Alpine
Testing Solutions

Security Deep Dive

Russell W. Smith, Ph.D.
VP Assessment Services
Senior Psychometrician

…you likely have a problem. So… now what?

- Differential Person Functioning (DPF)
- Differential Item Functioning (DIF & Drift)
- DPF & DIF used in conjunction
- Answer Similarity Index (ASI) Analyses
- Score Similarity Index (SSI) Analyses
- Person Residual Correlations (B3)
- Cluster Analyses

Eckerly, C., Smith, R. W., Lee, Y-H. (2018). An Introduction to Item Pre-Knowledge Detection with Real Data Applications. Presented at the 7th Annual Conference on Test Security. Park City, UT.

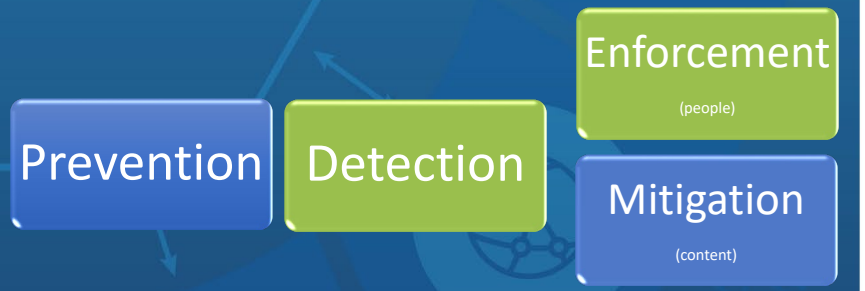
Foley, B. P. (2019). Collusion Detection Using an Extension of Yen's Q3 Statistic. Presented at the 8th Annual Conference on Test Security. Miami, FL.

O'Leary, L. S., & Smith, R. W. (2017). Detecting candidate preknowledge and compromised content using differential person and item functioning. In G. J. Cizek and J. A. Wollack, Eds., *Handbook of quantitative methods for detecting cheating on tests*. New York, NY: Routledge.

Smith, R. W. (2019). Comparing B3 to Answer Similarity Index for Detecting Collusion. Presented at the 8th Annual Conference on Test Security. Miami, FL.

Wollack, J. A. & Maynes, D. D. (2017). Detection of test collusion using cluster analysis. In G. J. Cizek and J. A. Wollack, Eds., *Handbook of quantitative methods for detecting cheating on tests*. New York, NY: Routledge.

Differential Person Functioning (DPF)



- Compare **a person's** performance on sub-sets of items, e.g.:
  - new/old
  - scored/unscored
  - exposed/unexposed
    - based on exposure rates or
    - based on found content
  - performance/selected response
- Probabilistic
- Validity focused
- Actionable
  - Use Contrast AND Probability
- Conceptually easy, computationally easy

DPF Examples

Prevention

Detection

Enforcement
(people)

Mitigation
(content)

