NORTH ATLANTIC TREATY
ORGANIZATION

SCIENCE AND TECHNOLOGY
ORGANIZATION

AC/323()

www.sto.nato.int

STO TECHNICAL REPORT

PUB REF STO-MP-SAS-114-PPD

ANNEX D

SAS-114 Experiment Update: Effect of Source Reliability, Information Credibility, and Classification Level on Analysts' Uncertainty about Information Accuracy

David R. Mandel, Mandeep K. Dhami, Greg Weaver, and Mark Timms

SAS-114 Experiment Update: Effect of Source Reliability, Information Credibility, and Classification Level on Analysts' Uncertainty about Information Accuracy

David R. Mandel (DRDC), Mandeep K. Dhami (Middlesex University, Greg Weaver (US Army Research Laboratory), and Mark Timms (DRDC)
Email: david.mandel@drdc-rddc.gc.ca

Defence Research and Development Canada | Recherche et développement pour la défense Canada

Canada

Acknowledgements

- **We thank Brenda Fraser, Sarah Gibbon, and William Kozey from DRDC SCS Section's Research Operations Group for assistance with this research. This research contributes to SAS-114, Project 05ad on Joint Intelligence Collection and Analytic Capability and Project CSSP-2016-TI-2224 on Improving Intelligence Assessment Processes with Decision Science.**

Background

- Three commonly used markings in intelligence production are (a) source reliability, (b) information credibility, and (c) classification.

- credibility speaks directly to information quality: i.e., probability that information received is accurate; reliability should be positively related, and classification is at best a weak indicator of accuracy.

- Our overarching goal was to examine how intelligence analysts' judgments of information accuracy are influenced by these meta-informational markings.

Scales from AJP 2.1 and other intelligence doctrine

Source Reliability		Information Credibility	
A	Completely reliable	1	Completely credible
B	Usually reliable	2	Probably true
C	Fairly reliable	3	Possibly true
D	Not usually reliable	4	Doubtful
E	Unreliable	5	Improbable
F	Reliability cannot be judged	6	Truth cannot be judged

Pertinent literature

- Baker et al. (1968) found that 87% of spot reports in an Army field exercise used A1, B2, C3, D4, E5, F6. More striking B2 comprised 72% of ratings!
- Samet (1975) studied 37 Army captains familiar with the scales and found using multiple methods that credibility had a stronger effect on assessed information accuracy than reliability.
- Travers et al. (2014) found that non-analysts exhibit a “secrecy heuristic” in which they assign more weight to classified than to unclassified information, so it is of interest to verify whether analysts are similarly biased.

Hypotheses

H1: Judged accuracy will increase with reliability and credibility

H2: Analysts will not be susceptible to the secrecy heuristic, and classification will have little or no effect on accuracy.

H3: The test-retest reliability of analysts will be proportional to the congruence of the reliability and credibility scales.

H4: Likewise, the inter-analyst reliability of accuracy judgments will be proportional to the congruence of the reliability and credibility scales.

Method

Sample

- *N* = 44 UK and US analysts/operators.
- 77% male
- *M* age = 41.8 y (*SD* = 12.7)
- *M* experience in operational community = 16.2 y (*SD* = 13.6)

