. For example, gambler's fallacy items require some statistical thinking, as do base rate neglect items. Test-takers are not required to do formal and exact calculations, but the tendency to think statistically/mathematically seemed likely to produce higher scores on items such as these, among other Representativeness items/tasks. With regard to Projection, there is no obvious reason why correlations with verbal ability should exceed those of quantitative ability. It should also be noted that, for two out of the three Projection forms, the difference in correlations with verbal and quantitative ability are very similar. In general, correlations between cognitive ability and ABC-2 scales are substantially lower for BE scales than for RD scales. Given the heavy reading load associated with RD scale- this is also entirely sensible.

Another key point from Table 33 is that cognitive ability is largely independent of Anchoring bias susceptibility. This may be a consequence of Anchoring being a less cohesive construct than Representativeness and Projection, together with greater item specificity in the Anchoring domain. That said, the lower positive correlations between Anchoring and cognitive ability warrant further investigation.

Table 34: Zero-Order Correlations between Biases and Relevant Demographic Variables

ABC-2 Scale	Age	Gender (1 = Male, 2 = Female)	Did Not Attend College (1 = Yes, 2 = No)	Hispanic Or Latino? (1 = Yes, 2 = No)	Are You Currently Employed? (1 = Yes, 2 = No)	Number Of Psychology Courses Taken	Cumulative GPA
ANC Form 1	.09*	-.04	-.01	.03	-.04	-.14**	-.05
ANC Form 2	.08*	-.06	-.10*	-.03	-.02	.03	-.02
ANC Form 3	.08*	-.10*	-.01	.05	-.07	.00	.06
REP Form 1	-.08*	-.08*	-.08	.03	-.04	.00	.10*
REP Form 2	-.07	-.16**	-.11**	.01	-.02	.06	.03
REP Form 3	.01	-.09*	-.07	.02	-.04	.01	.03
PRO Form 1	.00	-.03	-.09*	-.01	.03	.01	-.01
PRO Form 2	-.05	-.06	-.07	-.01	.01	.05	.07
PRO Form 3	-.06	-.06	-.04	.02	.07	.03	.07

Note. $n = 568-688$. ** Correlation is significant at the .01 level (one-tailed). * Correlation is significant at the .05 level (one-tailed).

Bias Susceptibility We had two measures of bias susceptibility in the Phase 2 Field Test: (1) a 10-item version of the Cognitive Reflection Test (CRT; Frederick, 2005, personal communication); and (2) the BICC. These are necessarily not clear “marker” tests because of the novelty of bias susceptibility measurement in the individual difference domain. However, both have been carefully developed and represent the best available measures for evaluation of convergent validity.

Table 35 presents correlations between CRT and ABC-2 scales. Zero-order correlations³¹ show that the CRT correlates at statistically and practically significant levels with all ABC-2 constructs, with its highest correlations being with Representativeness and RD. We note, however, that the CRT correlates highly with general cognitive ability ($r = .62$). Given this, we computed partial correlations³² between the CRT and ABC-2 scales, controlling for g ³³. After controlling for g , the correlations dropped substantially for RD and also, albeit to a lesser extent, for Representativeness and Projection. The drop in correlation for RD makes sense given its higher correlation with g . Despite these drops, the CRT retains practically significant correlations

³¹ A zero-order correlation measures the magnitude or strength of an association between two variables, without controlling for any other factors.

³² A partial correlation controls for the effects of a third variable to determine whether a zero-order correlation changes.

³³ There is a widely accepted view among psychometric experts that the structure of human cognitive abilities is hierarchical, with a single, highest-order factor usually called “general cognitive ability, or “ g ” (Neisser et al., 1996). General cognitive ability was defined by Humphreys (1979) as: “the resultant of the processes of acquiring, storing in memory, retrieving, combining, comparing, and using in new contexts information and conceptual skills...” (p. 115).

with Representativeness and RD. Its correlation with Projection, however, appears to be primarily explained by *g*.

Table 35: Zero-Order and Partial Correlations between Cognitive Reflection Test (CRT) and ABC-2 Scale-scores.

	Zero-Order Correlations with CRT	Partial <i>r</i> , Controlling for <i>g</i>
ABC-2 Scaled Score		
Anchoring	.14**	.10**
Representativeness	.35**	.27**
Projection	.20**	.07**
Recognition and Discrimination (RD)	.42**	.16**

Note. *n* = 1243-1370. All correlations are significant at the .01 level (one-tailed).

Table 36 shows correlations between ABC-2 and BICC scales. The correlations between comparable scales on the ABC-2 and BICC in Table 36 show convergent validity support for Representativeness and Projection and, to a lesser extent, Anchoring. Looking at the off-diagonal elements -- and setting aside RD -- we see discriminant validity support for Representativeness and Projection, but not for Anchoring. The ABC-2 RD scale correlates very highly with the BICC Representativeness scale, and modestly with the BICC Anchoring and Projection scales.

Table 36: Correlations between ABC-2 and BICC Scales

ABC-2 Scale	BICC Scale (Uncorrected)			BICC Scale (Corrected)		
	ANC	REP	PRO	ANC	REP	PRO
ANC	.16	.21	.06	.28	.36	.10
REP	.06	.39	.19	.10	.64	.30
PRO	.15	.12	.27	.25	.19	.42
RD	.23	.50	.18	.35	.74	.26

Note. The correlations in the last three columns are disattenuated for unreliability in both the ABC-2 and BICC. Convergent validity coefficients (correlations between analogous scales on the ABC-2 and BICC) are shown in bold.

Table 37 shows comparative correlations between ABC-2 and BICC scales, on the one hand, and personality, CRT, and cognitive ability variables, on the other hand. In general, the data in the table show a similar pattern of correlations with covariates for the ABC-2 and BICC. This is especially evident in the case of Projection, with the exception of correlations with Openness. Correlations involving Representativeness also are similar, with the exception of correlations between Representativeness and Emotional Stability, and smaller correlations between ABC-2 Representativeness and cognitive ability than were found for the BICC Representativeness scale. Correlations between ABC-2 and BICC scales and the CRT are roughly comparable, though

correlations between the CRT and ABC-2 were smaller than the correlations between the CRT and BICC in the case of Anchoring and Representativeness.

Table 38 shows correlations between ABC-2 and BICC scale-scores, on the one hand, and demographic variables, on the other hand. Although the correlations are uniformly low, we again see a similar pattern of correlations for the ABC-2 and BICC scales.

Table 37: Correlations between ABC-2, BICC, and Personality, CRT, and Cognitive Ability Variables.

	Openness	Emotional Stability	Conscientiousness	Extraversion	Agreeableness	CRT	g	QA	VA
ABC-2 Scale-score									
Anchoring	.04	.07**	.07**	.02	-.08**	.14**	.13**	.12**	.09**
Representativeness	.07**	.00	-.05*	-.10**	-.10**	.35**	.31**	.32**	.18**
Projection	.11**	-.08*	-.13**	-.08**	-.10**	.20**	.28**	.22**	.31**
BICC Scale-score									
Non-Focalism Average (Anchoring)	.13*	.13*	.04	-.06	-.02	.25*	.25**	.21**	.22**
Representativeness	.09	-.14*	-.17**	-.14*	-.14*	.46**	.50**	.42**	.44**
Projection	.01	-.11*	-.13*	-.11	-.02	.15*	.16**	.10	.20**

Note. ABC-2 $n = 1243-1370$. BICC $n = 235$. ** Correlation is significant at the .01 level (one-tailed). * Correlation is significant at the .05 level (one-tailed).

Table 38: Correlations between ABC-2, BICC, and Demographic Variables.

Variable	Age	Gender (Male = 1, Female = 2)	College? (Yes = 1, No = 2)	Currently Employed (Yes = 1, No = 2)	#Psych Courses	Father's Schooling	Mother's Schooling	GPA
ABC-2 Scaled Scores								
Anchoring	.08**	-.07**	-.03	-.04**	-.04*	.02	.00	.00
Representativeness	-.04**	-.11**	-.06**	-.03	.03	.08**	.09**	.05*
Projection	-.04*	-.05*	-.06**	.04*	.03	.06**	.10**	.04*
BICC Scale-scores								
Anchoring	.08	-.15**	-.03	-.06	-.04	.06	.03	.05
Representativeness	-.09	-.16**	-.07	-.03	.07	.10*	.11*	.14**
Projection	-.03	.11*	-.07	.02	.02	.02	-.08	.03

Note. ABC-2 n = 1728 – 2003. BICC n = 271 – 303. ** Correlation is significant at the 0.01 level (1-tailed). * Correlation is significant at the 0.05 level (1-tailed).

3.6.2.4 ABC-2 Intercorrelations

Table 39 shows intercorrelations between each of the ABC-2 scale-scores across forms.

Table 39: Intercorrelations between ABC-2 Scale-scores.

	ABC-2 Scale-score	ANC	REP	PRO	RD
1	Anchoring (ANC)		.20	.03	.09
2	Representativeness (REP)	.14**		.11	.31
3	Projection (PRO)	.02	.07		.27
4	Recognition and Discrimination (RD)	.06	.21**	.19**	

Note. $n = 2,012$. Correlations based on observed scores are below the diagonals and correlations based on scores disattenuated for measurement error are shown in bold above the diagonal.

Most notably, Table 39 does not show a positive manifold, making computation of an overall battery score inappropriate. There are several additional points to be made about these results. First, Table 39 indicates that the RD scale and the BE scales are largely independent, with an average corrected intercorrelation of $r = .22$.

The highest intercorrelations are between RD, on the one hand, and Representativeness and Projection ($r_s = .21$ and $.19$, respectively, both $p < .01$; corrected $r_s = .31$ and $.27$, respectively). By contrast, RD correlates only $.06$ with Anchoring, corrected $r = .09$. Among the BE scales, there is a small correlation between Anchoring and Representativeness ($r = .14$, $p < .01$; corrected $r = .20$).

3.6.2.5 Correlations between ABC-1 and ABC-2

Correlations between the ABC-1 and ABC-2 scales are presented in Tables 40 and 41. These tables show the following:

- ABC-1 CB is uncorrelated with ABC-2 scales
- ABC-1 FAE showed modest positive correlations with ABC-2 REP
- ABC-1 BBS showed modest negative correlations with ABC-2 PRO and RD, indicating greater BBS susceptibility is associated with more RD knowledge and less PRO susceptibility
- ABC-1 RD correlated highly with ABC-2 RD, which is consistent with the general finding that RD is relatively unidimensional, independent of bias content

First, with the exception of the correlation between the Phase 1 and Phase 2 RD scales, the correlations shown in Table 40 range from $r = 0$ to $.23$, indicating that the bias susceptibility measures are relatively independent. The Phase 1 and Phase 2 RD scales correlate highly ($r = .52$, $p < .01$, disattenuated³⁴ $r = .68$). There are, however, a few modest correlations worthy of mention. The correlation between BBS and PRO is $-.18$ ($p < .01$, disattenuated $r = -.27$). That is, the more people think that others think like themselves (showing greater PRO), the less likely they are to attribute more bias to others (showing less BBS). The correlation between REP and FAE is $.21$ ($p < .01$, disattenuated $r = .29$). That is, the more susceptible people are to REP the more susceptible they are likely to be to FAE. Interestingly, REP correlates about the same with FAE and RD ($r = .21$ and $r = .23$, respectively; both $p < .01$, disattenuated $r = .29$ and $r = .33$).

Table 40: Correlations between ABC-1 and ABC-2 Scale-scores

ABC-2 Scale-score	ABC-1 Scale-score			
	CB	FAE	BBS	RD
Anchoring	0.07	0.06	-0.06	0.09
Representativeness	0.10	0.21	-0.07	0.23
Projection	0.02	0.01	-0.18	0.21
Recognition and Discrimination	0.08	0.08	-0.22	0.52

Table 41: Correlations between ABC-1 and ABC-2 Scale-scores, Disattenuated for Unreliability

ABC-2 Scale-score	ABC-1 Scale-score			
	CB	FAE	BBS	RD
Anchoring	0.13	0.09	-0.10	0.13
Representativeness	0.18	0.29	-0.11	0.33
Projection	0.03	0.01	-0.27	0.29
Recognition and Discrimination	0.13	0.10	-0.31	0.68

Note. $n = 388-406$. CB = Confirmation Bias, FAE = Fundamental Attribution Error, BBS = Bias Blind Spot, and RD = Recognition and Discrimination.

Table 42 shows correlations between ABC-1 and ABC-2 scale-scores and the CRT, including partial correlations controlling for g . We controlled for g , because the CRT correlates quite highly with g . ABC scale-score correlations with the CRT were highly variable. They ranged from $r = -.15$ to $r = .43$; median $r = .17$. After controlling for g , the variability was preserved.

³⁴ A disattenuated correlation refers to a correlation to which a statistical correction is applied in order to produce an estimate of the “true” relationship between the variables being correlated. The statistical correction is intended to remove the influence of unreliability in the variables, which has the effect of “attenuating” (i.e., lowering the absolute value of) the correlation.

Controlling for g caused the zero-order correlations to drop substantially, but it should be noted that the partial correlation between REP and the CRT remained in the high .20s. The only other partial correlation of any magnitude was between the RD scale and the CRT.

Table 42: Correlations Between Cognitive Reflection Test (CRT) and ABC-1 and ABC-2 Scale-scores

	Zero-Order Correlations with CRT	Partial r, Controlling for g	Difference
ABC-1 Scale-score			
Confirmation	.04	.07	-.03
FAE	.09	.01	.08
BBS	-.15**	-.02	-.13
Recognition and Discrimination	.43**	.09	.34
ABC-2 Scale-score			
Anchoring	.14**	.10**	.04
Representativeness	.35**	.27**	.08
Projection	.20**	.07**	.13
Recognition and Discrimination	.42**	.16**	.26

Note. ABC-2 $n = 1243-1370$. ABC-1 $n = 399$. ** Correlation is significant at the .01 level (one-tailed). * Correlation is significant at the .05 level (one-tailed).

Table 43 shows correlations between ABC-1 scale-scores and personality and cognitive ability scores. ABC-1 BBS was the only ABC-1 scale that correlated with any of the Big-Five personality factors ($r = -.15$ with Openness) or cognitive ability ($r = -.23$ with g). FAE correlated $r = .13$ with g , but CB was essentially uncorrelated with cognitive ability. The ABC-1 RD scale correlated $r = .62$ with cognitive ability.

Table 43: Zero-order Correlations Between ABC-1 Scale-Scores and Personality and Cognitive Ability Factor-Scores.

ABC-1 Scale Score	Forced-Choice Personality Instrument					ETS Omnibus Ability Measure		
	ES	E	O	A	C	g	Quant.	Verbal
Confirmation Bias	.00	-.04	-.01	-.08	-.06	-.03	-.01	-.05
Fundamental Attribution Error	.03	-.02	.01	-.07	-.07	.13	.13	.09
Bias Blind Spot	-.06	.02	-.15	-.07	.05	-.23	-.16	-.28
Recognition & Discrimination	.08	-.14	.16	.06	-.08	.62	.55	.53

Note. For correlations between ABC-1 scale scores and covariates $n = 403-405$. g = General Cognitive Ability. ES = Emotional Stability. E = Extraversion. O = Openness. A = Agreeableness. C = Conscientiousness.

3.6.2.6 Conclusions Regarding Structure and Individual Difference Measurement of Biases

As in Phase 1, there is no support for a common bias susceptibility construct. Bias susceptibility appears to be formative rather than reflective; that is, it is best conceptualized as a linear combination of bias susceptibility scales, many of which are not correlated with one another. And, as in Phase 1, the RD scale is internally consistent and a good approximation of unidimensionality.

Once again, this score would be best understood as a concatenation of thematically related measures of the Phase 2 biases rather than a unidimensional scale measuring elicitation of ANC, REP, and PRO. And, once again, a label of overall bias susceptibility would be a label of convenience only.

As in Phase 1, the only way to achieve unidimensionality would have been to virtually eliminate the test's validity; for example, by limiting a great deal of critical content (attenuating content validity) and measuring so transparently that we would be measuring RD rather than BE (attenuating construct validity by introducing a confound). We believe that the ABC-2's validity renders it a fair measure of the content domain underlying the Phase 2 Sirius biases.

3.6.3 ABC-2 Pretest Sensitization Study

3.6.3.1 Purpose

Similar to the pretest sensitization study conducted prior to the Phase 1 IV&V, we conducted a study designed to investigate whether the ABC-2 creates a pre-test sensitization effect. Specifically, we asked whether taking the ABC-2 interacts with bias mitigating training interventions, thereby resulting in different post-intervention ABC-2 test performance than if no pretest were given.

3.6.3.2 Method

3.6.3.2.1 Participants

The Pretest Sensitization Study was administered to a total of 170 ETS employees, all of whom were located in the United States and recruited online through ETS's Performance Assessment Scoring Services (PASS). The sample of test-takers averaged 47 years of age ($SD = 11$ years); was 76% female; and approximately 83% Caucasian, 7% African-American, 4% Asian, and 6% multi-racial. Nine percent of the sample was Hispanic or Latino. All test-takers had graduated college, and the sample was high-achieving, with 73% reporting GPAs over 3.5. The majority of the sample reported having taken one or two psychology courses, though the vast majority had not taken more than four psychology courses. Participants were compensated at a rate of \$20/hour for completing the study.

3.6.3.2.2 Experimental Design and Procedure

Study participants were randomly assigned to the following three groups:

- (1) IARPA Video Experimental Group: Participants took the ABC-2, Form 1, as a pretest. Then they watched the IARPA instructional video about the Phase 2 cognitive biases before taking a posttest, which was the ABC-2, Form 2.

- (2) Control Video Group: Participants first took the ABC-2, Form 1, as a pretest; then watched an unrelated instructional video before taking the ABC-2, Form 2, as a posttest. The instructional video was the 30-minute lecture given by Dr. Steven Pinker about language and psychology that was administered in the ABC-1 Pretest Sensitization Study. The video was selected, because it is approximately the same length as and has audio-visual features comparable to the IARPA instructional video.
- (3) No Pretest Group: Participants did not take a pretest prior to watching the IARPA instructional video, but only took the ABC-2, Form 2, as a posttest. In place of the ABC-2 pretest, participants completed the BFI-44 personality and ETS Omnibus cognitive battery prior to watching the instructional video.

A demographics questionnaire was administered to all participants at the end of the study. Participants were emailed a link to access the assessments and instructional videos. Participants were instructed to complete all study activities in a single session lasting approx. 3 hours.

3.6.3.3 Results and Discussion

We computed and compared pretest and posttest scale-scores for each ABC scale and group in order to evaluate the following hypotheses:

- (1) Assuming pre-test scores are equivalent across groups, post-test scores will be higher in the IARPA Video Experimental Group and No Pre-test Group than in the Control Video Group for targeted ABC-2 scales.
- (2) Because the IARPA instructional video most directly targets explicit/declarative knowledge of cognitive biases, improvement will be shown for the ABC-2 RD scale, but there may be little or no improvement in the BE scales.