DPF Examples

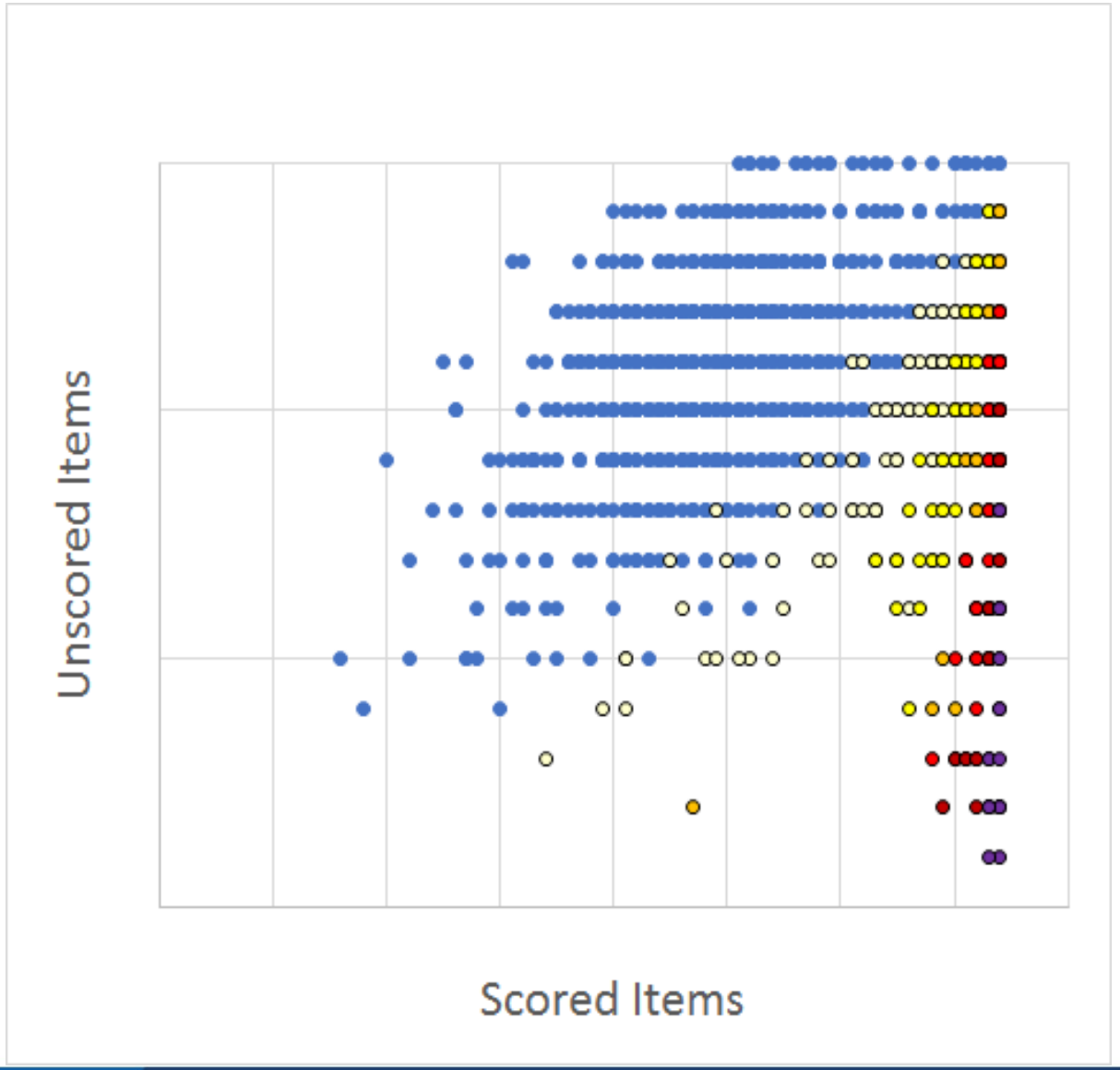
DPF Examples

Prevention

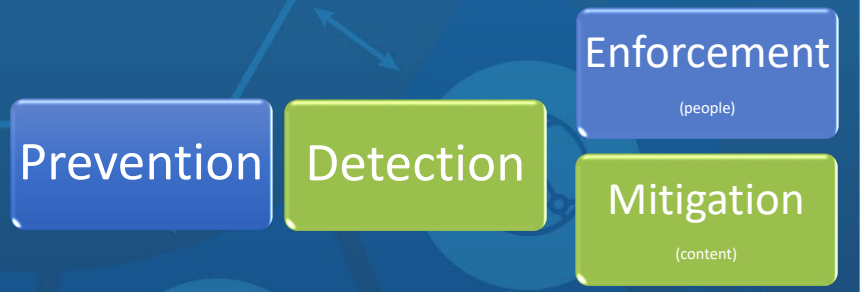
Detection

Enforcement
(people)

Mitigation
(content)



Differential Item Functioning (DIF)



- Compare **a item's** performance on sub-sets of people, e.g.:
  - Demographics, e.g. race, gender
  - Geography
  - Early test takers vs. recent (DRIFT)
  - Test takers identified as likely having pre-knowledge vs. those who weren't identified
- Probabilistic
- Validity and fairness focused
- Actionable
  - Mitigate by identifying and replacing exposed or biased content
- Conceptually easy, computationally easy