Method

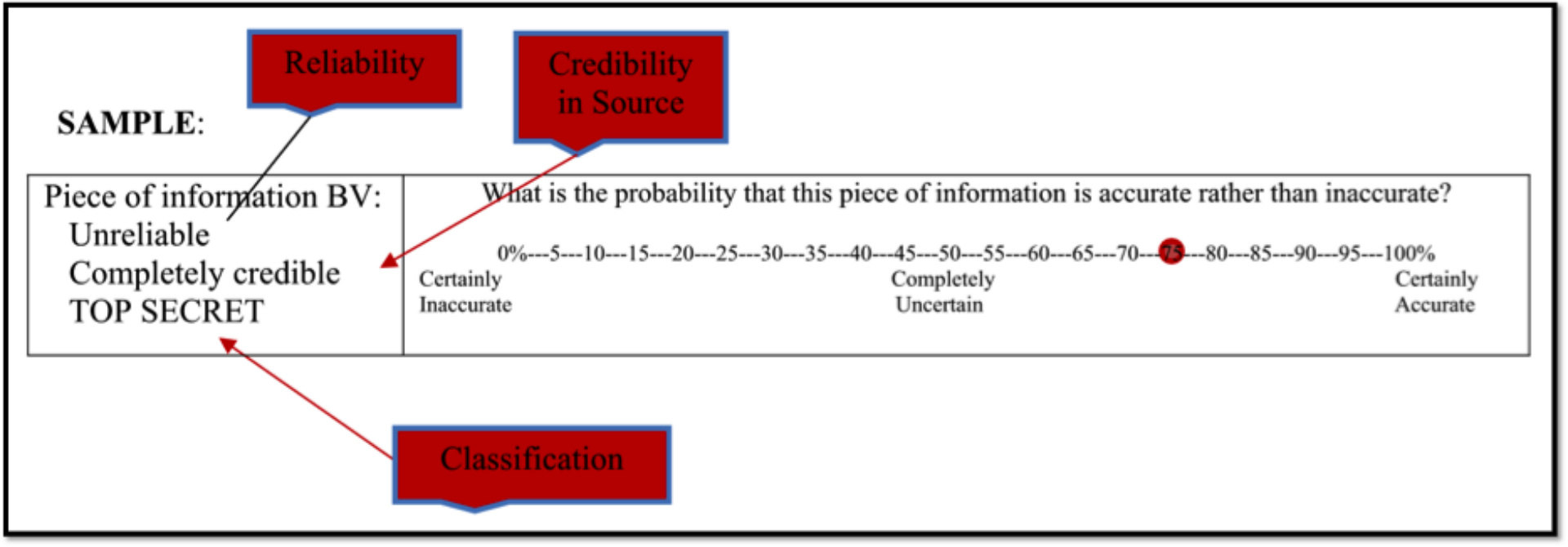
- 96% familiar with official/FOUO and TS distinction
  - 2 participants unfamiliar were omitted from analysis.
- 50.0% familiar with source reliability scale
- 53.0% familiar with information credibility scale

Method

Design and Stimuli

- Independent variables
  - Source reliability (all 6 levels)
  - Information credibility (all 6 levels)
  - Security classification (2 Official/FOUO and Top Secret)
- Full factorial repeated measures design: 6 x 6 x 2 = 72 cases plus 10 resampled cases, all presented in randomized order.
- Resampled cases all at official/FOUO level and varied degree of reliability-credibility scale congruence (low, med, high)

Method


Reliability
Credibility in Source
SAMPLE:
Piece of information BV:
Unreliable
Completely credible
TOP SECRET
What is the probability that this piece of information is accurate rather than inaccurate?
0%---5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80---85---90---95---100%
Certainly Inaccurate
Completely Uncertain
Certainly Accurate
Classification

Results testing H1 and H2

H1: Judged accuracy will increase with reliability and credibility

H2: Analysts will not be susceptible to the secrecy heuristic, and classification will have little or no effect on accuracy.