In order to evaluate these hypotheses, we computed pre-test and post-test scale and facet-level scores for each ABC-2 scale and condition and conducted an Analysis of Covariance (ANCOVA) to compare posttest performance in the IARPA Video group with performance in the Control Video group, controlling for pretest scores. As shown in Table 44, pretest performance was generally comparable across groups for each scale. Consistent with our hypotheses, watching the IARPA instructional video enhanced performance on the RD post-test relative to watching the Control video [$F(1,102) = 39.81, p < .01$]. RD scores improved by 39 points from 31 to 70 in the IARPA instructional video condition (Cohen's $d = 1.53$). It should be noted, however, that RD scores also improved in the Control Video condition, but to a substantially lesser extent (Cohen's $d = 0.51$). Importantly, RD post-test scores were equivalent in the IARPA Video and No Pre-Test groups (Means = 70 vs. 71).

In terms of performance on the BE measures, watching the IARPA instructional video enhanced performance on the ANC BE posttest relative to watching the Control video [$F(1,76) = 4.44, p < .05$, and $F(1,76) = 5.56, p < .05$, with Focalism items excluded]. ANC BE scores improved in the IARPA instructional video condition (Cohen's $ds = .21 - .42$). ANC post-test scores were a bit higher in the IARPA Video group as compared to the No Pre-Test group (Means = 56 vs. 54), which may either be due to sampling error or pre-test sensitization. In addition, watching the IARPA instructional video enhanced performance on the REP BE posttest relative to watching the Control video [$F(1,94) = 9.13, p < .01$, and $F(1,95) = 9.20, p < .01$, with BRN "Tanks" item removed]. Specifically, REP BE scores increased by 9 points in the IARPA instructional video condition (Cohen's $ds = .82 - .86$), and, as shown in Table 45, 3 out of 4 facets revealed

statistically significant bias mitigation effects. REP post-test scores were similar in the IARPA Video and No Pre-Test groups (Means = 58 vs. 59). By contrast, there were no statistically significant differences between PRO BE pretest and posttest scores.

Table 44: ABC-2 Anchoring Bias (ANC), Representativeness Bias (REP), Projection Bias (PRO), and Recognition and Discrimination (RD) Pretest and Posttest Scale-scores.

Group	Mean ANC Scores		Mean REP Scores		Mean PRO Scores		Mean RD Scores	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Control Video	55 (10) N = 47	54 (10) N = 41	52 (12) N = 56	54 (10) N = 45	50 (10) N = 57	51 (9) N = 49	31 (29) N = 57	46 (30) N = 49
IARPA Video	54 (9) N = 48	56 (10) N = 47	50 (10) N = 60	59 (11) N = 53	54 (9) N = 60	55 (10) N = 55	31 (26) N = 60	70 (25) N = 56
No Pre-Test	-	54 (10) N = 42	-	59 (9) N = 50	-	57 (9) N = 49	-	71 (27) N = 49

Note. Values in parentheses are standard deviations. Each scale ranges from 0-100.

Table 45: Summary of ANCOVA Results for ABC-2 BE and RD Measures: Did the IARPA Video Group Do Better Than the Control Group, Controlling for Pretest Scores?

Bias Scale	P-Value for Condition (IARPA Video Versus Control Video Treatment Group)	Higher Score in Experimental Group Than in Control Group?	Effect Size (Partial Eta- Squared)
Anchoring	.04	Yes	.06
Representativeness	< .01	Yes	.09
Projection	.38	Yes	.01
Recognition and Discrimination	< .01	Yes	.28

Table 46: Summary of ANCOVA Results for Anchoring Bias Facets: Did the IARPA Video Group Do Better Than the Control Group, Controlling for Pretest Scores?

Anchoring Bias Facet	P-Value for Condition (IARPA Video Versus Control Video Treatment Group)	Higher Score in Experimental Group Than in Control Group?	Effect Size (Partial Eta-Squared)
Numerical Priming	.01	Yes	.07
Selective Accessibility	.70	Yes	< .01
Comparative Judgment	.66	No	.00
Self-Generated Anchor	.30	Yes	.01
Focalism	.90	No	.01

Table 47: Summary of ANCOVA Results for Representativeness Bias Facets: Did the IARPA Video Group Do Better Than the Control Group, Controlling for Pretest Scores?

Anchoring Bias Facet	P-Value for Condition (IARPA Video Versus Control Video Treatment Group)	Higher Score in Experimental Group Than in Control Group?	Effect Size (Partial Eta- Squared)
Base Rate Neglect	.02	Yes	.06
Base Rate Neglect (minus “Tanks” item)	.01	Yes	.06
Sample Size Insensitivity	< .01	No	.08
Conjunction Bias	.02	Yes	.06
Non-Random Sequence Fallacy	< .01	Yes	.11

Why was there little change in PRO posttest scores relative to pretest scores? In order to address this question, we re-examined the scoring and results for each of the facets of projection bias. In addition, we re-scored the Knowledge Projection facet items using an alternative approach that measures the extent to which people impute their knowledge (or lack thereof) onto estimates of others' knowledge. In contrast to the ABC-2, this approach does not make use of confidence ratings. This was also the same approach proposed by the BICC for scoring their so-called "Curse of Knowledge" items.

As shown in Table 48, there were no significant differences between PRO pretest and posttest facet scores. The only facets that revealed bias mitigation effects that approached statistical significance were FCE and Social Projection – Affective Attribution. Moreover, as shown in Table 48, we also did not find any errors with the current PRO scoring approaches and the alternative "curse of knowledge" alternative scoring approach for Knowledge Projection yielded similar results. Perhaps there may not have been sufficient information presented in the IARPA instructional video to mitigate the facets of projection bias measured in the ABC-2.

Taken together, our findings suggest that exposure to the ABC pre-test does not appear to have a practically significant effect on BE or RD bias measures. Consistent with our hypotheses and the findings from the ABC-1 Pretest Sensitization Study, the results indicate that participants were learning from the IARPA instructional video, but acquiring more explicit/declarative knowledge than procedural knowledge of the Phase 2 biases. Nevertheless, watching the IARPA instructional video did produce moderate to large improvements in ANC and REP scores, and the dissociation in performance between BE and RD bias measures may be considered discriminant validity evidence for the ABC-2, providing additional support for its construct validity.

Table 48: ABC-2 pretest and posttest knowledge projection composite raw-scores for IARPA Video, Control Video and No Pretest IARPA Video groups.

Condition	Mean Knowledge Projection Scores (w/Confidence Ratings)			Mean Knowledge Projection Scores (w/Course of Knowledge Approach)		
	Pretest	Posttest	<i>d</i>	Pretest	Posttest	<i>d</i>
IARPA Video	795.10 (106.82) N = 60	774.50 (128.68) N = 57	-.17	5.94 (1.44) N = 60	5.74 (1.43) N = 56	-.14
Control Video	748.00 (122.77) N = 60	760.70 (127.16) N = 51	.10	5.37 (2.20) N = 57	5.40 (1.31) N = 49	.02
IARPA Video, No Pre-Test	-	792.43 (115.54) N = 50		-	5.80 (.97) N = 50	

Table 49: Summary of ANCOVA Results for Projection Bias Facets: Did the IARPA Video Group Do Better Than the Control Group, Controlling for Pretest Scores?

Projection Facet	P-Value for Condition (IARPA Video Versus Control Video Treatment Group)	Higher Score in Experimental Group Than in Control Group?	Effect Size (Partial Eta-Squared)
False Consensus Effect	.11	Yes	.03
Knowledge Projection	.49	No	< .01
Alternate Knowledge Projection Measure, Following Curse of Knowledge Scoring Approach	.40	Yes	< .01
Social Projection-Affective Attribution	.09	Yes	.03
Social Projection-Personality Attribution	.70	Yes	< .01

3.7 ABC-2 Implementation

3.7.1 Development of Equivalent Forms

For the Phase 2 equating, we used linear rather than equipercentile equating methodology. In the case of linear equating, a linear transformation is chosen such that scores on the equated test forms correspond to the same number of standard deviations above or below the mean (Peterson, Kolen, & Hoover, 1993). The reason we used linear equating in Phase 2 is that implementation of equipercentile equating yielded scale-score distributions that were highly bi-modal, and that were markedly different than their corresponding raw-score distributions, which were all approximately normal. Use of linear equating methodology preserved the raw-score distributions and was, therefore, used in place of equipercentile equating.

Equated scores for the ABC-2 BE scales were standardized to a mean of 50 and standard deviation of 10.³⁵ This was done to ensure that all test-takers' equated BE scores were between 0 and 100. In addition, the 0-100 metric is one that is familiar to most test-takers and test users.

3.7.1.1 Completion Time for ABC-2

In order to meet the IV&V operational requirement that the ABC-2 scales take between 45 and 60 minutes for test-takers to complete, we analyzed the timing data from each task administered in the Field Test. We calculated the mean and median completion times, as well as the SDs, for each ABC-2 task, and identified additional tasks for removal, because they took too long relative to the amount of information they provided. Table 50 reports the mean, median, and SDs for ABC-2 completion times for each of the three primary ABC-2 test forms in both the Field Test and Retest studies.

Table 50: ABC-2 completion times for three primary ABC-2 test forms in Field Test and Retest studies.

Test	ABC Form 1	ABC Form 2	ABC Form 3
Field Test	Mean = 50.3 Median = 48.4 Stand Dev. = 14.2 N = 645	Mean = 53.8 Median = 51.6 Stand. Dev. = 15.7 N = 678	Mean = 53.4 Median = 52.3 Stand Dev. = 15.6 N = 649
Re-Test (1 month later)	Mean = 50.4 Median = 47.2 Stand Dev. = 15.7 N = 261	Mean = 52.9 Median = 50.2 Stand Dev. = 17.9 N = 177	Mean = 51.5 Median = 47.8 Stand Dev. = 17.2 N = 238

Consistent with findings from the ABC-1 Field Test and Retest studies, the median completion times were 1-3 minutes lower than the mean completion times. This was due to the influence of

³⁵ The equated RD scale scores were standardized using higher means and SDs than were used for the BE scale scores. This was done because the RD scale scores were somewhat skewed and peaked relative to the more normally distributed BE scale scores.

extreme outlier response times. As such, the median values are a more appropriate estimate of the average ABC-2 completion time. Second, test-takers took less time to complete the ABC-2 during the retest study conducted approximately 1 month following the Field Test.

3.7.2 ABC-2 User Manual

As in Phase 1, we created a User Manual for use/adaptation of the ABC-2 test battery to facilitate a smooth handoff of the ABC-2 to JHUAPL and any other future users of the ABC-2. The User Manual was intended to provide important information and useful guidance for implementation of the ABC-2 in the IV&V phase. The main purposes of this User Manual were to:

1. Describe the content of the ABC-2, and provide illustrative ABC-2 items/tasks;
2. Describe and explain the scoring process for the ABC-2 scales and overall battery score;
3. Describe test equating methodology to link ABC-2 scores across test forms; and
4. Describe data processing and syntax files created to score the ABC-2 forms.

Along with the ABC-2 User Manual, we included a deployment package to further facilitate the implementation of the ABC-2 in the IV&V phase of the project. The deployment package included:

1. Python scripts and associated files configured to process raw data files from individual test-takers and transform them into a single, master data set (.csv format).
2. SPSS syntax files, one for each ABC-2 scale. Each such file has the syntax necessary to compute all the group-level total-scores for the ABC-2 scales.

3.7.2.1 Identification and Resolution of Implementation Issues

Subsequent to delivery of the ABC-2 User Manual and deployment package, the IV&V team identified issues that were relevant to successful implementation of the ABC-2. First, at the request of JHUAPL, we investigated ways to modify SPSS scoring syntax for the ABC-2 BE scales such that standardized scale-scores would not change with additions to the sample. We chose to “hard code” the standardization of each item/task comprising the ABC-2 BE scales; that is, we standardized using means and SDs unique to each item/task derived from our field test sample. The same item-specific means and SDs were then applied to each test-taker in the IV&V sample. In addition, we corrected a few minor errors in the SPSS scoring syntax files identified by the IV&V team subsequent to delivery of the deployment package.

Second, at IARPA’s request, we provided separate linear equating formulas (1) for ANC with and without the Focalism facet, and (2) for REP with and without the “Tanks” Base Rate Neglect task. Third, we investigated the impact of differential weighting on key Phase 2 BE scale statistics, because differential weighting was suggested as a possible way to better represent the content domain of the BE scales by ensuring that facets with differing numbers of items were equally represented. Based on our analyses, we revised the SPSS scoring syntax files to use differential weighting.

3.8 Integrative Summary for ABC-2

Development of the ABC-2 essentially required the same steps as development of the ABC-1 (e.g., partitioning the content domain of the biases, using the best extant literature to suggest task prototypes). As such, we will not repeat them here. That said, there were aspects of developing and validating the ABC-2 that were not part of development of the ABC-1. To a large extent, this was a result of capitalizing on lessons learned in Phase 1; other differences were related to the nature of the bias constructs to be measured in Phase 2. As with Phase 1, various implementation issues emerged. These involved reexamination of scoring, weighting, and equating, and were addressed in consultation with the IV&V team. Many of the key findings from Phase 1 were replicated in Phase 2. Specifically, the RD scale was relatively unidimensional whereas the BE scales were not. The dissociation between RD and BE was also replicated in Phase 2. Finally, as in Phase 1, our bias mitigation study with the IARPA instructional video revealed that the RD scale and BE scales most saturated with declarative knowledge showed the greatest mitigation effects.

4 Software Platform for Computer-based Delivery of Tests

The ABC involves the use of a variety of different item types including multimedia and interactive elements that are not available in any existing COTS survey or test delivery platform. We therefore developed the ABC test administration platform to support the authoring and administration of the ABC test. The test administration software is used for presenting the test items to subjects, collecting their responses, and exporting the subject data. Additionally, the test administration platform includes a separate authoring tool that allows test authors to create items and tests, add items to a test, and edit existing items.

4.1 System Overview

The ABC test administration platform is implemented as a web application, allowing any user with a web browser to connect and use the software. Users may be test developers, test administrators, researchers, or test participants. For the purposes of the ABC-1 and ABC-2 test development as well as Sirius IV&V testing, the administration platform was hosted on a secure web server on the MITRE DMZ.

4.2 System Architecture

The test administration software consists of three components:

1. The Authoring Tool for creating tests and test items
2. The Test-taker Interface for administering the test and collecting responses
3. The Administrator Interface to manage user accounts and participant data

The back end database serves both as storage for the tests themselves and for the participant responses.

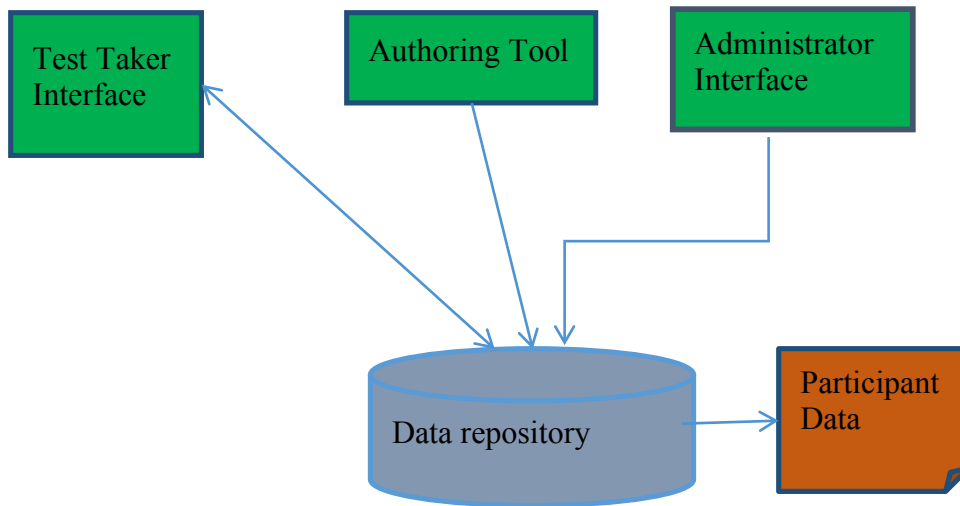


Figure 55: ABC delivery platform system architecture

Tests are created using the Authoring Tool, which provides a simple form-based interface for adding new items and creating the item content.

The Test-taker Interface is the interface that is used by participants taking a test. It pulls the test description from the database in order to display it to the subject, and writes the user responses back into the database.

Finally, the Administrator Interface contains several screens for administering different parts of the system. In order to use the Administrator Interface, a user must have a login account with a password. When logged in, the administrator is able to:

- launch a test-taking session
- add/create/delete user accounts
- enter the test authoring tool
- export participant data as a spreadsheet

User accounts in the Administrator Interface are associated with different roles, which are used to control access to the various parts of the system. The roles are: System Administrator, Editor and Research Assistant. Only System Administrators have permission to modify and create user accounts. Only System Administrators and Editors can edit tests.

4.3 Implementation

The ABC administration platform web application server is implemented in Java using the Spring web application framework, JavaServer Pages (JSP), and the Hibernate persistence framework. Application data is stored in a Postgres database. The client code is written in

Javascript with JQuery, Backbone, the Twitter Bootstrap library, and the Flowplayer library for playing video files.

4.3.1 Setup and Configuration

The ABC software platform requires a server with the following packages installed:

- PostgreSQL 8.1.23
- Java 7
- Tomcat 7.0

The database should be configured by creating a database with the name “abc_db” for the ABC system and loading the database schema included in the distribution. Additionally, the ABC WAR (Web Application Archive) file named “stiEditor.war” must be deployed on Tomcat. The WAR file includes all of the custom ABC code, and all necessary Java and Javascript libraries (including JQuery, JQuery-UI, Flowplayer media player, all necessary stylesheets, JSPs and the Java Spring Framework).

To connect to the system requires a client machine with a web browser. Any reasonably modern web browser will work, but Google Chrome and Mozilla Firefox are preferred. The browser must have an up to date Adobe Flash plugin installed in order to play videos and audio. The war file is configured to connect over HTTP on port 8080. To connect to the running web application, point the browser to: <http://server.host.name:8080/stiEditor>.

The database schema file creates an initial administrator account with the password “Changeme123!”. Upon logging in for the first time, the system administrator should change the password. They can then create accounts for any additional administrators or users who need access to the administrator interface.

4.3.2 Browser Requirements

- The ABC application works with a wide range of browsers on any operating system
 - Operating Systems: Windows, Linux, Mac OS-X
 - Browsers: Firefox 3.0 or higher, MS Internet Explorer 8.0 or higher, Safari 3.0 and higher, Chrome 5 and higher
 - Plug-in: Adobe Shockwave Flash 8 or higher.