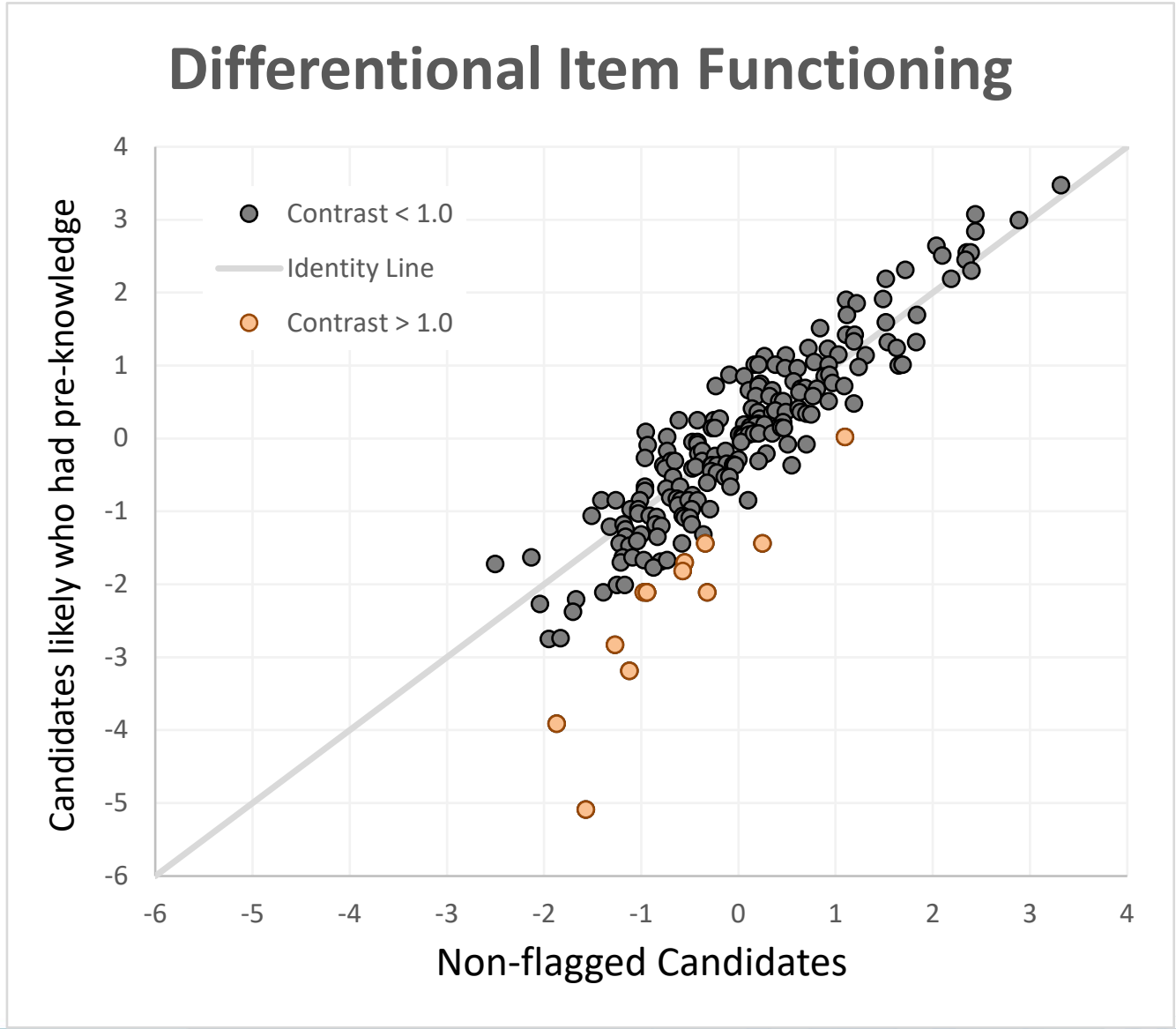
DIF Examples

Prevention

Detection

Enforcement
(people)

Mitigation
(content)