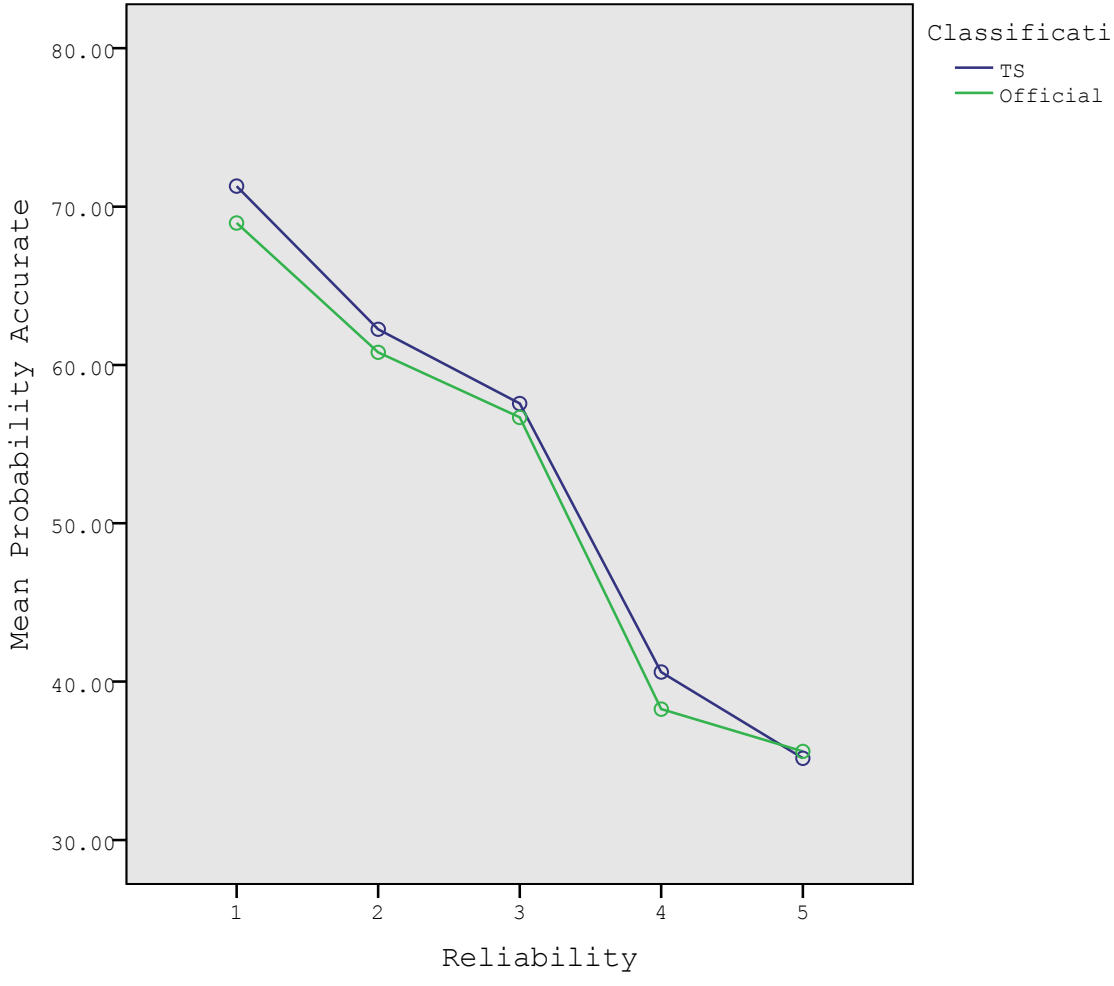
Results

Multivariate Tests

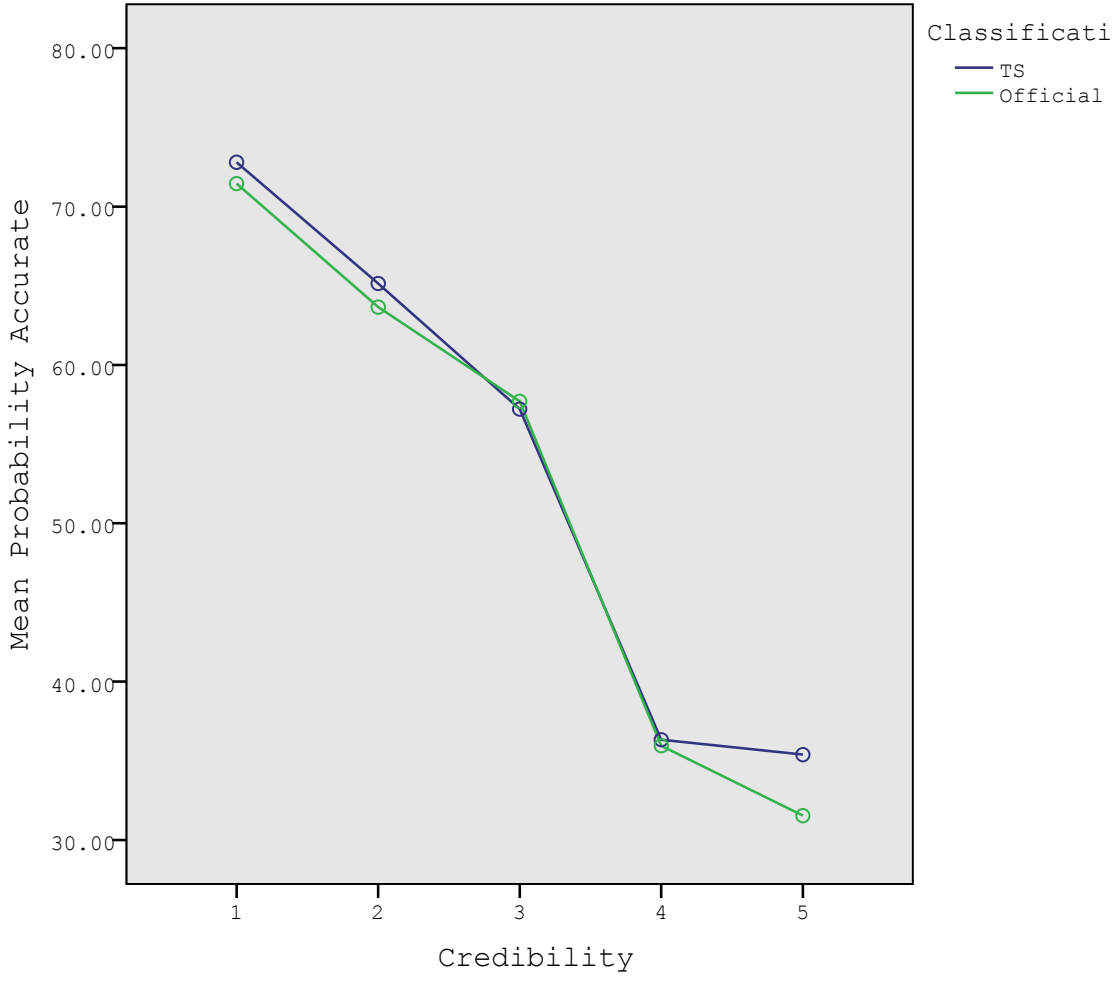
Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Reliability	Pillai's Trace	.873	65.062[b]	4.000	38.000	.000	.873
	Wilks' Lambda	.127	65.062[b]	4.000	38.000	.000	.873
	Hotelling's Trace	6.849	65.062[b]	4.000	38.000	.000	.873
	Roy's Largest Root	6.849	65.062[b]	4.000	38.000	.000	.873
Credibility	Pillai's Trace	.893	79.689[b]	4.000	38.000	.000	.893
	Wilks' Lambda	.107	79.689[b]	4.000	38.000	.000	.893
	Hotelling's Trace	8.388	79.689[b]	4.000	38.000	.000	.893
	Roy's Largest Root	8.388	79.689[b]	4.000	38.000	.000	.893
Classification	Pillai's Trace	.096	4.365[b]	1.000	41.000	.043	.096
	Wilks' Lambda	.904	4.365[b]	1.000	41.000	.043	.096
	Hotelling's Trace	.106	4.365[b]	1.000	41.000	.043	.096
	Roy's Largest Root	.106	4.365[b]	1.000	41.000	.043	.096
Reliability * Credibility	Pillai's Trace	.709	3.962[b]	16.000	26.000	.001	.709
	Wilks' Lambda	.291	3.962[b]	16.000	26.000	.001	.709
	Hotelling's Trace	2.438	3.962[b]	16.000	26.000	.001	.709