4.3.3 Security and Authorization

- The instantiation of the test platform used during the Sirius program is deployed outside MITRE’s corporate firewall for remote access and as such must comply with all of MITRE’s relevant security requirements regarding server configuration and password strength.
- User accounts and roles: The platform provide four login roles: (a) System Administrator, (b) Site Administrator, (c) Test Editor and (d) Research Assistant. There is also a Participant role for test-takers, but they can only access the test-taker interface and they do not have to log in with a password. The following table summarizes the permissions for each user role.

Table 51: Access permissions for each role

	System Administrator	Site Administrator	Test Editor	Research Assistant
Create/Modify/Delete user accounts	Yes	Yes (for their site only)	No	No
Launch tests	Yes	Yes	Yes	Yes
Create/Edit tests	Yes	No	Yes	No
Export participant data	Yes	Yes (for their site only)	Yes	Yes (for their site only)

- All user accounts have a login id and password. Passwords must conform to the MITRE password guidelines: passwords expire after 120 days, passwords may not repeat any of the previous 12 passwords used, and every password must be at least eight characters and contain at least one number, one capital letter and one special character.
- Lab testing: A test administrator starts each test in the lab by entering the id of the participant (no passwords needed for subjects).
- Remote testing: Participants taking the test remotely may be provided a unique link to access their test session. The participant must enter the correct participant ID to access the test using this link. The link will expire after a specified period of time.
- If an expiration date is not required for the test link, remote participants can be simply given a link to the start page for a test. In that case, they are responsible for entering their participant ID correctly in order to receive credit.
- All authorization attempts and site activity are logged on the server.

4.3.4 Test-taker User Interface

- The test platform supports test items that have audio and video prompts as well as text.
- The test platform can accommodate several different graphic layouts for item content, using a split screen configuration to display information on the left and right sides, and using tables to lay out multi-part information.
- The system tracks and displays test time elapsed in the upper right corner of the screen.
- The system supports a range of question types (see below) including single choice (single-select multiple choice), multiple choice (multi-select multiple choice), Likert scales, semantic differential and forced-choice multiple choice. The full range of question types is described in Section 4.4.4.
- Questions may be required or optional. The test-taker cannot proceed to the next page until they have answered all required questions on the current page.
- The user interface does simple input validation on text entry fields, such as requiring numeric input for number fields, and making sure that user inputs add up to the requested total when required.

4.3.5 Test Authoring

- The authoring system supports authoring of all item and response types that can be included in a test.
- The authoring system includes the ability to edit and reorder existing test items.
- The ABC platform stores all available items in an “item bank”. Items from the item bank can be added to tests in any combination. A given item can be used in more than one test.
- The authoring tool includes the ability to copy whole items as well as copying individual pages, page stems, or questions.

4.3.6 Instrumentation and Logging

- The test platform records all participant behavior, including mouse clicks, text entry, interaction with multimedia elements, and answer submission.

4.3.7 Data Management

- The platform stores user account data as well as participant data in a single Postgres database.
- The test platform stores all of the data recorded from multiple test-takers over time. When the participant data is no longer needed, a button is provided for the system administrator to export the old data to a file for long-term storage and expunge it from the running system. On the MITRE DMZ installation, the database is backed up nightly.
- The platform saves the progress of a participant in real time so if there is a computer failure they can begin at the point they left off. When a participant starts a test with a participant ID that has taken that test before, they are offered the choice of starting the test at the point where the last session left off, or starting the test from the beginning.
- The system logs all errors that it encounters in the Tomcat logs.

4.3.8 Scoring

- The platform exports all participant data in a spreadsheet format for scoring purposes. See Section 2.6.4 and Section 3.7 for details on implementation of scoring procedures for the ABC-1 and ABC-2 respectively.

4.4 Data Model

This section provides an overview of the data model used by the ABC test administration platform to represent tests and user responses. The structure of this data model is reflected directly in the structure of the underlying database tables.

4.4.1 Top Level Data Model

In the ABC data model there are two top-level data structures: tests and items. Each test has a unique ID, a name and a status indicating whether the test is under development or deployed.

Associated with each test is an ordered list of *Items*. An Item can be included in multiple tests, as indicated by the many-to-many mapping. Each item is categorized according to a Bias-Paradigm-Task hierarchy which is used in the authoring tool to organize and browse items.

A test item consists of an ordered list of *Pages*. A page can be split into two *HalfPages*, or remain as a single page. Pages and HalfPages contain an ordered list of *StemParts* containing text and multimedia elements that serve to introduce the item, followed by an ordered list of *Questions*.

4.4.2 Data model for stem parts and questions

Stem parts and questions are what make up the content on each page of a test. Stem parts contain the text and multimedia prompts that introduce the question content to test-takers and instruct them on how to answer the questions. Questions are the interactive elements of the page that ask the test-taker to make some kind of response.

Both stem parts and questions have several sub-types that require different information to define their content. The details of each is given below.

4.4.3 Data model for stem parts

There are 11 stem part types that are used to display instructions and information related to test items. Each of these stem part types has its own properties. Each stem part contains a *Content* property which is used to store the basic content that will be displayed for that stem part.

- **Text Stem Parts** There are four different ways of formatting and displaying blocks of text.
 - **Text** stem parts just display the text in their *Content* field.
 - **Highlighted Text** displays the content in a brightly colored box, to draw the user's attention.
 - **Directions** are used to instruct the test taker on what needs to be done with the item. They have a gray background and always include the word "Directions" at the beginning.
 - **Scrollable Text Area** is used when large amounts of text are to be displayed. In this case the text is displayed in a fixed-size text area with a scrollbar.
- **Multimedia Stem Parts** The *Content* property of these stem parts is the internal ID number of the image in the ABC database.
 - **Image** stem parts display a static image.
 - **Audio** stem parts display an audio player widget which the test taker can click to hear the audio file.
 - **Video** stem parts display a video player widget which the test taker can click on to view the video. The "Submit" or "Continue" button will be disabled until the user watches the video to the end.
- **Dynamic Stem Parts** These stem parts change as the user interacts with the item.

- **Score Counter** stem parts are used in conjunction with questions that require the user to select or click on multiple items. The score counter can be configured to count up or down, and can use either points or currency values (e.g. dollars) as its units.
- **Timer** stem parts are used to impart a sense of time pressure for an item. The timer has a *Duration* property which specifies the amount of time to allow, and can be configured to count up or down. When a page with a timer is loaded, the timer immediately begins displaying the count of seconds on the screen. When the time runs out, all of the stem parts on the screen are hidden and the test taker must enter responses to all questions in order to proceed.
- **Complex Stem Parts**
 - **Table** stem parts display multi-part information. Each cell in a table may contain an image or a block of text.
 - **Conditional Estimate** stem parts depend on the answer that was given to an earlier numeric entry question in the test. The conditional estimate stem part has a *Question* property that points to the question it is conditioned on. It also has a list of *Baseline* numbers that are used to compute the output. When the conditional estimate stem part is displayed, it calculates the numbers to display based on the answer that was given to the earlier question. The options in a conditional estimate stem part include whether to display the numbers in floating point format or as a four digit year, and whether to multiply or add the earlier numeric response to the baseline numbers.

4.4.4 Data model for questions

Each Question requires different information to define its content, depending on the question type:

- **Multiple choice** questions include an ordered list of *MultipleChoiceOptions*, each of which represents a response to the multiple choice question and includes some text to be displayed for that option.
- **Likert** questions include an ordered list of *ColumnLabels*, one for each Likert option, as well as optionally an ordered list of *RowStartLabels*, which may be used to describe the Likert scale being presented, if multiple Likert scale questions are to be displayed as a single table. (A table containing multiple Likert scale questions is considered a single *question* in our model, although it represents multiple participant responses).
- **Semantic differentials** do not include column labels, but must include ordered lists of both *RowStartLabels* and *RowEndLabels* of which there must be the same number, and *numColumns*, an integer denoting how many “points” the scale should contain.
- **Text entry** questions include a *numChars* field for the number of characters to display.
- **Numeric entry** questions may include an optional *minValue* and *maxValue* field to be used to validate the entries, as well as a *numChars* field to define the size of the entry field.
- **Point Distribution** questions require the test-taker to distribute a given number of “points”, recorded as *MaxPoints*, among a set of options (*PointItems*). For example, the question might

ask the test-taker to specify the probability out of 100% of each of five possible outcomes given a description of a situation.

- **Forced-choice** questions present the test-taker with a list of options, stored as *RowStartLabels*, and a set of labels, stored as *ColumnLabels*, and ask the test-taker to assign each label to the option that best matches it.
- **Wason** questions include a list of *WasonItems*, each of which has two parts, *sideA* and *sideB*, as well as a *startSide*, which indicates which of sideA or sideB starts out being displayed. For each of the Wason items, the test-taker can click on the hidden side to reveal the information it contains.
- **Info Item** questions contain a set of *InfoItems*, which are links that will be presented to the test-taker in order to retrieve additional information. Each InfoItem includes an *infoTitle*, which is the text that the user will click on to retrieve the information, and an *infoText*, which is the text that is displayed when the user clicks on the title. There is an optional *delay*, which if present will cause the test-taker to have to wait a given amount of time before the information is displayed.

4.4.5 Test tracker for participant responses

Participant responses are stored in the database using the *TestTracker* data model. A test tracker is associated with a pair of *participantId* and *testId*. To record participant activity, the test tracker includes a list of *TestEvents*, which records the id of the answer selected by the participant, and associates that answer with the question being answered.

TestEvents have several possible types:

- **MouseClicked**: When a mouse is clicked on a button or link
- **TextEntry**: When text is entered into a text field
- **QuestionAnswer**: When an answer to a question is submitted
- **PageLoad**: When a new page is loaded
- **VideoEvent**: When a video is started or stopped
- **TestStart**: When a test is started
- **TestRecover**: When a test that had previously been started is re-started at the point where the test-taker left off.

All of these events are recorded and stored in the database while a test-taker is taking a test.

Most of the TestEvent types are very simple, with the exception of the QuestionAnswer type, which records all of the information associated with the test-taker's answer to a question. Because several of the question types have multiple parts, a QuestionAnswer object contains a list of Response objects, each corresponding to the response to one of the parts of the question. In the case of a MultipleChoice question, each Response represents one of the options selected. In the case of Likert, SemanticDifferential, and ForcedChoice questions each Response represents a (row label, column label/column number) pair.

The contents of a Response varies depending on the question type:

- For single choice and multiple choice questions, the response contains the ID of the multiple choice option that was selected
- Semantic differential, Likert and Forced choice responses contain the ID of the row being answered (because these question types can contain multiple rows, each of which represents a different response), and the number of the column that was selected.
- Text responses contain the text that was entered
- Numeric responses contain the number that was entered
- Point distribution responses contain the numeric value that was entered for the item.
- WasonResponse contains the Wason item that was clicked on
- Info item response contains the info item that was clicked on

4.5 Data Export Format

Participant data is exported in a Zip file containing individual CSV files for each participant session. The columns in the CSV file are:

- Tracker ID: The unique session ID for the session
- Time: A human readable timestamp, in date format
- Test ID: The unique id of the test form being taken
- Participant ID: The ID entered by the participant at the start of the session
- Timestamp: The number of milliseconds since the start of the test session
- Event Type: One of MOUSE_CLICK, TEXT_ENTRY, ANSWER, PAGE_LOAD, VIDEO_EVENT, TEST_START, or TEST_RECOVER
- Object ID: The unique ID of the object that was interacted with. Depending on the event type the object type will vary. For mouse click it is the ID of the button or field that was clicked. For question answers it is the ID of the question that was answered.
- Item: The index (0-based) of the item within the test that a question answer event is part of.
- Item Name: The name of the item that a question answer event is part of.
- Page: The index (0-based) of the page within the current item that the test event took place on.
- Question: The index of the question (1-based) within the current page, for question answer and mouse click events.
- Question Type: The type of question being answered.
- Row: The index of the row (1-based) within the question that the event is recording, for Likert, Semantic Differential, and Forced Choice questions.
- Response: The entry that the user made, including mouse clicks, text and numeric entry, and question answers. The response is formatted as follows:

- Single choice and Multiple Choice: response contains the number of the option selected
- Text entry: response contains the text entered
- Numeric entry: response contains the numeric value entered
- Likert and Semantic Differential: response contains the number of the column selected
- Info items: response contains a semi-colon separated list of numbers corresponding to the info items that were selected
- Point distribution: response contains a semi-colon separated list of numbers that were assigned to each point distribution option.

Table 52: Example response data export file

Tracker ID	Time	Test ID	Participant ID	Timestamp	Event Type	Object ID	Item	Item Name	Page	Question	Question Type	Row	Response
5533		244		3652797	ANSWER	83572	25	FC-BFI-v2	8	1	forcedChoice	2	1
5533		244		3652797	ANSWER	83573	25	FC-BFI-v2	8	1	forcedChoice	3	2
5533		244		3652797	ANSWER	83574	25	FC-BFI-v2	8	2	forcedChoice	1	1
5533		244		3652797	ANSWER	83576	25	FC-BFI-v2	8	2	forcedChoice	3	2
5533		244		3653021	PAGE_LOAD		25		8				
5533		244		3660659	MOUSE_CLICK	0	26		1	1	NumericEntry		0
5533		244		3663306	MOUSE_CLICK	1	26		1	2	SingleChoiceOption		1
5533		244		3671908	MOUSE_CLICK	5	26		1	3	MultipleChoiceOption		5
5533		244		3674250	MOUSE_CLICK	2	26		1	4	SingleChoiceOption		2
5533		244		3678370	MOUSE_CLICK	5	26		1	5	SingleChoiceOption		5
5533		244		3688751	MOUSE_CLICK	6	26		1	6	SingleChoiceOption		6
5533		244		3697325	ANSWER	820296	26	Demographics-AMT-v2	1	1	numericEntry		50

•

4.6 User Interfaces

4.6.1 Administrator UI

The administrator UI is used for configuring the ABC test platform settings, launching tests and editing tests, and exporting response data from the database. An administrator account is required to login to this interface (see Section 4.3.3).

The administrator interface has four main sections, organized into tabs at the top of the screen.

- The Test tab presents a list of the tests currently in the database, with links to either launch or edit each test. The listing shows the name, ID, type, and status of each test. From this interface, the administrator can launch a test by clicking the “Launch test” link. The participant login screen will then appear and the test participant can enter their ID (assigned by the administrator) and begin the test.
- The User tab lists the registered user accounts on the system and includes links for the administrator to add, modify and delete accounts.
- The Site tab lists registered sites and includes links for the administrator to add, modify and delete sites.
- The Participants tab lists all participant sessions recorded in the system and provides links to export the full data or a summary of the data. The user of this screen can select individual participant sessions or a group of sessions by shift-clicking the mouse. They can then export the data for the sessions they have selected. The participants tab also provides a dialog to generate participant links for remote participant sessions, as well as a link to archive and expunge old data from the database.

4.6.2 Test-taker interface



Figure 56: Participant login screen

The test-taker interface is the interface that is used to display the test and record participants' responses.


A participant first logs in to begin taking a test, using the ID that was given to them by the test administrator (Figure 56). After pressing the Begin button, the first page of the test appears.

Each page optionally begins with *page stem*, which is a combination of text and multimedia content followed by a list of questions. Some pages only contain stem content and do not include any questions. In that case, the test-taker simply clicks the "Continue" button to proceed to the next page. If there are questions, the participant must select a response for all non-optional questions before pressing the submit button. Clicking submit will cause the selected response to be stored in the database and the test UI will move to the next page.

Test 2 of 10 Test Time Elapsed 0:01:56

Below is a picture showing one person helping a friend prepare for a job interview by asking her a series of questions. The person in the role of the questioner is asking questions that she has made up on the spot.

Directions: Listen to the audio clip. You will NOT be able to listen to the audio clip again after you have clicked on continue.



[Click to Play Audio](#)






Figure 57: Page with Text, Direction, Image, and Audio stem parts. (No Questions)

Test 2 of 10 Test Time Elapsed 0:04:17

Directions: Answer the questions below.

Questioner
Answerer

Compared to the average college student,

	Much less intelligent	Less intelligent	Neither less nor more intelligent	More intelligent	Much more intelligent
how intelligent is the questioner?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
how intelligent is the answerer?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>

Who would you say is more intelligent?

Definitely the answerer		←—————→					Definitely the questioner	
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	




Figure 58: Direction and Table (with image and text) stem parts, Likert and Semantic Differential questions

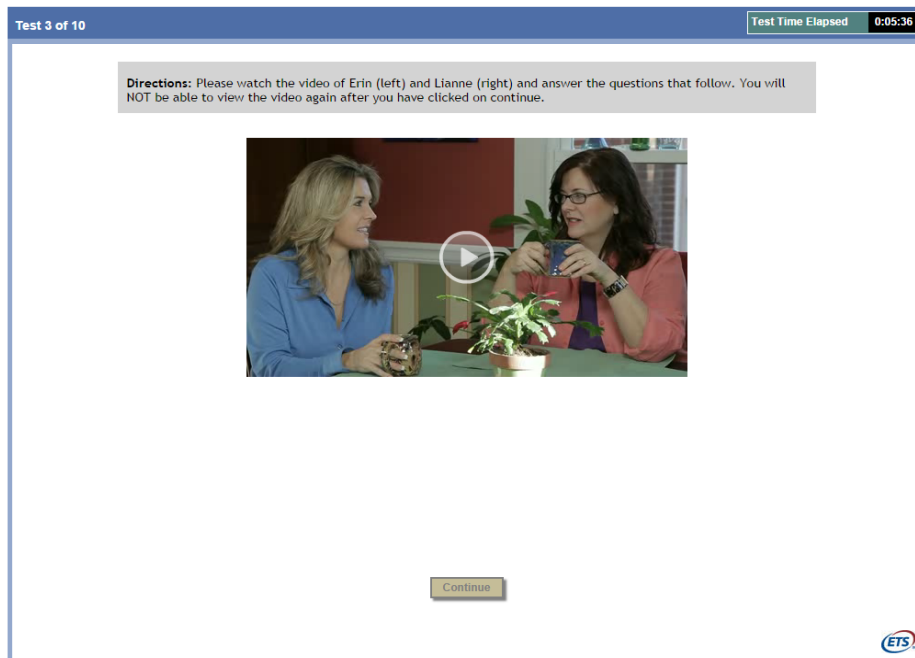


Figure 59: Directions and video stem part. The test taker must watch the entire video before the Continue button becomes enabled

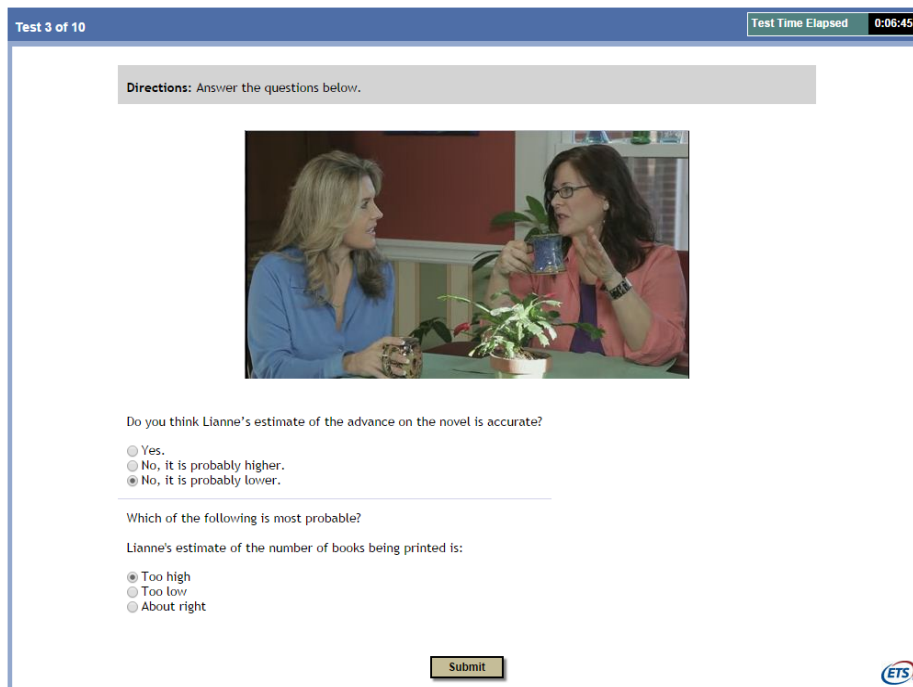


Figure 60: Direction and Image stem parts and two Single Choice questions

The test item count is presented in the upper left corner of every screen, while the amount of time elapsed is displayed in the upper right.