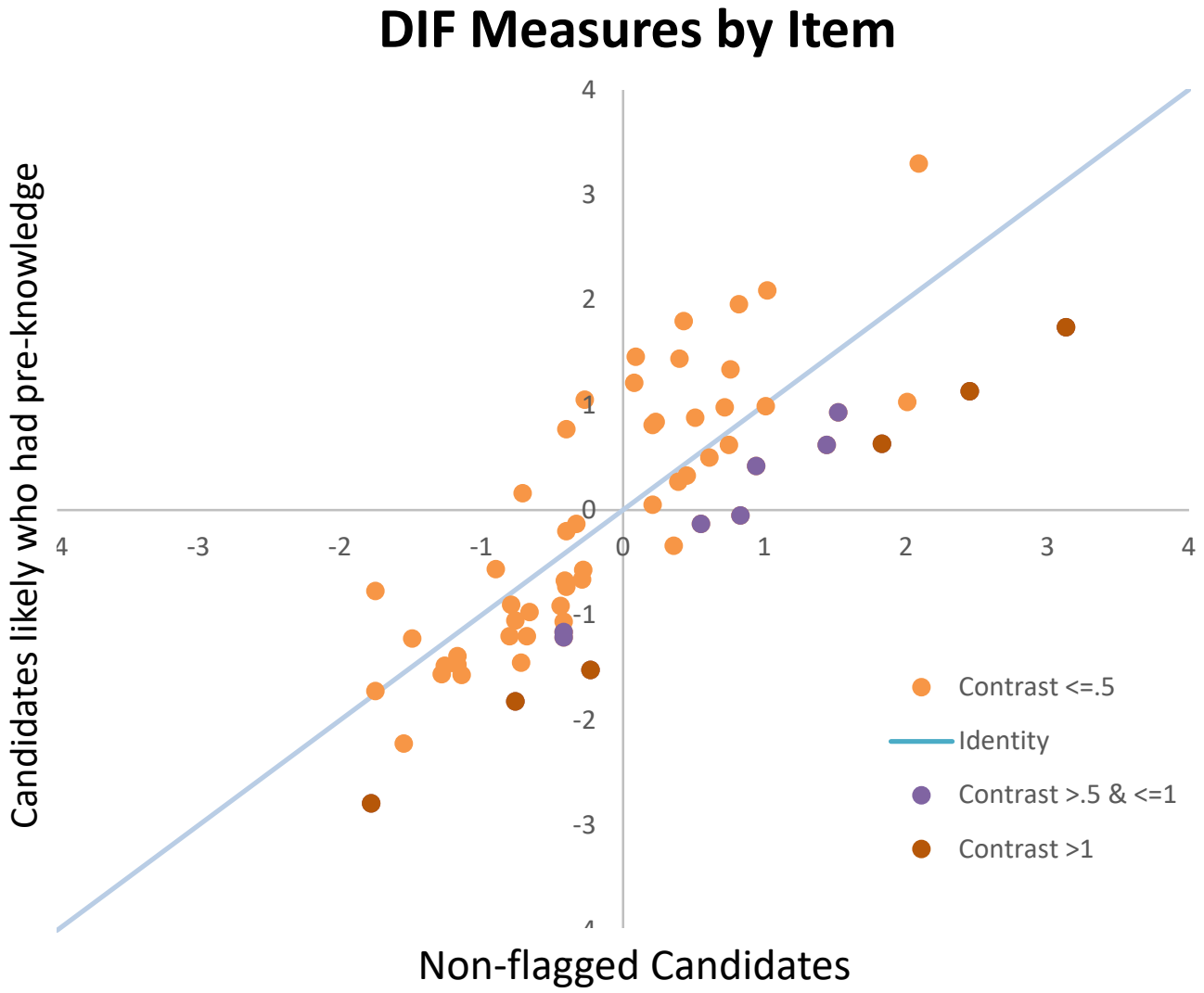
DIF Examples

Prevention

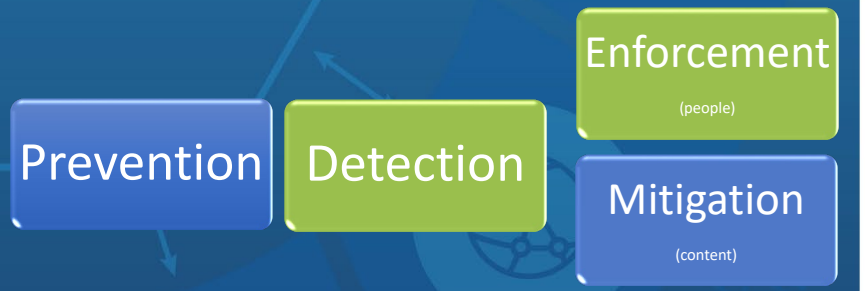
Detection

Enforcement
(people)