	Roy's Largest Root	2.438	3.962[b]	16.000	26.000	.001	.709
Reliability * Classification	Pillai's Trace	.174	2.001[b]	4.000	38.000	.114	.174
	Wilks' Lambda	.826	2.001[b]	4.000	38.000	.114	.174
	Hotelling's Trace	.211	2.001[b]	4.000	38.000	.114	.174
	Roy's Largest Root	.211	2.001[b]	4.000	38.000	.114	.174
Credibility * Classification	Pillai's Trace	.230	2.842[b]	4.000	38.000	.037	.230
	Wilks' Lambda	.770	2.842[b]	4.000	38.000	.037	.230
	Hotelling's Trace	.299	2.842[b]	4.000	38.000	.037	.230
	Roy's Largest Root	.299	2.842[b]	4.000	38.000	.037	.230
Reliability * Credibility * Classification	Pillai's Trace	.507	1.668[b]	16.000	26.000	.119	.507
	Wilks' Lambda	.493	1.668[b]	16.000	26.000	.119	.507
	Hotelling's Trace	1.027	1.668[b]	16.000	26.000	.119	.507
	Roy's Largest Root	1.027	1.668[b]	16.000	26.000	.119	.507

a. Design: Intercept
Within Subjects Design: Reliability + Credibility + Classification + Reliability * Credibility + Reliability * Classification + Credibility * Classification + Reliability * Credibility * Classification