The test interface requires the test-taker to interact with the material on each page before moving to the next page. When a page loads, initially the Submit/Continue button is disabled. It will be

enabled when all non-optional questions on the page are answered, as well as all multimedia clips played through at least once.

4.6.3 Test authoring tool

The ABC authoring tool is a browser-based graphical interface that is used to enter tests and test items into the test delivery platform.

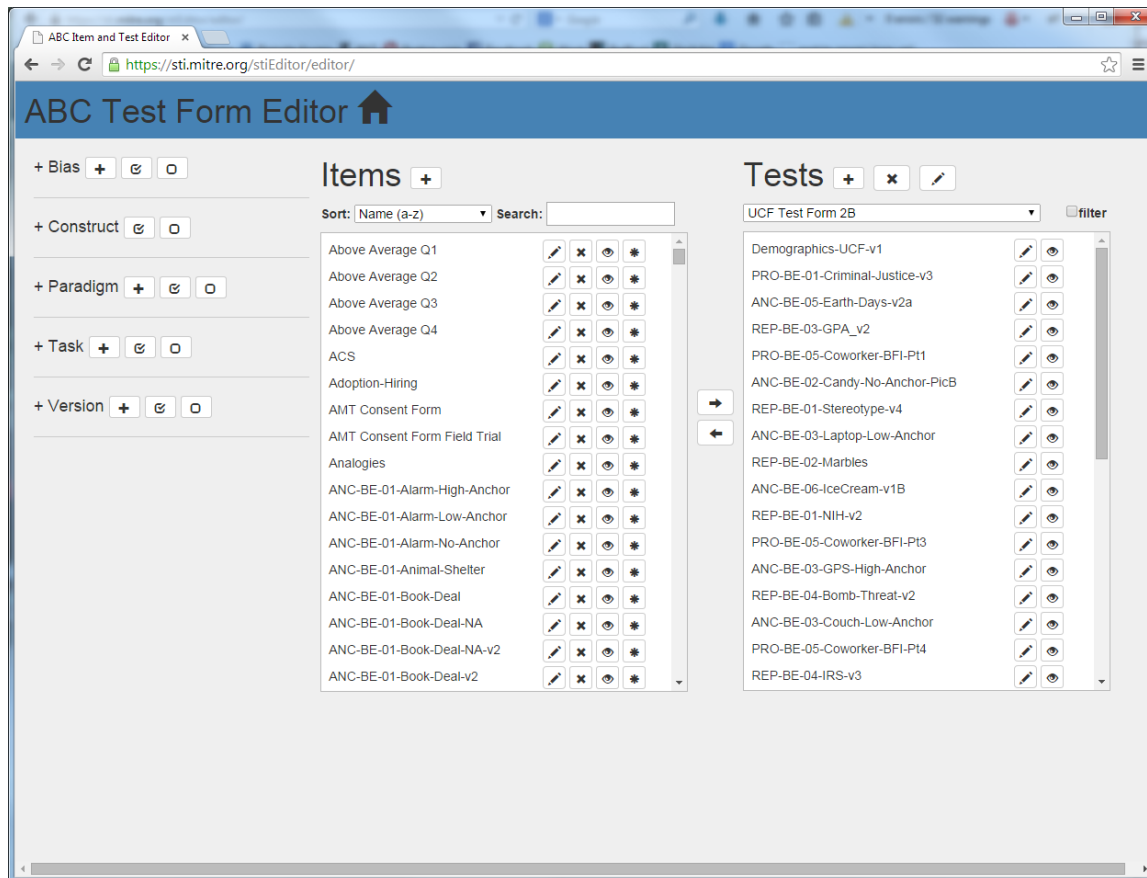


Figure 61: Test Editor start page

The authoring tool may be entered from any test administrator account. On entering the authoring interface the test author can choose between creating a new item or test, or editing any of the items or tests that exist in the database. Items can be edited, deleted, previewed, or copied using the respective buttons next to the name of the item in the list. Items can be added to or removed from a test by selecting the item and pressing the left or right arrow.

4.6.4 Editing an Item

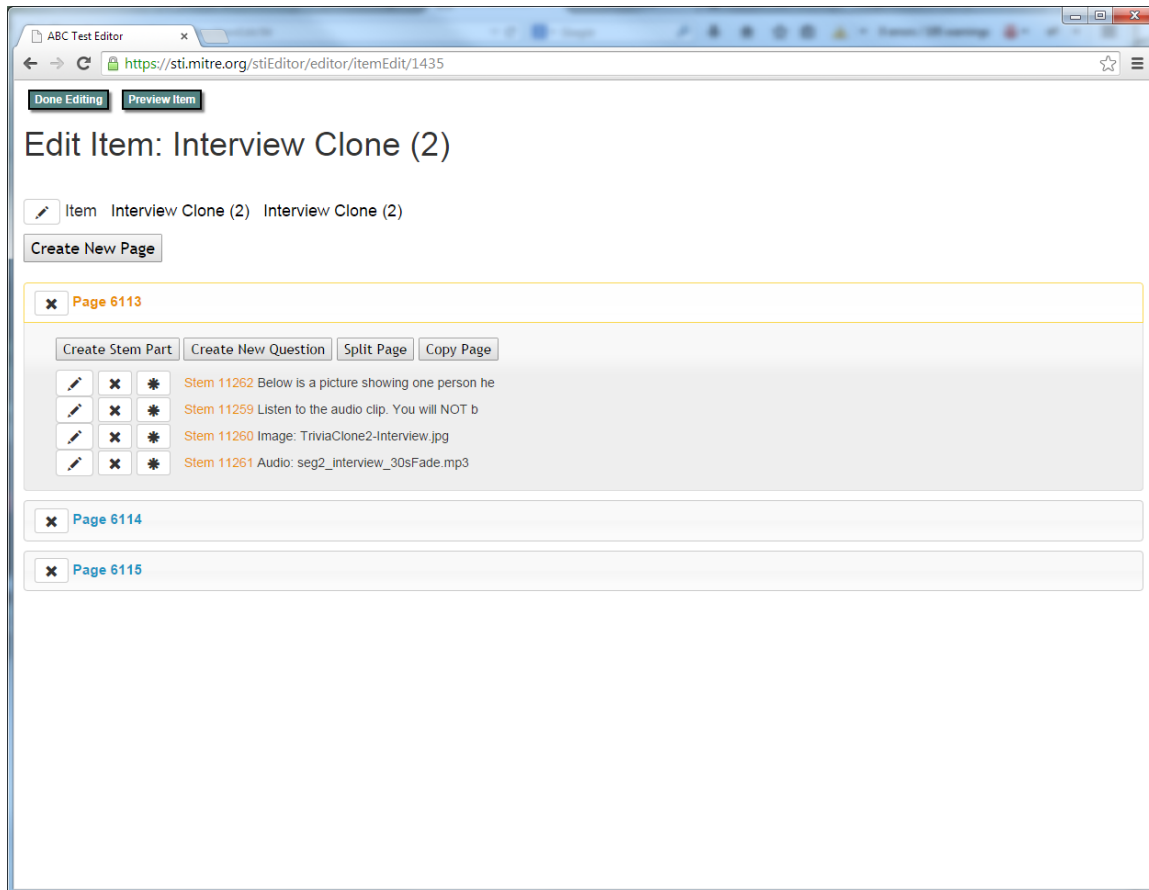


Figure 62: Editing an Item

An item contains an ordered set of pages. Pages contain an ordered set of stem parts and an ordered list of questions. The editing interface allows test authors to create, edit, delete, and reorder all the components of an item.

4.6.5 Editing Stem Parts

There are a number of different types of stem parts that can be added to a page, including Text, Highlighted Text, Directions, Scrollable Text, Table, Timer, Image, Video, Audio, ScoreCounter, and Conditional Estimate. For each of these types, the authoring tool provides a form for entering the relevant information for that stem part type.

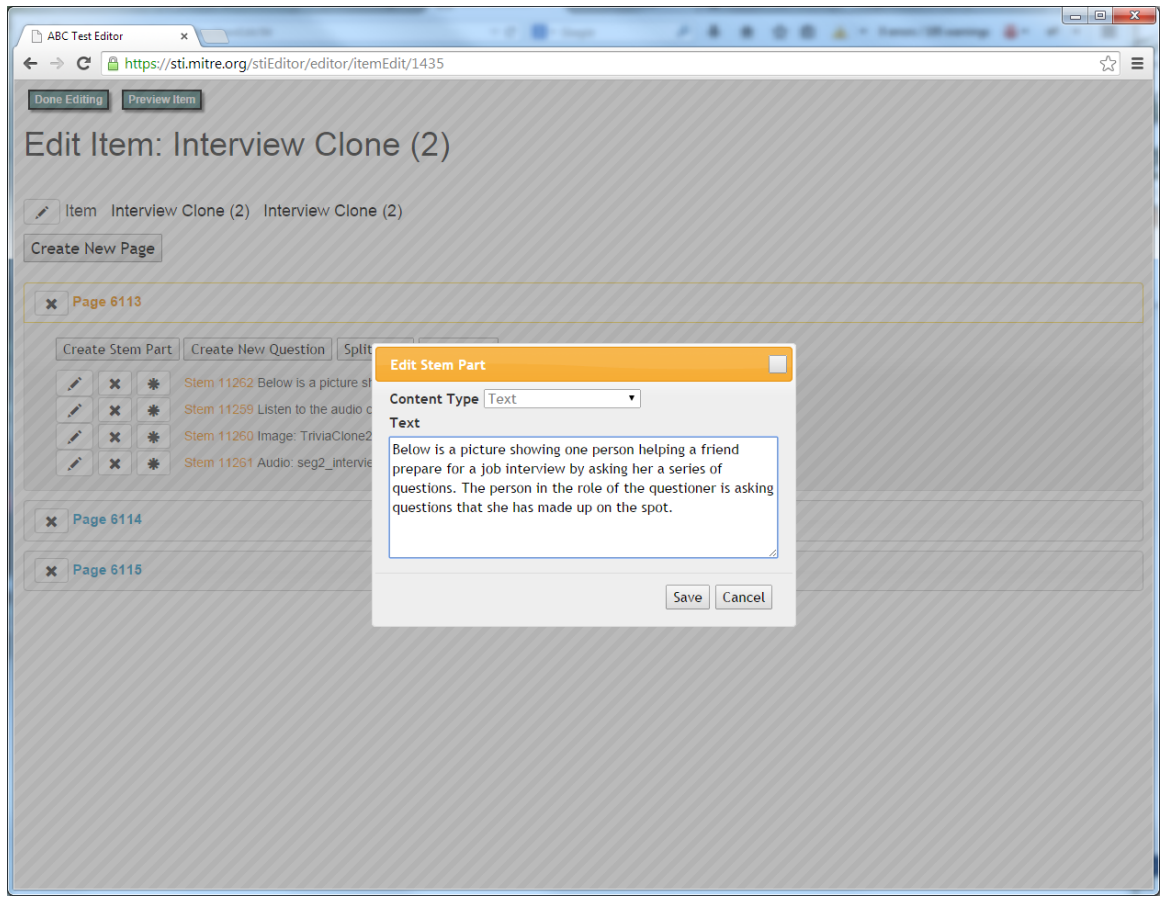


Figure 63: Editing a Text stem part

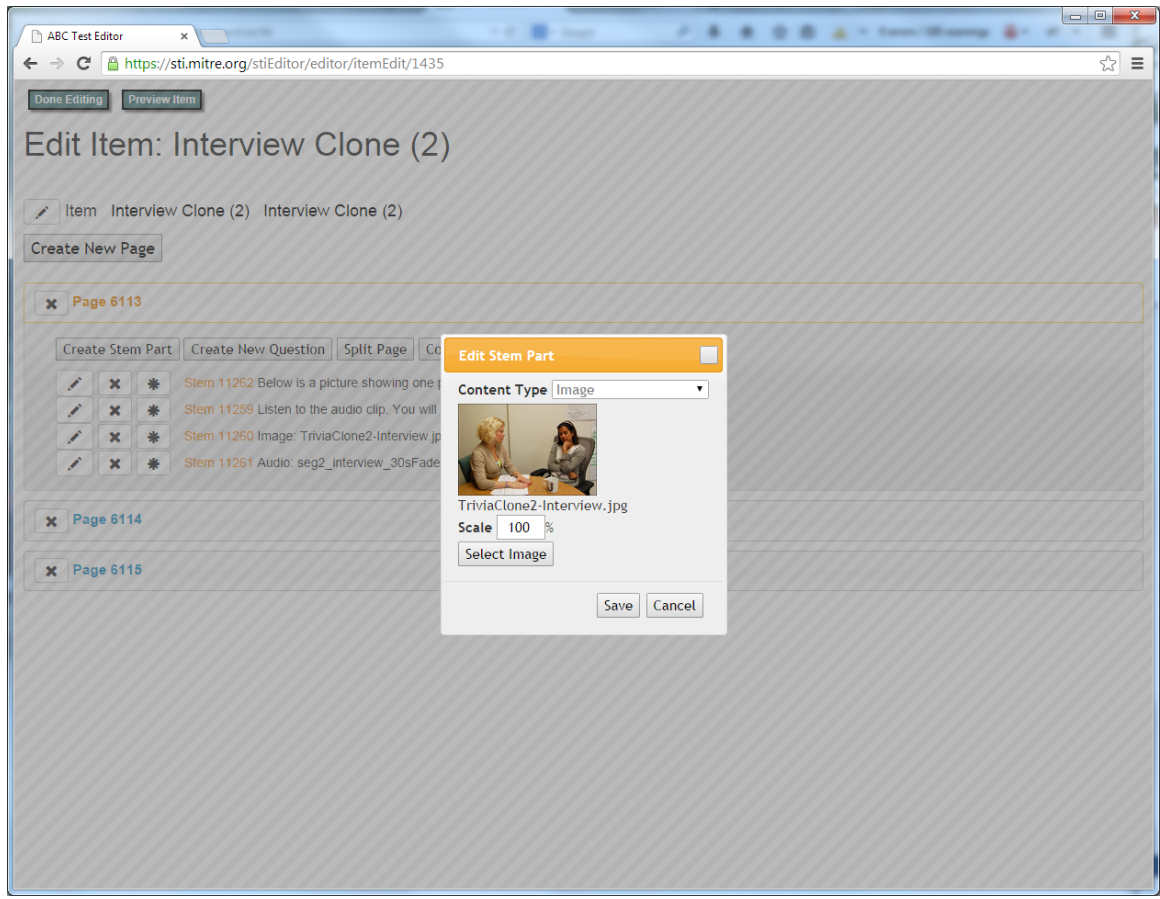


Figure 64: Editing an Image stem part

The user can set the scale of an image to reduce its size when it is displayed during the test. The size of the image is not changed in the authoring tool.

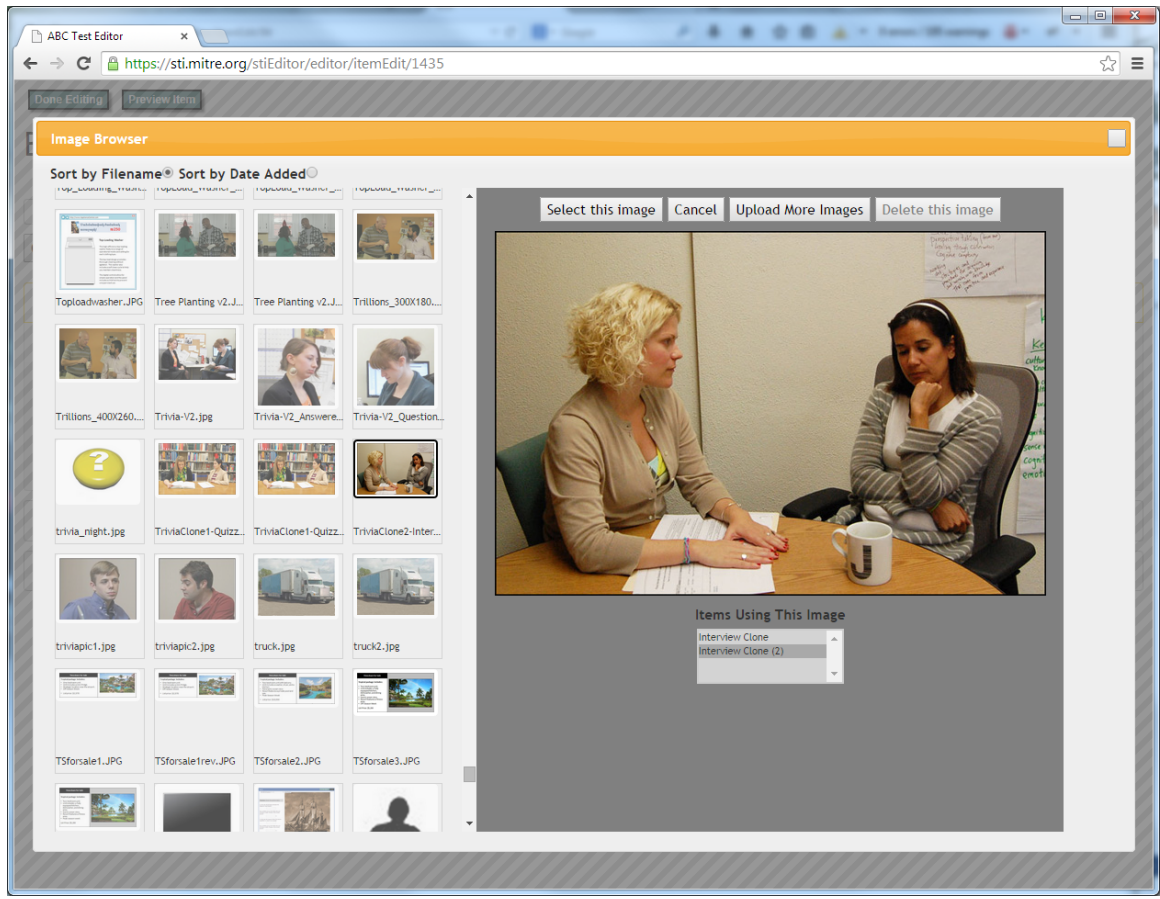


Figure 65: Image Browser

The user can select from the images already included in the database or upload new images.