Mitigation
(content)



Answer similarity index (ASI)
Score similarity index (SSI)
Person Residual Correlations (B3)



- Collusion statistics
- Compare **a person's** scores/answers to another **person's** scores/answers
- Actionable, particularly when combined with cluster analyses
- Conceptually can be somewhat challenging and computationally intense
- Don't work well at/near 100% scores
- ASI: single response 4 or 5 option MC
  - Compares **responses/answers**
  - Model: Bock's NRM & generalized binomial
  - Has an underlying distribution
- SSI: dichotomous items
  - Compares **scores**
  - Model: Rasch or IRT & generalized binomial
  - Has an underlying distribution
- B3: any items
  - Correlates **score** residuals
  - Model: Rasch or IRT, raw or standardized residuals
  - Does NOT have an underlying distribution

ASI Steps

- Calibrate the items (responses) and people
- Calculate the ind. prob. for each person on each item
- Calculate the joint prob. of a matching response for every person with every other person for every item
  - Which includes the person ability and the likelihood of each response for each item
- For each pair of people
  - Calculate the expected number of matches
  - Count the actual number of matches
  - Calculate the probability of the actual number given the expected number

RegID1	RegID2	Theta 1	Theta 2	Expected	Matching	Prob
1	2	3.6021	4.1727	191.90	193	0.4356
1	3	3.6021	3.7555	191.15	194	0.2146
1	4	3.6021	2.8610	188.01	190	0.3417
1	5	3.6021	4.0331	191.68	192	0.5470
1	6	3.6021	3.0426	188.88	188	0.6788
1	7	3.6021	4.5391	192.35	194	0.3535
1	8	3.6021	4.5109	192.32	195	0.2168
1	9	3.6021	2.7577	187.44	188	0.5104
1	10	3.6021	3.3095	189.92	189	0.6907

Bock, R.. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Pyschometrika 37, 29-51.

van der Linden, W. J. & Hambleton, R. K. (1997), Handbook of modern item response theory Springer-Verlag, New York.

SSI Steps

- Calibrate the items ~~(responses)~~ and people
- Calculate the ind. prob. for each person on each item
- Calculate the joint prob. of a matching score for every person with every other person for every item
  - Which includes the person ability and the likelihood of a correct response for each item
- For each pair of people
  - Calculate the expected number of matches
  - Count the actual number of matches
  - Calculate the probability of the actual number given the expected number

RegID1	RegID2	Theta 1	Theta 2	Expected	Matching	Prob
1	2	3.6021	4.1727	191.90	193	0.4356
1	3	3.6021	3.7555	191.15	194	0.2146
1	4	3.6021	2.8610	188.01	190	0.3417
1	5	3.6021	4.0331	191.68	192	0.5470
1	6	3.6021	3.0426	188.88	188	0.6788
1	7	3.6021	4.5391	192.35	194	0.3535
1	8	3.6021	4.5109	192.32	195	0.2168
1	9	3.6021	2.7577	187.44	188	0.5104
1	10	3.6021	3.3095	189.92	189	0.6907