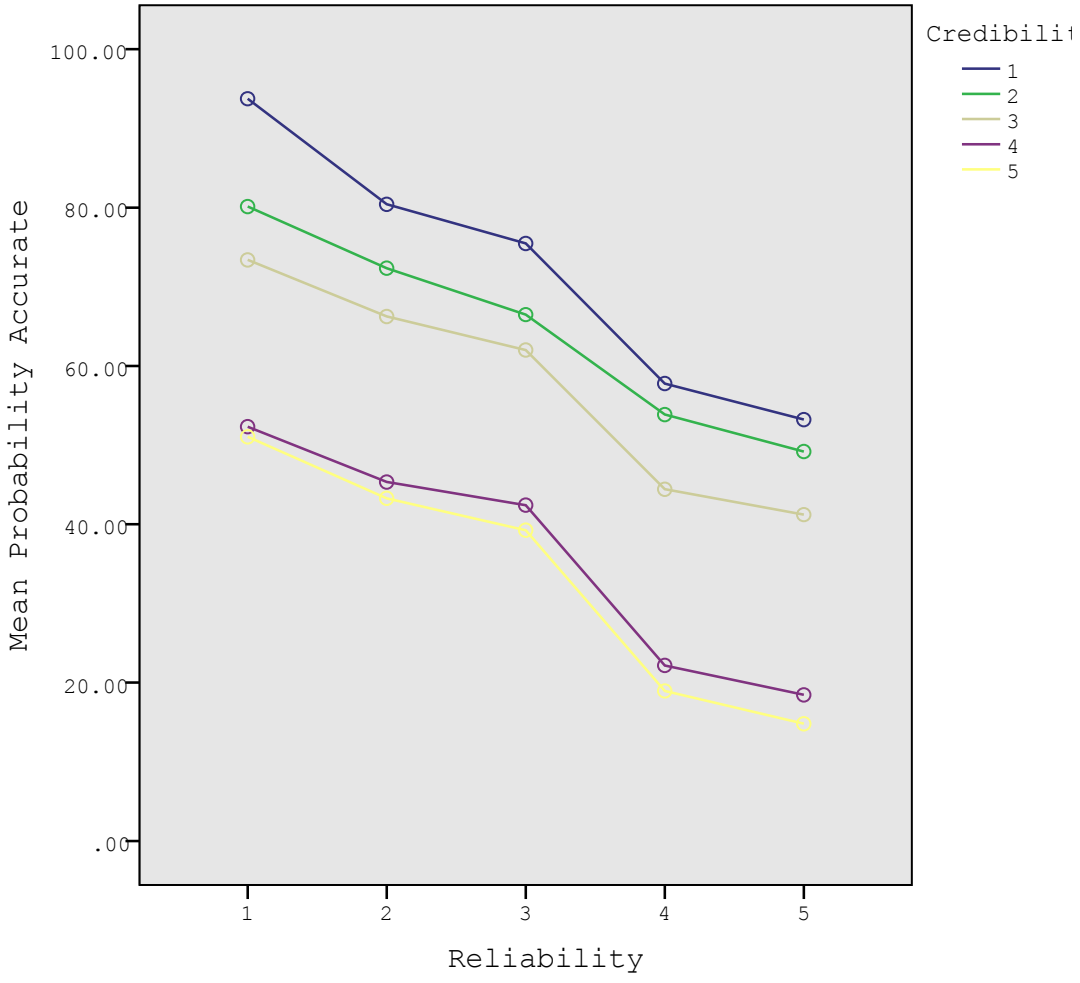
Results


Classificati
TS
Official
Mean Probability Accurate
80.00
70.00
60.00
50.00
40.00
30.00
1
2
3
4
5
Reliability

Results


Classificati
TS
Official
Mean Probability Accurate
80.00
70.00
60.00
50.00
40.00
30.00
1
2
3
4
5
Credibility

Results


Credibili
1
2
3
4
5
100.00
80.00
60.00
40.00
20.00
.00
Mean Probability Accurate
1
2
3
4
5
Reliability

Results

T-Test

[DataSet3] /Users/davidrmandel/Desktop/Dataset3_Allvariables_replaced_8Nov (1).sav

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
EffectClassification	42	1.3130	4.07267	.62843

Cohen's d = .3

One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
EffectClassification	2.089	41	.043	1.31298	.0439	2.5821

Results

Bootstrap for Coefficients

Model		B	Bootstrap[a]				
			Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval	
						Lower	Upper
1	(Constant)	1.208	-.030	.837	.155	-.468	2.692
	YearsofExp	.005	.004	.051	.931	-.089	.117
2	(Constant)	4.121	-.749	5.230	.448	-5.545	12.650
	YearsofExp	.125	-.030	.163	.497	-.184	.351
	Age	-.135	.035	.178	.498	-.463	.297
	Gender	.617	-.090	1.314	.623	-2.219	3.061

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Results testing Hypothesis 3

H3: The test-retest reliability of analysts will be proportional to the congruence of the reliability and credibility scales.