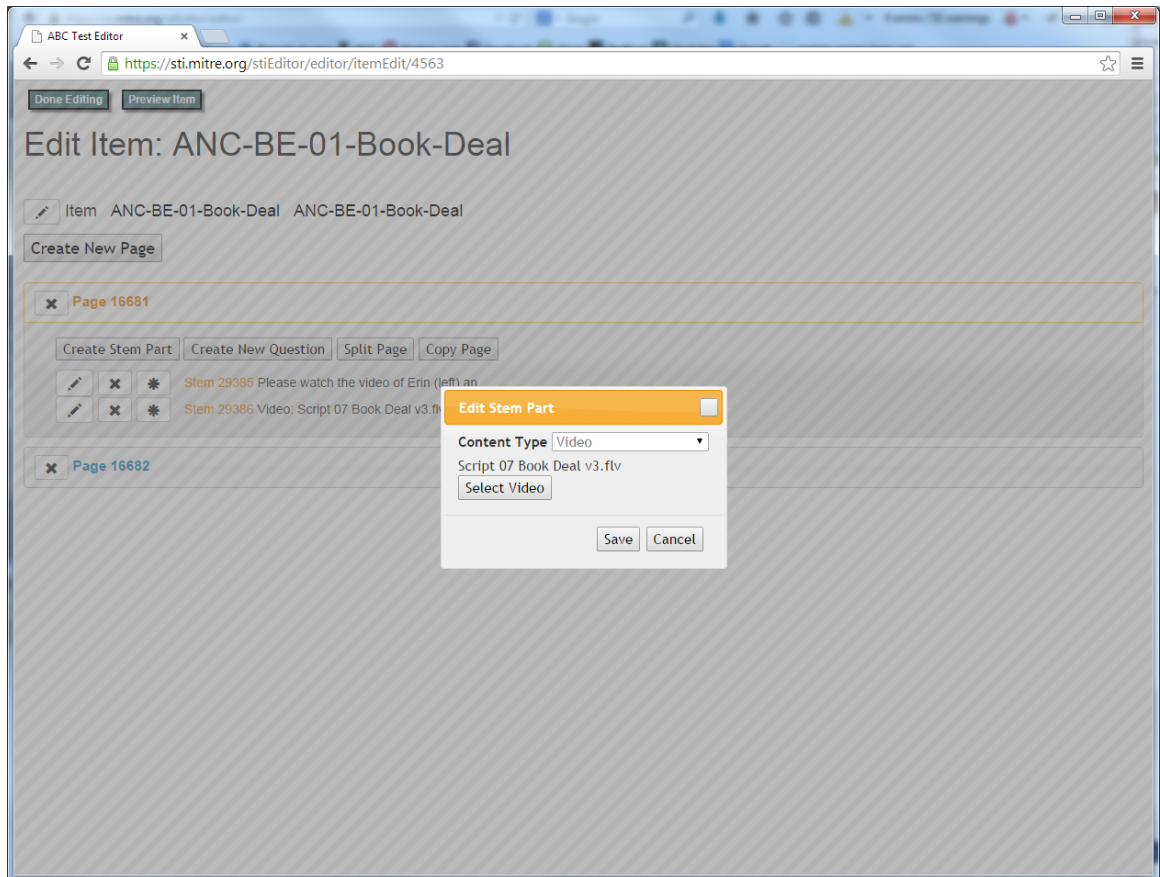


Figure 66: Editing Video stem part

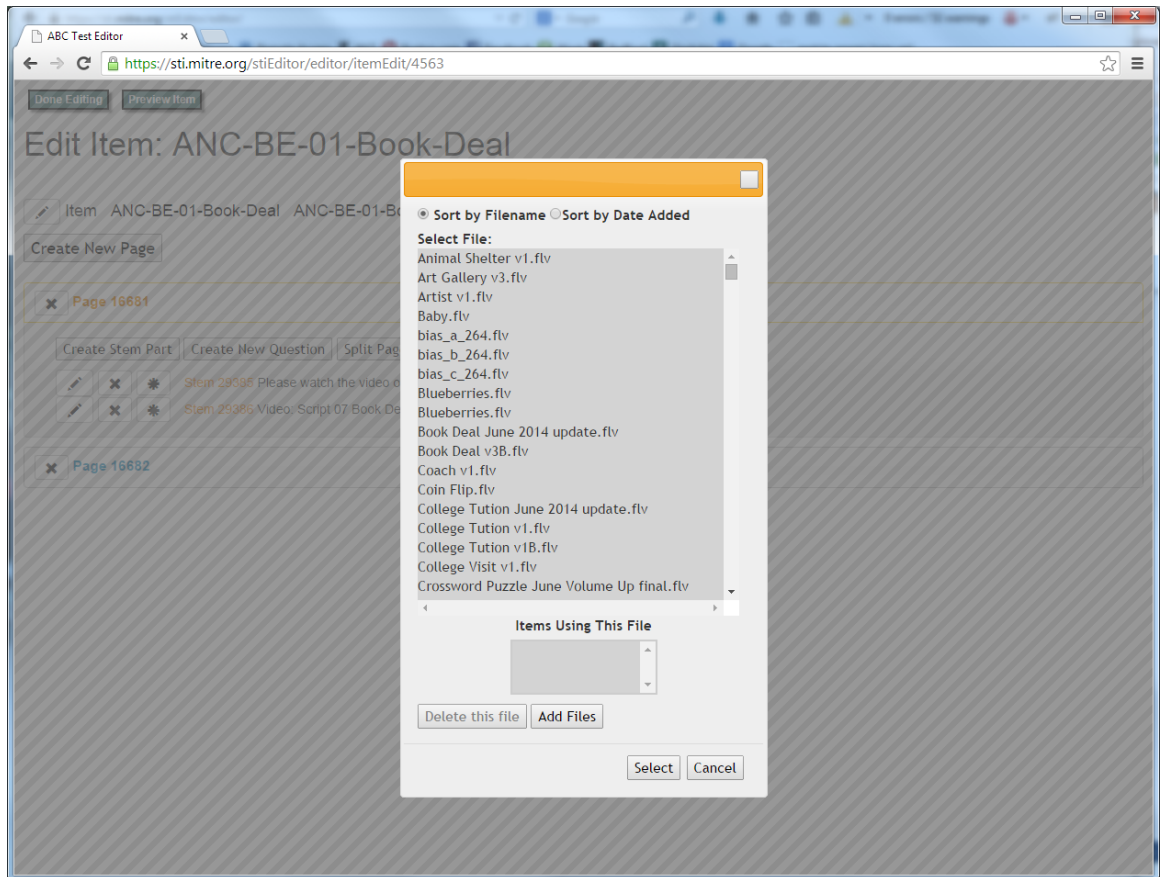


Figure 67: Video file browser

The user can select from the video files already stored in the database or upload new ones. A similar interface is provided for audio files.

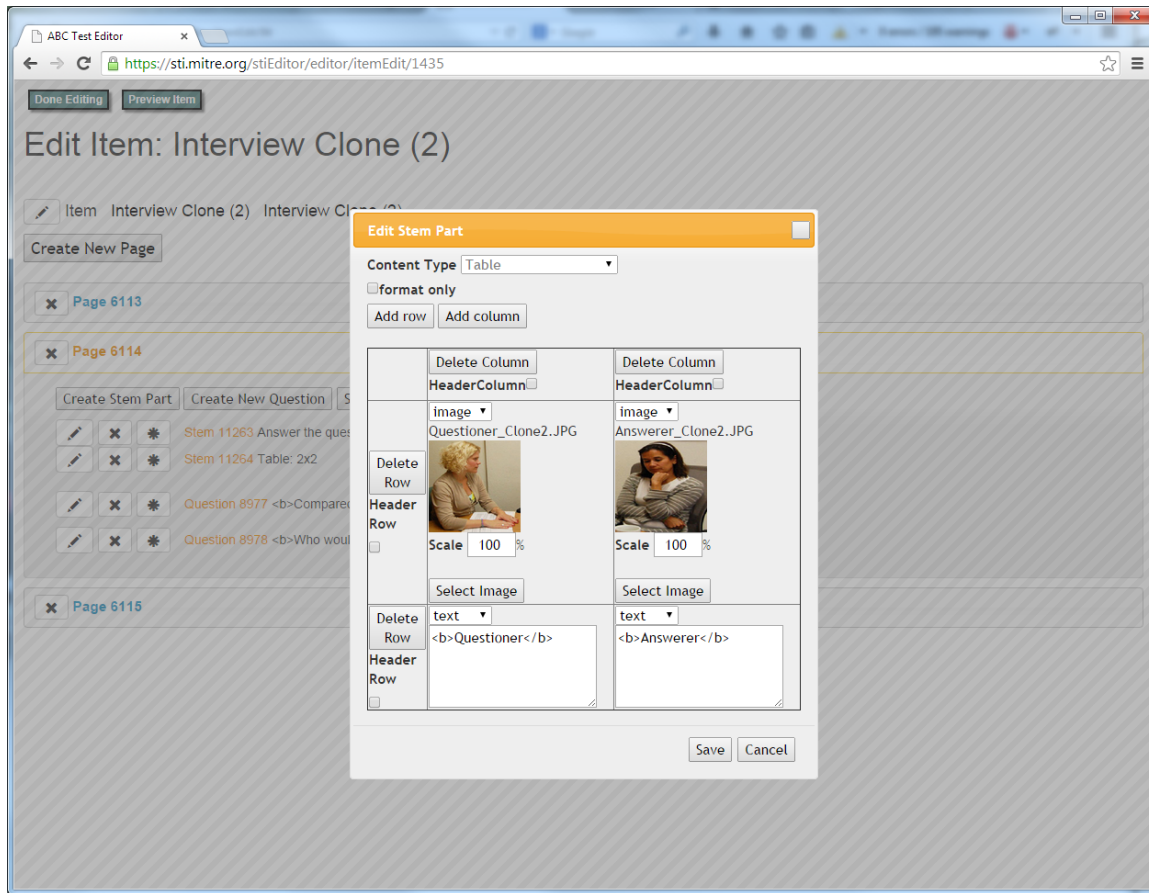


Figure 68: Editing Table stem part with images and text

4.6.6 Editing Questions

There are a number of different types of questions that can be created, including Single Choice, Multiple Choice, Likert, Semantic Difference, and Text or Numeric Entry. For each of these question types, the authoring tool provides a form for entering the relevant information for that question type.