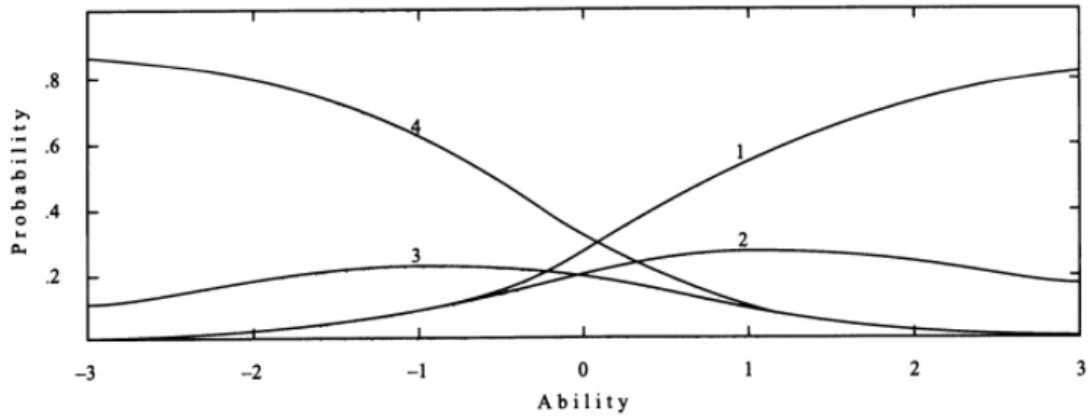
Bock, R.. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Pyschometrika 37, 29-51.

van der Linden, W. J. & Hambleton, R. K. (1997), Handbook of modern item response theory Springer-Verlag, New York.

ASI: Bock's (1972,1997) Nominal Response Model

Limitations:

- Types of items, 4 or 5 option MC
- Software
- Sample size requirements
- The model itself, with one monotonically increasing and one monotonically decreasing response
- When the model fits, it works well

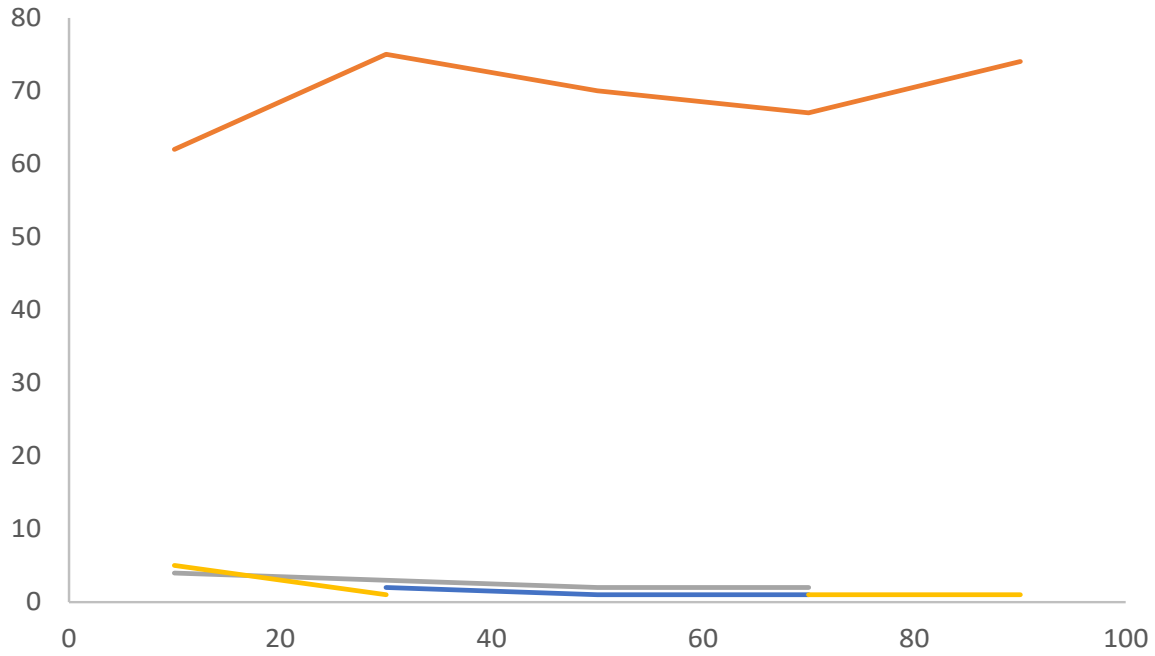


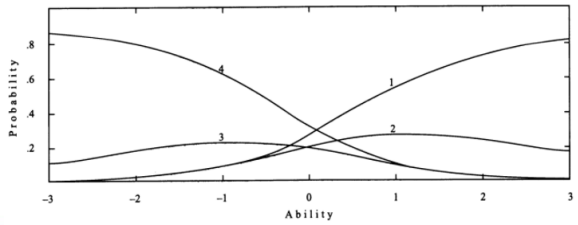
$$P_{u_g}(\theta) = \frac{\exp(a_{u_g}\theta + c_{u_g})}{\exp(a_{1_g}\theta + c_{1_g}) + \exp(a_{2_g}\theta + c_{2_g}) + \ldots + \exp(a_{m_g}\theta + c_{m_g})}$$

Bock, R.. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Pyschometrika 37, 29-51.