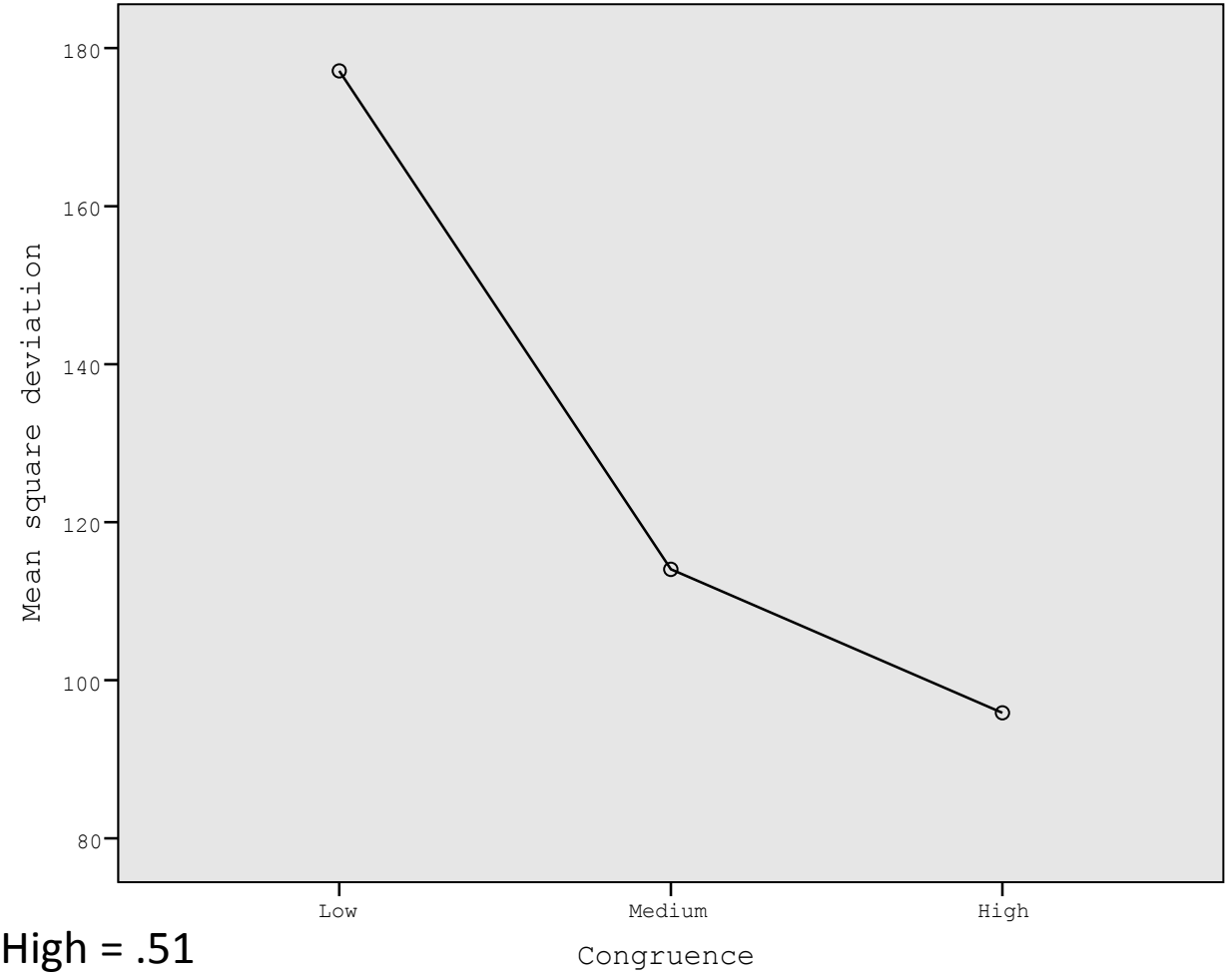
Results

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Congruence	Pillai's Trace	.178	4.345[b]	2.000	40.000	.020	.178
	Wilks' Lambda	.822	4.345[b]	2.000	40.000	.020	.178
	Hotelling's Trace	.217	4.345[b]	2.000	40.000	.020	.178
	Roy's Largest Root	.217	4.345[b]	2.000	40.000	.020	.178