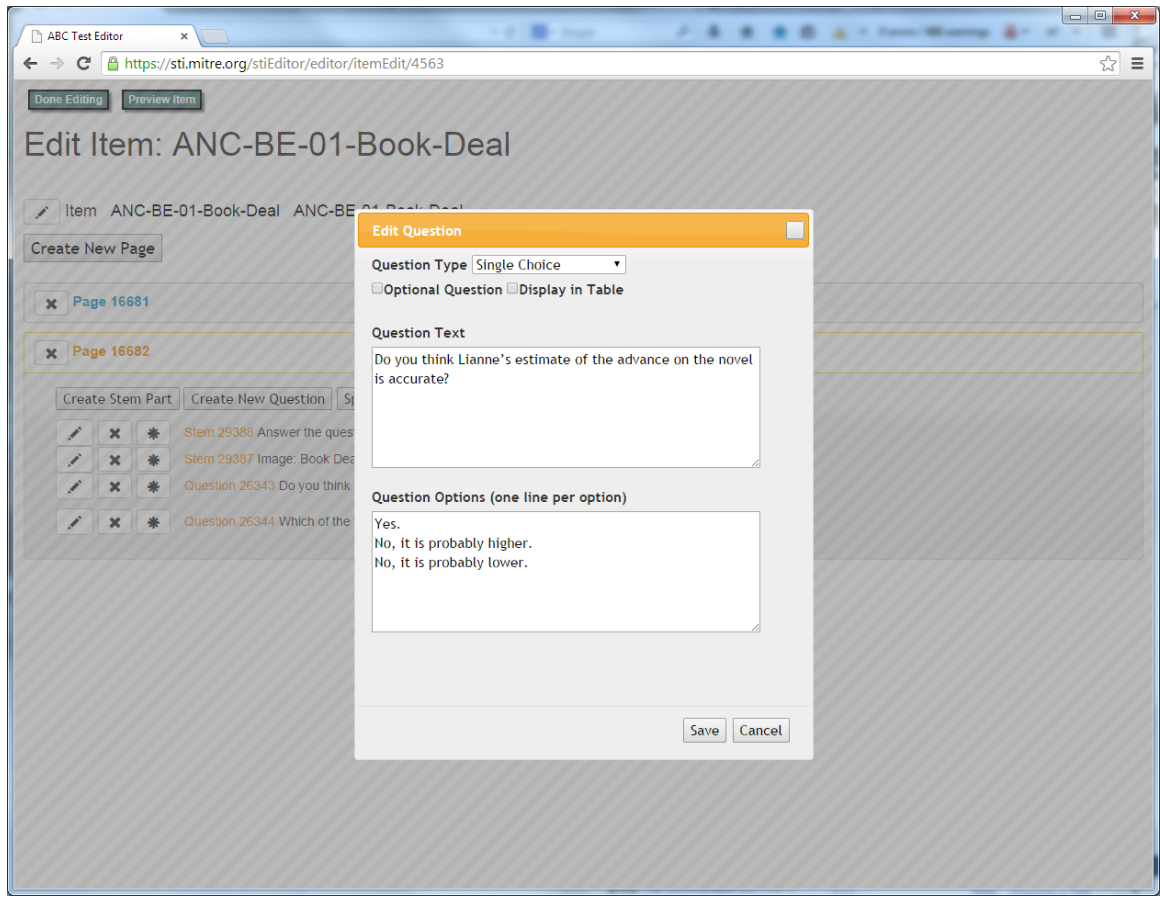


Figure 69. Editing Single Choice question

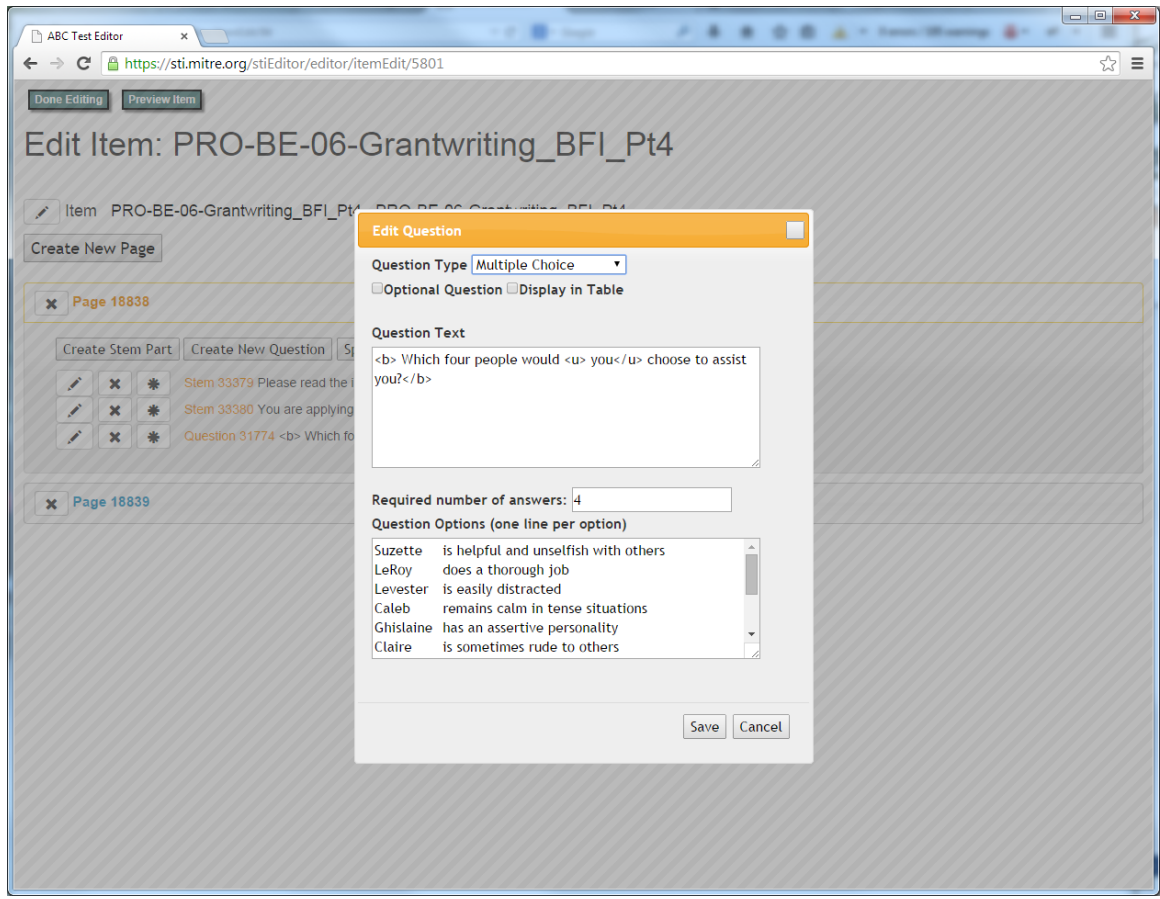


Figure 70: Editing Multiple Choice question

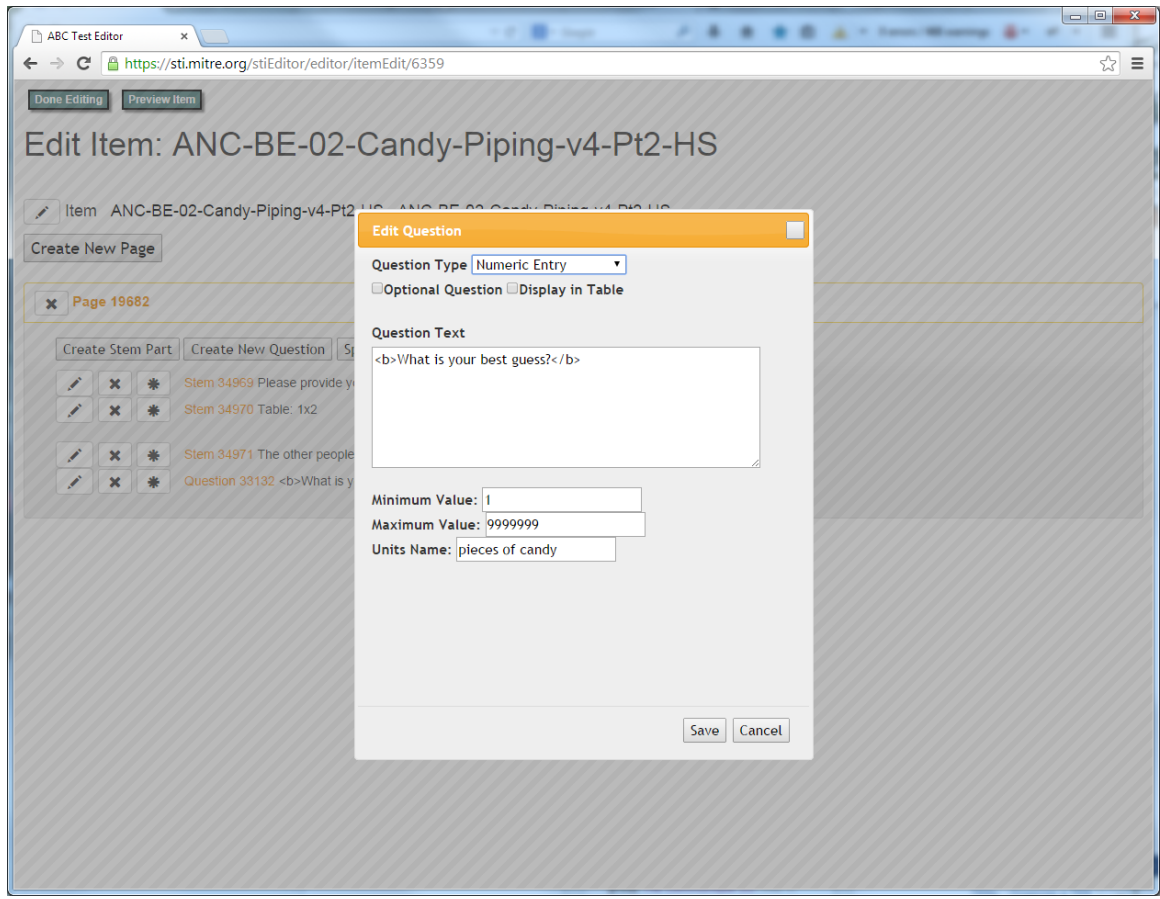


Figure 71: Editing Numeric Entry question

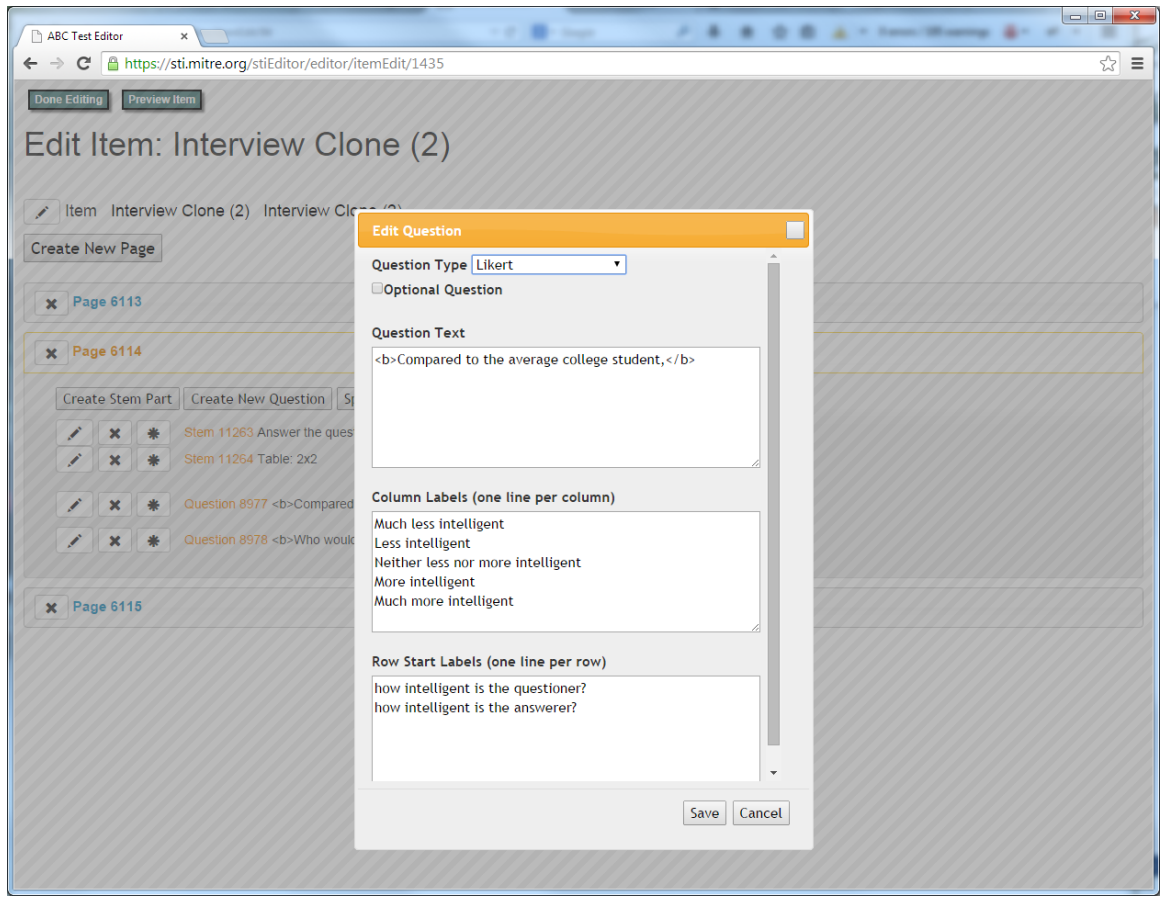


Figure 72. Editing Likert question

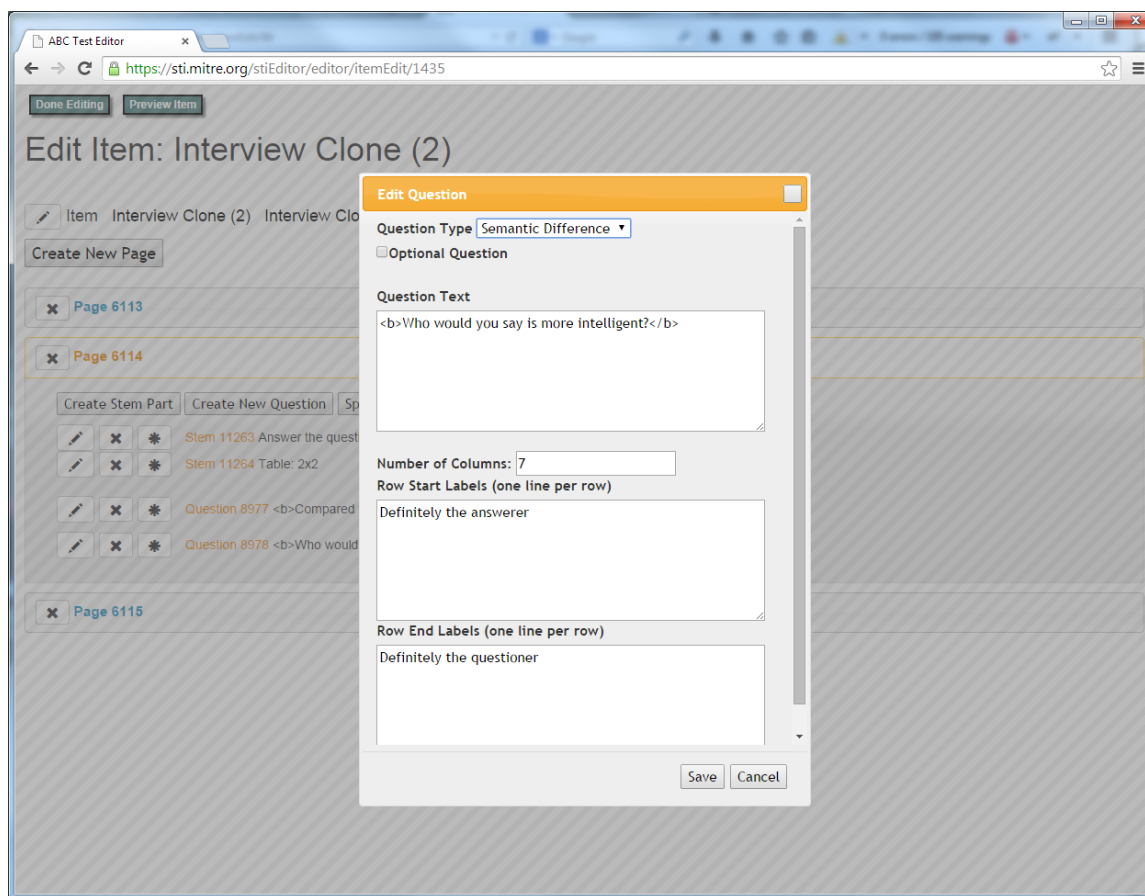


Figure 73. Editing Semantic Difference question

5 Validity Argument and Evidence for the ABC

According to the AERA/APA/NCME *Standards*, “Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests... The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations” (2014, p. 11). It should first be noted that much of the research encompassed by this project was necessarily exploratory. The constructs targeted for measurement were not well understood, especially from the standpoint of individual differences, nor were there any “gold standard” measures to facilitate evaluation of convergent validity within the domain of bias susceptibility or knowledge, or interventions to facilitate evaluation of sensitivity to bias mitigation. In addition, the nature of the behavior elicitation constructs was such that item development often required context and non-transparency. Indeed, during much of the project, we did not even have a basis for knowing whether the targeted bias constructs were unidimensional, multi-dimensional, formative, or simply epiphenomena of other constructs. As such, it was not even clear what the appropriate reliability metrics should be. This background should inform --in fact is critical to -- our validity argument.

According to the *Standards* (2014), the most critical source of validity evidence is not whether a test is valid in general, but whether it is valid for its intended use. In this project, the ABC’s

immediate intended use was not to simply measure its targeted constructs, but to detect changes in a group's standing on those constructs when they occurred.

Our validity argument, informed by our literature review (Gertner et al., 2013), is grounded in specific definitions of the targeted constructs, including facets making up those constructs to help ensure comprehensive measurement of the construct domains; and evaluation of a set of propositions logically and/or empirically implied by the validity (or lack of validity) of the ABC for its intended use. Each proposition is accompanied by a rationale for its inclusion. After stating each proposition and its associated rationale, we evaluate evidence regarding the degree of empirical support for each proposition from this project.

5.1 Propositions

In this section of the report, we assert several propositions that, if true, would support the validity of the ABC for its intended use. For each proposition, we provide a rationale and evaluate relevant evidence.

5.1.1 Proposition 1

The constructs targeted by the ABC will be modestly correlated with the cognitive reflection test (CRT; Frederick, 2005).

5.1.1.1 Rationale

The CRT was developed to measure the tendency to override an incorrect, *prepotent* response. What characterizes the items in the CRT is that although a quick, intuitive answer springs to mind, this quick answer is incorrect. The key to deriving correct answers to the problems is to suppress and/or re-evaluate the first solution that springs to mind (Frederick, 2005). It should be noted that the CRT is not a “gold standard” marker test in terms of its ability to measure bias susceptibility. It is, however, one of the few relevant measures available and, as such, was included to evaluate convergent validity of the ABC scales. Similarly, the ABC BE scales were developed to evaluate the extent to which an individual is susceptible to the same general type of cognitive process; namely, the tendency to provide an incorrect, biased response. The overlap between the CRT and ABC was not expected, however, to be especially large, given that the CRT and ABC are not targeted to the same biases. Nevertheless, a moderate correlation was expected, based on the conceptual ground described above.

An alternate hypothesis is that the CRT and ABC scales might correlate due to the fact that both correlate with cognitive ability—especially quantitative ability in the case of the CRT. As such, where ABC scales correlated with cognitive ability in general, and quantitative ability, in particular, we deemed it necessary to evaluate whether the correlations were due to these cognitive ability constructs as opposed to measurement of bias susceptibility.

5.1.1.2 Relevant Evidence

Results from our field tests showed partial support for Proposition 1. Specifically, (1) the CRT was found to correlate at statistically and practically significant levels with ABC-1 BBS and RD, and each of the four ABC-2 scales; and (2) when controlling for cognitive ability, practically significant correlations remained in the case of the ABC-2 ANC, REP, and RD scales. It should be noted, however, that the zero-order correlations between the ABC and the CRT were largely due to cognitive ability. Nevertheless, non-trivial partial correlations between CRT and ABC-2

variables showed that—especially for RD and REP—the scales showed some convergent validity. This is true, to a lesser extent, for ANC.

The pattern of results is generally supportive of our hypotheses. This statement rests on our post-hoc interpretation of the data as reflecting a mismatch in breadth between the CRT and ABC. More specifically, it seems reasonable to expect the highest partial correlation to be with the ABC REP scale, because (1) both are saturated with quantitative content, (2) the ABC REP scale correlated more highly with general cognitive ability (and quantitative ability) than the other ABC BE scales, and (3) the ABC REP scale correlated highly with the ABC-2 RD scale, indicating that it may be more saturated with bias-related declarative knowledge than the other ABC BE scales. It is somewhat strange that the partial correlation between the ABC-1 RD scale did not retain statistical significance when controlling for *g*, whereas the partial correlation between the ABC-2 RD scale did retain significance. The significant partial correlation in the case of the ABC-2 RD scale can be explained as follows: to score highly on the CRT, it would seem to be necessary to recognize (i.e., have declarative knowledge of) biases. Without such knowledge, it would seem unlikely that a prepotent System 1 response could be overridden by a System 2 response requiring explicit knowledge. The ABC-2 RD scale was intended to measure declarative knowledge of the Phase 2 biases. As such, both the zero-order and partial correlations between the CRT and the ABC-2 RD scales make sense. Similarly, the lack of overlap between the CRT and the ABC-1 RD scale is likely due to the CRT being more limited in the range of biases to which it is relevant. For example, the CRT seems most relevant to quantitatively-based cognitive biases, such as ABC-2 REP, and least relevant to the social cognitive biases measured without need for quantitative ability, such as the ABC-1 FAE. This also explains the non-significant partial correlations between the CRT and ABC-1 BE scales. ABC-1 FAE and BBS are very limited in their quantitative demands, and are the most socially-oriented biases in the ABC, along with ABC-2 PRO, which also had a very low partial correlation with the CRT. With regard to CB, the lack of correlation with the CRT was likely due to measurement issues with the CB scale.

5.1.2 Proposition 2

Proposition 2a: The ABC-2 scales should correlate modestly with their analog scales on the BICC (e.g., ABC-2 Anchoring should correlate modestly with BICC Anchoring, ABC-2 Representativeness should correlate modestly with BICC Representativeness).

Proposition 2b: Each ABC-2 scale should correlate higher with its analog scale on the BICC than with its non-analog BICC scales, thereby providing evidence of discriminant validity (Campbell & Fiske, 1959).

5.1.2.1 Rationale

The BICC was developed specifically to operationalize the same bias susceptibility construct as the ABC-2. As with the CRT, the BICC is not a “gold standard” marker test. It is, however, an excellent test to provide convergent and discriminant validity evidence for the ABC-2 given that it is the only extant test designed to measure precisely the same constructs. Nevertheless, it should be noted that, as with the ABC-2, the BICC was developed under significant time constraints, and with the same measurement and logistical challenges as we experienced in developing and implementing the ABC-2. The BICC was developed independently of the ABC-2, which meant that we did not pool our resources, though such independence has the virtue of

making convergent and discriminant validity coefficients more meaningful. Given the foregoing, we predicted lower-than-typical convergent validities for analog constructs (Proposition 2a); discriminant validity predictions (Proposition 2b) were not expected to approach zero, nor even to be non-significant, but simply to be lower than convergent validity coefficients. Correlations between the ABC and BICC non-analog scales—our discriminant validity coefficient—are a function of the true correlations between bias constructs, which were unknown to us as we began the Phase 2 research.

5.1.2.2 Relevant Evidence

Proposition 2a was supported in our field trial data. Correlations between the ABC-2 and BICC analog scales were .16, .39, and .27 (all $p < .01$), for the ANC, REP, and PRO BE scales, respectively³⁶. After disattenuating the correlations for unreliability in both measures, these convergent validities rise to .28, .64, and .42 for ANC, REP, and PRO, respectively.

Proposition 2b was generally supported as well. REP and PRO had substantially higher convergent validities than the highest discriminant validity coefficients, particularly when correcting for unreliability. The ANC convergent validity coefficient was only trivially larger than the discriminant validity coefficient between the ABC-2 ANC and BICC PRO scale. This was true in both the uncorrected and corrected results.

5.1.3 Proposition 3

In the presence of a construct-valid bias mitigation intervention, all ABC scales will be higher on posttest than on pretest in a study that includes a control group, and with a study design that controls for pretest sensitization.

5.1.3.1 Rationale

The updated Standards (2014) emphasize that validity can only be discussed and claimed within the context of the intended use of a test. In the Sirius project, the ABC is to be used for detecting bias mitigation after an intervention intended to reduce the targeted biases. Proposition 3 makes this prediction. The inclusion of the control group and evaluation of pretest sensitization in the wording of the proposition acknowledge that the validity of this claim is scientifically meaningful only when certain alternate explanations can be ruled out.

5.1.3.2 Relevant Evidence

Our results, and results from the IV&V studies implemented by JHUAPL, collectively provide solid support for Proposition 3. In our Phase 1 and 2 field tests, we were provided with an instructional video developed by IARPA for purpose of evaluating sensitivity to a bias mitigation intervention prior to the IV&V stage for both Phases 1 and 2. In both of our bias mitigation studies, our experimental design included a control group and allowed us to evaluate whether a pretest sensitization effect could be confounding our results. It should be noted that the IARPA video, while carefully developed, was not intended to serve as a “gold standard” bias mitigation

³⁶ We did not collect data for the BICC RD scales due to logistical and timing constraints.

intervention. Indeed, the purpose of the Sirius project was to yield one or more bias mitigation interventions to fill a gap in the literature.

As described and explained more fully in Sections 2.6.3 and 3.6.3, our results showed that for the Phase 1 and 2 biases, test-takers scored higher on the ABC after being shown the IARPA video, compared to the control group, after taking into account any pretest sensitization effect. This was true for 7 out of the 8 bias constructs. Our results showed that the RD scales from both phases 1 and 2 changed the most, and that REP also changed substantially. This was also true, albeit to a lesser extent, for ANC. It should be noted that the IARPA video was not an especially strong bias mitigation intervention given the amount of material that needed to be covered, which likely accounts for the larger effect sizes associated with constructs most highly saturated with declarative knowledge.