a. Design: Intercept
Within Subjects Design: Congruence

Results



ES(d) for Low vs. High = .51

Results

Bootstrap for Coefficients

Model		B	Bootstrap[a]				
			Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval	
						Lower	Upper
1	(Constant)	148.555	-.875[b]	90.351[b]	.126[b]	-34.448[b]	340.394[b]
	FamRel	47.659	-.899[b]	88.981[b]	.592[b]	-135.327[b,c]	231.080[b]
	FamCred	-25.826	3.630[b]	93.582[b]	.782[b]	-187.871[b]	177.572[b]
2	(Constant)	-202.806	17.689[b]	247.766[b]	.420[b]	-673.514[b]	374.254[b]
	FamRel	16.262	-3.055[b]	64.390[b]	.780[b]	-103.745[b,c]	159.107[b]
	FamCred	37.245	-2.904[b]	67.523[b]	.572[b]	-118.905[b]	154.111[b]
	Age	7.811	-.422[b]	5.347[b]	.163[b]	-3.941[b]	17.342[b]
	Gender	-4.082	2.476[b]	85.638[b]	.960[b]	-149.280[b]	174.726[b]
	YearsofExp	-.593	.415[b]	4.654[b]	.904[b]	-9.395[b]	9.728[b]

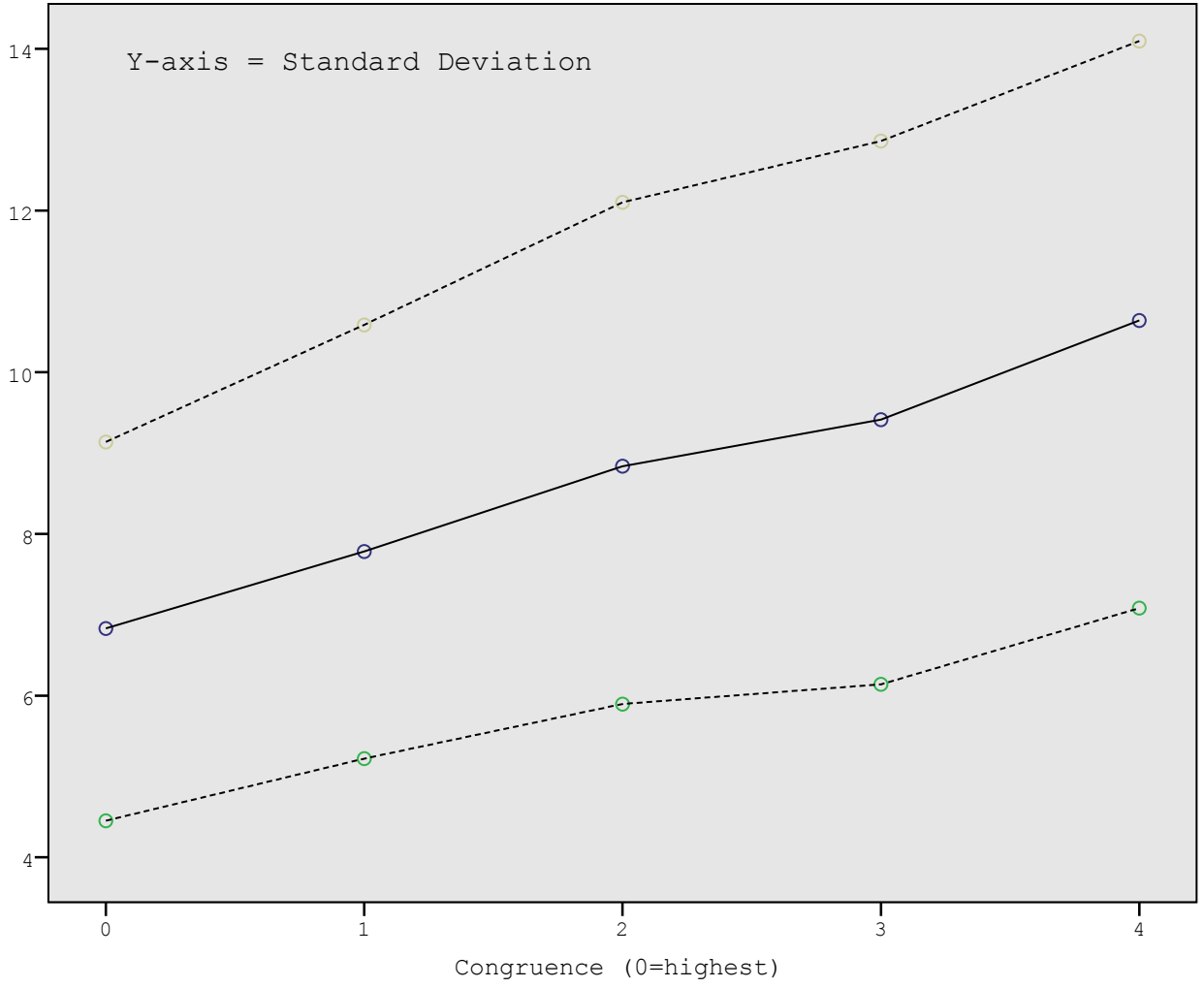
a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

b. Based on 949 samples

Results testing Hypothesis 4

H4: Likewise, the inter-analyst reliability of accuracy judgments will be proportional to the congruence of the reliability and credibility scales.

Results


Y-axis = Standard Deviation
14
12
10
8
6
4
0
1
2
3
4
Congruence (0=highest)

Results

Bootstrap for Coefficients

Model		B	Bootstrap[a]				
			Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval	
						Lower	Upper
1	(Constant)	12.034	.022	.657	.001	10.606	13.338
	CongValue	1.453	-.021	.312	.001	.850	2.004
2	(Constant)	12.995	.041	.733	.001	11.446	14.491
	CongValue	.876	-.028	.399	.040	.050	1.643
	CongPolarity	.961	.014	.332	.005	.255	1.624
3	(Constant)	12.995	.044	.716	.001	11.444	14.547
	CongValue	.876	-.028	.372	.041	.107	1.562
	CongPolarity	.961	.010	.306	.008	.293	1.611
	RminC	-.327	.002	.114	.015	-.537	-.092

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

CongValue = ABS(R-C). CongPolarity: 0= RC Polarity congruent; 1= RC polarity incongruent.
RminC = R – C.

Results

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.626[a]	.392	.380	2.21538
2	.670[b]	.449	.426	2.13080
3	.710[c]	.505	.472	2.04329

a. Predictors: (Constant), CongValue

b. Predictors: (Constant), CongValue, CongPolarity

c. Predictors: (Constant), CongValue, CongPolarity, RminC

CongValue = ABS(R-C). CongPolarity: 0= RC Polarity congruent; 1= RC polarity incongruent. RminC = R – C.

Discussion

H1: Judged accuracy will increase with reliability and credibility

- H1 confirmed, and information credibility had only a slightly larger effect size than source reliability (cf. Samet, 1975).
- There was a polarity effect such that moving from a positive polarity term to a negative polarity term led to the largest decline in assessed probability.
- The latter finding suggests that a scale without a polarity change in the middle might be better. This could be tested (next?).

Discussion

H2: Analysts will not be susceptible to the secrecy heuristic, and classification will have little or no effect on accuracy.

- H2 disconfirmed; there was a small effect of classification on assessed accuracy in line with the secrecy heuristic.
- The magnitude of the effect was not predictable on the basis of years of experience, age, or gender.

Discussion

H3: The test-retest reliability of analysts will be proportional to the congruence of the reliability and credibility scales.

- H3 confirmed; low vs. high RC-value congruence yielded a medium size effect on test-retest reliability.
- The result is impressive given that retest happened within a single 30 min session.
- What does this say about reliability in the field over much longer timeframes?

Discussion

H4: The *inter*-analyst reliability of accuracy judgments will be proportional to RC value congruence.

- H4 confirmed. Inter-analyst variability increased with RC value incongruence.
- Variability also increased with RC *polarity* incongruence (i.e., when one scale had positive polarity and the other had negative polarity).
- Variability also increased with information credibility exceeded source reliability (consistent with the view that R enables C).

Extra material: Specific resampled cases used for test-retest reliability

- Completely reliable, confirmed by other sources
- Completely reliable, possibly true
- Completely reliable, improbable
- Fairly reliable, confirmed by other sources
- Fairly reliable, possibly true
- Fairly reliable, improbable
- Unreliable, confirmed by other sources
- Unreliable, possibly true
- Unreliable, improbable
- Reliability cannot be judged, truth cannot be judged