The IV&V results reported by JHUAPL are also relevant to evaluating this proposition.

For Phase 1, the IV&V results for several of the game-based training interventions showed medium to large effect sizes for all of the Phase 1 bias constructs. For Phase 2, the IV&V results for game-based training showed very large effect sizes for RD and REP, medium to large effects for PRO, and small to medium effects for ANC. It is also noteworthy that the effect sizes varied considerably for the three serious video games tested in the Phase 2 IV&V. In general, the pattern of results for the Sirius video games mirrors the ABC field test results using the IARPA video. It is entirely sensible that the magnitude of the effects would be substantially larger for the Sirius video games evaluated in the IV&V studies given that the interventions were much more intensive and developed over multiple rounds of iterative testing. It is also noteworthy that the differences in effect sizes between the IV&V studies and the ABC field test intervention studies were lowest in the case of the RD tests. The declarative knowledge assessed by the RD tests was far more (1) accessible and (2) isomorphic with the IARPA video and serious video game content. By contrast, the BE scales required substantial additional knowledge beyond the declarative knowledge assessed by the RD. The general support for this proposition is perhaps the most important part of our validity argument. It speaks not merely to construct validity, but to the ability of the ABC to detect changes in construct level before and after bias mitigation interventions.

5.1.4 Proposition 4

For each construct, BE scales will be uncorrelated with RD scales most relevant to those constructs.

5.1.4.1 Rationale

Proposition 4 is based on the following: (1) the presence of declarative knowledge on a given topic is necessary but not sufficient for demonstration of procedural knowledge in the same domain; and (2) there is a substantial extant literature in cognitive psychology showing that declarative and procedural knowledge are relatively independent (Roediger, Zaromb, & Goode, 2008; Squire, 1992).

5.1.4.2 Relevant Evidence

Results from the Phase 1 and 2 field tests support Proposition 4. Even when disattenuated for unreliability, correlations between BE and RD scales seldom rose above .3. The correlation between the ABC-1 RD scale and the BBS scale was slightly higher than .30, but when

controlling for *g*, partial correlation was reduced to .22. These results are consistent with the notion that declarative knowledge accounts for a portion of the BE scale-scores, but that BE scale-scores also require procedural knowledge/skill much of which is unrelated to declarative knowledge.

5.1.5 Proposition 5

Each ABC scale will show a pattern of correlations with measures of cognitive ability and personality that permits the inference that the ABC scales measure constructs distinct from both the cognitive ability and personality domain.

5.1.5.1 Rationale

This proposition evaluates the discriminant validity of the ABC from the two individual-difference domains with which they perhaps have the most in common. As such, measures of the Big-5 personality factors, together with other relevant personality scales (e.g., need for cognition; Cacioppo, Petty, & Kao, 1984) and measures of general cognitive ability should not correlate so highly with ABC scales that one could argue that the ABC bias constructs are simply personality and/or cognitive ability. This, along with convergent validity, is a cornerstone of construct validity.

Our review revealed little extant literature regarding correlates of bias constructs, probably because bias has not been programmatically researched within an individual-difference framework. One exception is the work of Stanovich and his colleagues (e.g., Stanovich & West, 2008; Stanovich & West, 1998; West, Meserve, & Stanovich, 2012). That research, which has investigated correlates of versions of classic experimental paradigms of various biases, has found modest correlations between proxies of general cognitive ability, such as the SAT, and measures of various bias-related tasks. Similarly, Stanovich and colleagues have investigated the relationships between certain personality variables they judged most relevant to cognitive biases (e.g., those included in what they refer to as Actively Open-Minded Thinking (Stanovich & West, 1997, 2007) and again found only modest relationships.

5.1.5.2 Relevant Evidence

Results from our two field tests provide strong support for the independence of both the ABC BE and RD scales and major personality factors. While correlations between the ABC scales and measures of general cognitive ability, as well as verbal and quantitative sub-factors of general cognitive ability, are higher than those involving personality factors, they are sufficiently modest to support the inference of discriminant validity with respect to the ABC scales; especially the BE scales.

An even stronger test of discriminant validity would be to investigate the overlap between each ABC scale and the entire personality and ability domains. To that end, we regressed each ABC scale on a set of variables that included (1) *g*, and (2) the FC-BFI personality factors. For the BE scales, the squared multiple correlations never exceeded .10. For the RD scales, the squared multiple correlations were .40 and .26 for the ABC-1 and ABC-2 RD scales, respectively. As such, even for the RD scales, there is no evidence to refute a claim of discriminant validity for the ABC scales.

5.2 Summary

The accumulated evidence is consistent with the inference that the ABC is valid for its intended use. Despite the lack of “gold standard” marker tests and bias mitigation interventions, the available evidence indicates that the ABC scales show both convergent and discriminant validity and are sensitive to bias mitigation interventions. Moreover, the extensive literature review conducted for this project enabled us to partition the content domain for each of the six bias constructs measured by the ABC into a set of facets that are both meaningful and comprehensive. That said, we emphasize that validation, especially for novel constructs such as those measured by the ABC, is an ongoing process. While the research record assembled during the course of this project is extensive and provides strong validity argument, additional validity research is needed to solidify and extend our understanding of the constructs measured by the ABC. The topic of future research is addressed in the next section of this report.

6 Future Research

Although a great deal of research was done in the course of developing and evaluating the validity of the ABC, the study of bias within an individual difference framework is still largely in its infancy. As such, the research documented in this report can serve as a springboard for many other potential research programs. We list several possibilities below.

The Sirius project encompassed six biases deemed important for intelligence analysis work. It should be noted, however, that there are many more cognitive biases that seem worthy of investigation. The might include such constructs as hindsight bias, planning fallacy, and susceptibility to sunk costs, among others.

A prominent method for evaluating the validity of a test in applied work settings is criterion-related validation; that is, identifying and operationalizing major work performance dimensions and then correlating test performance with work performance. Work performance may consist of both subjective (e.g., supervisor ratings) and quasi-objective criteria (e.g., quantification of errors committed). Work performance may also consist of overt or covert (i.e., cognitive) behaviors. Identification of relevant performance constructs could be derived from case studies in the literature (e.g., Beebe & Pherson, 2011; Heuer, 1999), but more fruitfully, SMEs with knowledge of the work domain under investigation (e.g., job). Subsequently, surveys or more in-depth cognitive interviews can be used to complete the job/work analysis and yield performance constructs. The job/work analysis can then be used to facilitate development of work performance measures against which instruments such as the ABC can be validated. This would enable us to know whether higher test performance is associated with better job/work performance.

It should be noted that, relevant to personnel selection applications such as the one described above, the ABC-2 was shown to have a smaller effect size than measures of general cognitive ability with respect to differences between Caucasian and certain EEOC racial/ethnic protected class subgroups.

In addition to personnel selection applications, instruments such as the ABC might also be fruitfully used for training and development purposes; that is, as part of a formative assessment system. In this way, the ABC could be used not just predictively, but also diagnostically. This might be done not merely within a work context, but within a clinical context as well.

In order, however, for the ABC to be used diagnostically, it would be necessary to conduct additional research to establish the ABC not only as an effective group-level measure, but also as an individual-level measure. Some of the research conducted as part of this project suggested that the ABC would make for an effective individual-level measurement tool, but more work needs to be done before we could recommend it for use at this level.

Another fruitful area for future research involves the nature of the bias mitigation interventions. For example, we, in conjunction with JHUAPL, are in the process of conducting formative evaluations of the ABC and the Sirius video game and instructional video interventions to determine what aspects of the best-performing interventions produced mitigation.

Some of the cross-bias correlations have implications for enhancing bias mitigation interventions. For example, note that the correlation between BBS and PRO was $-.18$ ($p < .01$, disattenuated $r = -.27$). That is, the more people think that others think like themselves (showing greater PRO), the less likely they are to attribute more bias to others (showing less BBS). This may have implications for bias mitigation interventions in that training one of these biases may counteract training of other biases. As such, training must stress the point that mitigating BBS, in the absence of the acknowledgment of the modest negative correlation with PRO, may accentuate PRO susceptibility.

Similarly, the correlation between REP and FAE was $.21$ ($p < .01$, disattenuated $r = .29$). That is, the more susceptible people are to REP the more susceptible they are likely to be to FAE. This raises the possibility that bias mitigation could be made more efficient as a result of the positive relationship between these two bias susceptibility constructs.

Extending the ABC into applied personnel selection would also require investigation of its validity for different job groups. If criterion-related validities are promising, it would be desirable to build up a database for major job families using the O*NET classification system. This would facilitate meta-analytic research³⁷ that would enable us to identify job families for which different biases are especially predictive and potential moderators of the bias-performance relationships that might be found. The applied research domain is, of course, not limited to the workplace. It seems likely that the study of bias within an individual difference framework would also be useful within the clinical and counseling domains.

Several biases would seem relevant to the burgeoning literature on cross-cultural competence. So, for example, projection bias, or as it is often referred to in the intelligence community, “Mirroring,” may affect people’s ability to relate to individuals in different cultures. Culture may also play a role in the study of cognitive and social biases to the extent that people in various cultures differ in their knowledge of and susceptibility to different biases. For example, findings from several studies suggest that people from different cultures differ in their susceptibility to FAE (e.g., Chua, Leu, & Nisbett, 2005; Morris & Peng, 1994).

As our knowledge of individual differences in bias susceptibility matures, it would likely be interesting and useful to study them within a multi-level framework (e.g., Klein & Kozlowski, 2000). This would involve studying bias susceptibility in aggregates of individuals, such as teams, organizations, and other meaningful social groups. It may be that some of the same biases

³⁷ A statistical method of research in which the results from independent, comparable studies are combined to determine the size of an overall effect or the degree of relationship between two variables.

that we have studied in this project would be relevant to such groups, or it may be the case that biases unique to aggregates will be discovered.

7 References

- Albano, A. D. (2011). *Equate: Statistical methods for test equating* [Computer software manual]. Retrieved from <http://CRAN.R-project.org/package=equate> (R package)
- Almond, P.J., Cameto, R., Johnstone, C.J., Laitusis, C., Lazarus, S., Nagle, K., Parker, C., Roach, A.T., & Sato, E. (2009). *Cognitive interview methods in reading test design and development for alternate assessments based on modified academic achievement standards (AA-MAS)* [White paper]. Dover, NH: Measured Progress and Menlo Park, CA: SRI International.
- American Education Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Education Research Association.
- American Education Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Education Research Association.
- Bagozzi, R. P. (2007). On the meaning of formative measurement and how it differs from reflective measurement: Comments on Howell, Breivik, and Wilcox. *Psychological Methods, 12*, 229–237.
- Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*, 211–233.
- Barrett, F., & Russell, J. A. (1998). Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology, 74*, 967–984.
- Beebe, S. M., & Pherson, R. H. (2011). *Cases in intelligence analysis: Structured analytic techniques in action*. Thousand Oaks, CA: CQ Press.
- Birns, J. H., Joffre, K. A., Leclerc, J. F., & Paulsen, C. A. (2002). *Getting the whole picture: Collecting usability data using two methods – concurrent think aloud and retrospective probing*. Paper presented at the annual Usability Professionals' Association Conference, Orlando, FL. Abstract retrieved from <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.111.5466>
- Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*, 305–314.
- Braver, M. W., & Braver, S. L. (1988). Statistical treatment of the Solomon four-group design: A meta-analytic approach. *Psychological Bulletin, 104*, 150–154.
- Brown, A., & Maydeu-Olivares, A. (2011). Forced-Choice Five Factor Markers [Database record]. Retrieved from PsycTESTS. doi: 10.1037/t05430-000
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3–5.
- Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI-2 manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.

- Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, 48, 306–307.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validity by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Chua, H.F., Leu, J., & Nisbett, R.E. (2005). Culture and diverging views of social events. *Personality and Social Psychology Bulletin*, 31, 925-934.
- Cook, M. B. & Smallman, H. S. (2008). Human factors of the confirmation bias in intelligence analysis: Decision support from graphical evidence landscapes. *Human Factors*, 50, 745–754.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 98–104.
- Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how human reason? Studies with the Wason selection task. *Cognition*, 31, 187–276.
- Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58, 1–73.
- Cronbach, L. J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Darley, J. M, & Batson, C. D. (1973). From Jerusalem to Jericho: A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology*, 27, 100–119.
- Del Missier, F., Bonini, N., & Ranyard, R. (2007). The euro illusion in consumers' price estimation: An Italian-Irish comparison in the third year of the euro. *Journal of Consumer Policy*, 30, 337-354.
- Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, 38, 269–277.
- Dumas, J. S. and Redish, J. C. (1993). *A practical guide to usability testing*. Portland, OR: Intellect Books.
- Edwards, J. R. (2001). Multidimensional constructs in organizational behavior research: An integrative analytical framework. *Organizational Research Methods*, 4, 144–192.
- Ekstrom, R.B., French, J.W., & Harman, H.H. (1979). Cognitive factors: Their identification and replication. *Multivariate Behavioral Research Monographs*, 79, 3-84.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science*, 17, 311–318.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.
- Evans, J. St. B. T. (1982). *The psychology of deductive reasoning*. London, England: Routledge & Kegan Paul.

- Evans, J. St. B. T., Newstead, E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Hove, England: Erlbaum.
- Fingar, T. (2011). *Reducing uncertainty: Intelligence analysis and national security*. Stanford, CA: Stanford University Press.
- Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, *22*, 323–327.
- Fischer, P., Greitemeyer, T., & Frey, D. (2008). Self-regulation and selective exposure: The impact of depleted self-regulation resources on confirmatory information processing. *Journal of Personality and Social Psychology*, *94*, 382–295.
- Fischer, P., Kastenmüller, A., Greitemeyer, T., Fischer, J., Frey, D., & Crelley, D. (2011). Threat and selective exposure: The moderating role of threat and decision context on confirmatory information search after decisions. *Journal of Experimental Psychology: General*, *140*, 51–62.
- Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, *19*, 25–42.
- Frey, D., & Rosch, M. (1984). Information seeking after decisions: The roles of novelty of information and decision reversibility. *Personality and Social Psychology Bulletin*, *10*, 91–98.
- Gawronski, B. (2003). On difficult questions and evident answers: Dispositional inference from role-constrained behavior. *Personality and Social Psychology Bulletin*, *29*, 1459–1475.
- Gertner, A., Zaromb, F., Burrus, J., Roberts, R., Bowen, C., & Schneider, R. (2011). *Developing a standardized assessment of cognitive bias for the IARPA Sirius program*. (MITRE Technical Report). McLean, Virginia: The MITRE Corporation.
- Gertner, A., Zaromb, F., Schneider, R., Rhodes, R., Matthews, G., Burrus, J., Roberts, R. D., & Bowen, C. (2013). *Developing a standardized assessment of cognitive bias for the IARPA Sirius program: A review of the literature*. (MITRE Technical Report). McLean, Virginia: The MITRE Corporation.
- Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*, 295–314.
- Goldstein, I. L. (1991). Training in work organizations. In M. D. Dunnette & L.M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., vol. 2, pp. 507–619). Palo Alto, CA: Consulting Psychologists Press.
- Goodie, A. S., & Fantino, E. (1995). An experientially derived base-rate error in humans. *Psychological Science*, *6*, 101–106.
- Hamill, R., Wilson, T. D., & Nisbett, R. E. (1980). Insensitivity to sample bias: Generalizing from atypical cases. *Journal of Personality and Social Psychology*, *39*, 578–589.
- Heuer, R.J., Jr. (1999). *Psychology of intelligence analysis*. Washington, DC: Center for the Study of Intelligence, Central Intelligence Agency.
- Hoffman, R., Henderson, S., Moon, B., Moore, D. T., & Litman, J. T. (2011). Reasoning difficulty in analytical activity. *Theoretical Issues in Ergonomics Science*, *12*, 225–240.
- Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods*, *12*, 205–218.

- Humphreys, L. G. (1979). The construct of general intelligence. *Intelligence*, 3, 105-120.
- Hutchins, S. G., Pirolli, P. L., & Card, S. K. (2007). What makes intelligence analysis difficult? A cognitive task analysis. In R. R. Hoffman (Ed.), *Expertise out of context*. (pp. 281–316). Mahwah, NJ: Lawrence Erlbaum Associates.
- Intelligence Advanced Research Projects Activity (2012). *Unbiasing your biases I*. Alexandria, VA: 522 Productions.
- Intelligence Advanced Research Projects Activity (2013). *Unbiasing your biases II*. Alexandria, VA: 522 Productions.
- Irvine, S. H., & Kyllonen, P. (Eds.). (2002). *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The “Big Five” inventory-Version 4a and 54*. Retrieved from <https://www.ocf.berkeley.edu/~johnlab/bfi.htm>.
- Johnstone, C. J., Bottsford-Miller, N. A., & Thompson, S. J. (2006). Using the think aloud method (cognitive labs) to evaluate test designs for student with disabilities and English language learners (Technical Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, 3, 1–24.
- Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 11, 527–535.
- Kassin, S. M., & Sukel, H. (1997). Coerced confessions and the jury: An experimental test of the “harmless error” rule. *Law and Human Behavior*, 21, 27–46.
- Kebbell, M. R., Muller, D., & Martin, K. (2010). Understanding and managing bias. In G. Bammer (Ed.), *Dealing with uncertainties in policing serious crime* (pp. 87–97). Canberra, Australia: Australian National University.
- Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94, 211–228.
- Klein, K. J., & Kozlowski, S.W.J. (Eds.) (2000). *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions*. San Francisco, CA: Jossey-Bass.
- Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, 19, 1–53.
- Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.
- Krueger, J., & Clement, R. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality and Social Psychology*, 67, 596–610.
- Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, 23, 6–15.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3, 635–694.

- Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.
- MacCallum, R., & Browne, M. W. (1993). The use of causal indicators in covariance structure models: Some practical issues. *Psychological Bulletin*, *114*, 533–541.
- MacCann, C., & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: Theory and data. *Emotion*, *8*, 540–551.
- Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review*, *78*, 333–368.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.
- Mochon, D., & Frederick, S. (2013). Anchoring in sequential judgments. *Organizational Behavior and Human Decision Processes*, *122*, 69–79.
- Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*, 502–517.
- Morris, M.W., & Peng, K. (1994). Culture and cause: American and Chinese attributions for social and physical events. *Journal of Personality and Social Psychology*, *47*, 949-971.
- Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Broday, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*, 77-101.
- Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin*, *125*, 737–759.
- Obrecht, N. A., Chapman, G. B., & Gelman, R. (2009). An encounter frequency account of how experience affects likelihood estimation. *Memory & Cognition*, *37*, 632–643.
- Oppenheimer, D. M., & Monin, B. (2009). The retrospective gambler's fallacy: Unlikely events, constructing the past, and multiple universes. *Judgment and Decision Making*, *4*, 326–334.
- Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, *5*, 343–355.
- Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. *Journal of General Psychology*, *129*, 257–299.
- Peterson, N. S., Kolen, M. J., & Hoover, H. D. (1993). Scaling, norming, and equating. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221-262). New York: American Council on Education and McMillan.
- Plous, S. (1993). *The psychology of judgment and decision making*. New York, NY: McGraw-Hill.
- Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*, 369–381.

- R Core Development Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from <http://www.R-project.org/>.
- Reise, S. P., & Duan, N. (Eds.) (2003). *Multilevel modeling: Methodological advances, issues, and applications*. Mahwah, NJ: Erlbaum.
- Riggio, H. R., & Garcia, A. L. (2009). The power of situations: Jonestown and the fundamental attribution error. *Teaching of Psychology, 36*, 108–112.
- Roediger, H. L., & Butler, A.C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*, 20–27.
- Roediger, H. L. III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181–210.
- Roediger, H. L., Zaromb, F. M., & Goode, M. K. (2008). A typology of memory terms. In R. Menzel (Ed.), *Learning theory and behavior*. Vol. 1 of *Learning and memory: A comprehensive reference*, 4 vols. (pp. 11–24). Oxford, UK: Elsevier.
- Ross, L., Amabile, T. M., & Steinmetz, J. L. (1977). Social roles, social control, and biases in social-perception processes. *Journal of Personality and Social Psychology, 35*, 485–494.
- Sato, E., Rabinowitz, S., Gallagher, C., & Huang, C. W. (2010). *Accommodations for English language learner students: the effect of linguistic modification of math test item sets*. (NCEE 2009-4079). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Snyder, M. L., & Frankel, A. (1976). Observer bias: A stringent test of behavior engulfing the field. *Journal of Personality and Social Psychology, 34*, 857–864.
- Snyder, M. L., & Swann, W. B. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology, 36*, 1202–1212.
- Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin, 46*, 137–150.
- Squire, L. R. (1992). Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. *Journal of Cognitive Neuroscience, 4*, 232–243.
- Stankov, L., & Lee, J. (2008). Confidence and cognitive test performance. *Journal of Educational Psychology, 100*, 961–976.
- Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89*, 342–357.
- Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127*, 161–188.
- Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking and Reasoning, 13*, 225–247.
- Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology, 94*, 672–695.

- Stolarz-Fantino, S., Fantino, E., & Van Borst, N. (2006). Use of base rates and case cue information in making likelihood estimates. *Memory and Cognition*, *34*, 603–618.
- Tecuci, G., Schum, D., Boicu, M., Marcu, D., & Hamilton, B. (2010). Intelligence analysis as agent-assisted discovery of evidence, hypotheses and arguments. In G. Phillips-Wren, L. C. Jain, K. Nakamatsu, & R. J. Howlett (Eds.), *Advances in intelligent decision technologies* (pp. 1–10). New York, NY: Springer.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. In P. S. D. Kahneman & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 3–20). Cambridge, UK: Cambridge University Press.
- Tversky, A., & Kahneman, D. (1982). Evidential impact of base-rates. In P. S. D. Kahneman & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153–160). Cambridge, UK: Cambridge University Press.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293-315.
- Tweney, R. D., & Doherty, M. E. (1983). Rationality and the psychology of inference. *Synthese*, *57*, 139–161.
- Wason, P. C. (1966). Reasoning. In B. M. Foss (Eds.), *New horizons in psychology*. Harmondsworth, England: Penguin.
- Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*, 273–281.
- Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content*. Cambridge, MA: Harvard University Press.
- Wedell, D. H. (2011). Probabilistic reasoning in prediction and diagnosis: Effects of problem type, response mode, and individual differences. *Journal of Behavioral Decision Making*, *24*, 157–179.
- West, R. F., Meserve, R. J., & Stanovich, K. E. (2012). Cognitive sophistication does not attenuate the bias blind spot. *Journal of Personality and Social Psychology*, *103*, 506-519.
- Williams, B. S. (September-October, 2010). Heuristics and biases in military decision-making. *Military Review*, pp. 40–52.

Appendix A Overall Project Organization

In the following sections, we list and describe the roles for key project personnel and sub-contractors involved in the project.

MITRE Project Organization

Program Management

Dr. James McDonald is a Project Leader at MITRE. Dr. McDonald was involved with oversight of the ABC development project, including development of Informal Task Statements and staff plans, management of subcontracts, and review of all deliverables

The following MITRE staff members were involved with the management of contracts, financial support, tracking deliverables, and monthly status reporting.

- Lisa Nudi
- Mary Raffa
- John Whitenack
- Michael Fine

Technical Staff

Dr. Abigail Gertner is a Principal Artificial Intelligence Engineer and Associate Department Head at MITRE. She was the MITRE task lead for the development of the ABC. As task lead, Dr. Gertner was responsible for overseeing the development of the test instruments, serving as liaison between the ETS development team and the Government Independent Validation and Verification (IV&V) team, and reviewing the research teams' results at regular site visits. She also led a team of software developers at MITRE who developed a web-based software platform for administering the ABC test instrument.

The following MITRE staff members contributed to the development of the ABC software platform:

- Susan Lubar (Phase 1 and 2)
- Maria Casipe (Phase 1)
- Joel Korb (Phase 1)
- Amanda Anganes (Phase 1)
- Ariel Abrams-Kudan (Phase 2)
- James Winston (Phase 2)

Additionally, **Charles Bowen** contributed to the writing of the literature review in Phase 1.

Subject matter experts

Several MITRE staff provided subject matter expertise during the development of the ABC.

Dr. Paul Lehner is a cognitive psychologist and has done research on confirmation bias in intelligence analysts. He reviewed and provided feedback on several of the items designed to measure confirmation bias in Phase 1.

Becky Lewis, Michael Maskaleris, David Merrill, and Mark Zimmerman provided insight into the impact of cognitive biases on the analytic process and reviewed several of the prototype test items with intelligence analysis themes for accuracy and relevance.

Subcontractors

Dr. Richard Roberts, the Vice President and Chief Scientist at Professional Examination Service, was a sub-contractor to MITRE during the second half of Phase 2 of the Sirius Program, providing oversight of all aspects of the project, including assessment design and statistical analysis, and information dissemination. During Phase 1 and the first half of Phase 2, Dr. Roberts was a Managing Principal Research Scientist at ETS and provided oversight of the ABC test development, including managing and mentoring ETS staff and subcontractors.

ETS Project Organization

Scientific Leadership and Staff

Dr. Patrick Kyllonen is the Senior Research Director for the Center for Academic & Workplace Readiness and Success at ETS. Dr. Kyllonen was concerned with oversight of all aspects of the project, including assessment design and statistical analysis.

Dr. Franklin Zaromb is a research scientist in the Center for Validity Research at ETS. He was responsible for literature reviews, assessment design, and preparation of the ABC test development research plan, PowerPoint summaries of findings from research studies, technical reports, research reports (and associated peer-review publications), design and implementation of research studies, and preparation of the ABC User Manual and deployment package for the Phase 1 and Phase 2 IV&V.

Dr. Jeremy Burrus is a Principal Research Scientist at ACT. Prior to joining ACT, Dr. Burrus was a Research Scientist in the Center for Academic and Workforce Readiness and Success at ETS. During Phase 1, he was responsible for literature reviews, assessment design, and preparation of pilot study reports, technical reports, research reports (and associated peer-review publications), and PowerPoint decks.

Drs. Jonathan Weeks is an Associate Research Scientist in the Center for Global Assessment at ETS and was responsible for conducting test-equating analyses to prepare six equated test forms for both phases of the Sirius Program.

Dr. Johnny Lin is an Associate Psychometrician at ETS and was responsible for developing and evaluating scoring models for anchoring and projection bias items and scales developed for the ABC-2.

The following ETS research scientists also contributed to the development of item prototypes, design and analysis of pilot studies, and preparation of PowerPoint decks:

- Dr. Jan Alegre (Phase 2)
- Dr. Bridgid Finn (Phase 1)

- Dr. Michelle Martin (Phase 1 and Phase 2)
- Dr. Kevin Petway (Phase 2)
- Dr. Rebecca Rhodes (Phase 2)
- Dr. Jacob Seybert (Phase 1)

Program Management

The following current and former ETS personnel were responsible for planning and oversight of the ABC test development project schedule, staffing, budget proposals, allocation of financial resources, and contract preparations:

- Dr. Meghan Brenneman (Phase 1)
- Paola Heincke (Phase 1)
- Andrew Latham (Phase 1)
- William Monaghan (Phases 1 and 2)
- Heather Walters (Phases 1 and 2)
- Zhitong Yang (Phase 2)

Assessment Development and IT Staff

Dr. Peter Cooper is a Principal Assessment Designer in the Assessment Development Division at ETS and was responsible for item writing, item review, and test form review.

Kasey Jueds is an Assessment Specialist in the Assessment Development Division at ETS. Her responsibilities included item writing, item review, and test form review.

Debra Pisacreta and **Thomas Florek** are Research Systems Specialists at ETS who were responsible for doing advance design and coding of experimental item types during. The Research Systems Specialist and Technology Director served as the primary liaisons between ETS and the IT area within The MITRE Corporation.

Marc Rubin is a database specialist who developed a centralized database for cataloguing ABC items and all of their associated documentation and data.

Mike Wagner, ETS's Technology Director during Phase 1, reviewed materials as needed and provided expert consultation on the protocol to follow in assessment design and item development to successfully achieve computer delivery.

In addition, the following Assessment Development staff were responsible for reviewing all ABC items to ensure that their content conformed to established testing standards for editorial quality, fairness, and sensitivity to test-takers:

- Barbara Suomi
- Cassandra Lang
- Courtney Craig
- Josh Crandall

Data Analysis

Dr. Phillip Leung is a Director of Data Analysis and Computation who was responsible for developing Python scripts³⁸ for processing data files generated by and exported from the ABC test administration platform developed by MITRE.

In addition, the following ETS staff members were responsible for data cleaning and coding, as well as conducting descriptive statistical analyses to evaluate the psychometric properties of ABC items and scoring methodologies:

- Hezekiah Bunde (Phase 1)
- Steven Holtzman (Phase 1 and Phase 2)
- Jun Xu (Phase 2)
- Fred Yan (Phase 1)
- Ningshan Zhang (Phase 1)

Research Support

The following ETS Research Assistants and Associates contributed to developing item prototypes, providing logistical support for pilot studies, preparing project documentation, updating and maintaining the ABC item databases, assisting with analyses of pilot study data, and supporting research dissemination activities:

- Meirav Attali
- Patrick Barnwell
- Lauren Carney
- Dr. Cristina Anguiano Carrasco
- Elizabeth Coppola
- Chelsea Ezzo
- Patrick Houghton
- Teresa Jackson
- Christopher Kurzum
- Travis Leibtag
- Gabrielle Moore
- Sarah Ohls
- Margaret Redman

In addition, **Heather Fell** and **Andrea Napoli** were responsible for managing recruitment and compensation for pilot studies conducted with ETS employees, and **Heather Walters** was

³⁸ Python is a widely used general-purpose, high-level programming language

responsible for managing recruitment and compensation for pilot studies conducted with Amazon Mechanical Turk workers.

Administrative Support

Mary Lucas, Joan Nimon, and Kitty Sheehan were responsible for coordinating travel and meeting logistics for project personnel to attend Technical Advisory Group meeting, Sirius Program PI meetings, and conferences.

Subcontractors to ETS and Their Roles

Dr. Robert Schneider (Research & Assessment Solutions, Ltd.) was responsible for providing scientific oversight, as well as contributing to the assessment design, item writing, administration of the pilot studies and field trials, and preparation of pilot study reports, technical reports, and research reports.

Dr. Gerald Matthews is a research professor at the Institute for Simulation & Training at the University of Central Florida. Dr. Matthews contributed to the assessment design, item writing, administration of the pilot studies and field trials, and preparation of pilot study reports, technical reports, research reports. Graduate students under the direction of Dr. Matthews also participated in the administration and analyses of data collected in pilot studies.

Creative Media for Learning (CML) of Louisville, KY is a private video production and taping service responsible for editing scripts and filming and editing videos for situational judgment tests.

Drs. Yana Weinstein (current an assistant professor at University of Massachusetts–Lowell) and **Jonathan Jackson** (currently a postdoctoral fellow at Brandeis University) designed and administered pilot studies to college students during Phase I at Washington University in St. Louis.

Technical Advisory Group and Its Role

We empanelled a technical advisory group (TAG) to advise us throughout the course of the project. The Phase 1 TAG consisted of the following individuals:

- **Dr. Larry Jacoby** (Washington University in St. Louis, <http://psych.wustl.edu/amccclab/AMCC%20Jacoby.htm>) is a faculty member of the Psychology Department at Washington University in St. Louis. His primary interests are in the areas of memory and cognition. Much of Dr. Jacoby's research has contrasted automatic vs. cognitively-controlled forms or uses of memory. For example, his research has shown that age-related differences in memory reflect a decline in cognitively-controlled uses of memory in combination with preserved automatic influences of memory. He has developed procedures to enhance older adults' memory performance by means of training cognitive control. The above work has been highly cited. Dr. Jacoby has also recently become interested in the development of computer games as a means of education. In that vein, he has served as a consultant for Cogniciti, a group that is developing computer games to enhance the memory performance of executives and other games to enhance memory performance of older adults. Currently, with support from the McDonnell Foundation, Dr. Jacoby is working with others to develop a computer game to teach

natural concepts. An initial version of the game has the goal of training ability to identify different species of birds. Other research is aimed at investigating the power of examples for teaching psychology concepts.

- **Dr. Emily Pronin** (Princeton University, <http://psych.princeton.edu/psychology/research/pronin/index.php>) is Associate Professor at Princeton University in the Department of Psychology and the Woodrow Wilson School of Public and International Affairs (she was Assistant Professor from 2003-2009). She was a Postdoctoral Fellow at Harvard University, she received her PhD from Stanford, and graduated with a B.A. from Yale. Pronin's research is in the area of social cognition. She is best known for her experimental work in areas including self-perception and judgmental bias, and for her theoretical contributions in areas including bias perception, self-knowledge, and thought speed. She originated the concept of the "bias blind spot" in a journal article published in 2002, with coauthors Daniel Lin and Lee Ross. Pronin's research is supported by a grant from the National Science Foundation. Her work has been featured in major media outlets (*The New York Times*, *Los Angeles Times*, *Chicago Tribune*, *Washington Post*, *ABC News 20/20*, *PBS Nightly Business Report*, and others). She is a contributor to the *Situationist*, a forum associated with the Project on Law and Mind Sciences at Harvard Law School, and to the *Edge Annual Question*, a forum for world-class scientists and creative thinkers. She serves on the Editorial Board of the *Journal of Experimental Social Psychology*, and is an elected member of the Society for Experimental Social Psychology. She has lectured about her research at major universities including Columbia, Cornell, Duke, Harvard, North Carolina, NYU, Penn, Wisconsin, and Yale.
- **Dr. Steve Reise** (UCLA, <http://aqm.gseis.ucla.edu/reise.html>) is a professor in the Psychology Department at UCLA. His research has primarily focused on the application of latent variable modeling techniques to psychological test data, including structural equation modeling, hierarchical (multilevel) linear modeling, and item response theory modeling. Among his publications are popular texts on item response theory (Embretson & Reise, 2000) and multilevel modeling (Reise & Duan, 2003). Most recently, Dr. Reise's research has focused on using the bifactor model to understand the latent structure of important assessment instruments. Dr. Reise is currently co-director of the Applied Quantitative Methods training grant funded by the Institute of Educational Science, and consultant on two large-scale assessment projects located at the Educational Testing Service.
- **Dr. Barbara Spellman** (University of Virginia, <http://people.virginia.edu/~bas6g/basvita.html>) is Professor of Psychology and Professor of Law at the University of Virginia. She received a J.D. from New York University School of Law in 1982 and a Ph.D. in Psychology from UCLA in 1993. Her empirical research is in the area of higher-order cognition – thinking, reasoning, and metacognition – with particular emphasis on causal, counterfactual, analogical, and inductive reasoning. She has also worked on applications of that research to the legal system and intelligence analysis. Her research has been published in leading psychology journals including: *Psychological Review*, *Journal of Social and Personality Psychology*, *Cognitive Psychology*, *Journal of Experimental Psychology: General*,

Developmental Psychology, Perspectives on Psychological Science, and Psychological Science. She has written many invited chapters on the intersection of psychology and law and has also published in various legal journals. Spellman is an elected fellow of the American Association for the Advancement of Science, the Society of Experimental Social Psychology, and the Association for Psychological Science; she has served on the Board of the Directors of the latter and on the Governing Board of the Psychonomic Society. She was a member of the National Academies Committee on Behavioral and Social Science Research to Improve Intelligence Analysis for National Security and contributed a chapter entitled “Individual Reasoning” to the companion volume to the committee report. She is currently the Editor in Chief of *Perspectives on Psychological Science*.

These researchers were selected for the TAG because they have widely acknowledged expertise specifically relevant to development and validation of the ABC (e.g., knowledge of the biases to be measured, implicit measurement methods, and / or relevant psychometric models).

Three TAG meetings were held during Phase 1 in which TAG members reviewed and provided input on project documentation, such as the literature review, research plan, descriptions of item prototypes, results from pilot research, and designs for proposed studies and additional item types.

Appendix B List of Abbreviations

ABC – Assessment of Biases in Cognition

AERA – American Educational Research Association

AMT – Amazon Mechanical Turk

ANC – Anchoring Bias

ANCOVA – Analysis of Covariance

APA – American Psychological Association

BAA – Broad Agency Announcement

BBS – Bias Blind Spot

BE – Behavioral Elicitation

BFI – Big Five Inventory

BFI-FC – Big Five Inventory – Forced Choice

BICC – Bias Instrument Coordinating Committee

CB – Confirmation Bias

CML – Creative Media for Learning

COTS – Commercial-off-the-shelf

CRT – Cognitive Reflection Test

CSV file – Comma Separated Values file

DMZ – Demilitarized Zone (sometimes referred to as a “perimeter network”)

EEOC – Equal Employment Opportunity Commission

FAE – Fundamental Attribution Error

FCE – False Consensus Effect

g – General Cognitive Ability

Gc – Crystallized Intelligence

IARPA – Intelligence Advanced Research Projects Activity

IC – Intelligence Community

IRT – Item Response Theory

IV & V – Independent Validation and Verification

JHUAPL – Johns Hopkins University Applied Physics Laboratory

JSP – JavaServer Pages

NASA-TLX – National Aeronautics and Space Administration – Task Load Index

NCME – National Council on Measurement in Education

O*NET – Occupational Information Network

PASS – Performance Assessment Scoring Services
PRO – Projection Bias
RD – Recognition and Discrimination
REP – Representativeness Bias
SJT – Situational Judgment Test
SME – Subject Matter Expert
SPSS – Statistical Package for the Social Sciences
STEM – Situational Test of Emotional Understanding
STEU – Situational Test of Emotional Management
TAG – Technical Advisory Group
UMUX– Usability Metric for User Experience
WAR – Web Application Archive