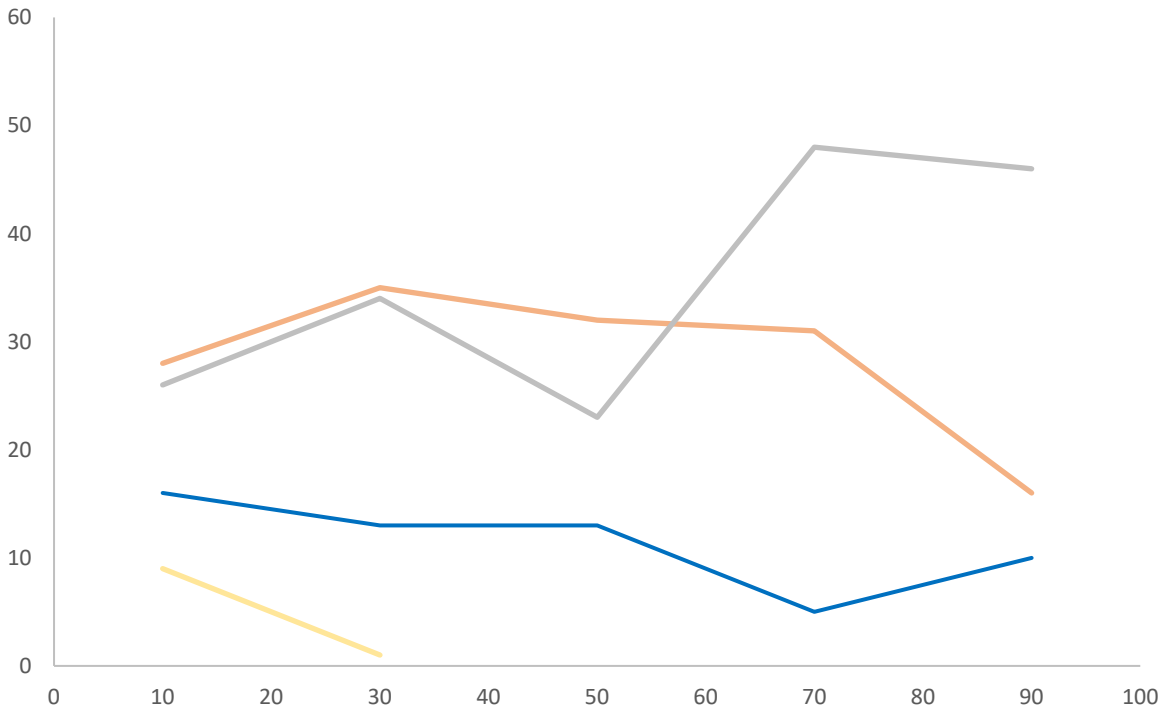
van der Linden, W. J. & Hambleton, R. K. (1997), Handbook of modern item response theory Springer-Verlag, New York.

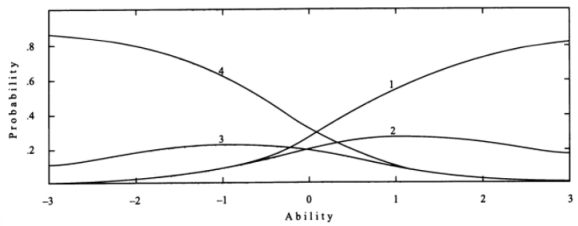
option	p-value	correlation	avg. time	0-20	21-40	41-60	61-80	81-100
A	0.011	-0.006	35		2	1	1	
B	0.926	0.303	26	62	75	70	67	74
C	0.029	-0.112	37	4	3	2	2	
D	0.021	-0.103	69	5	1		1	1



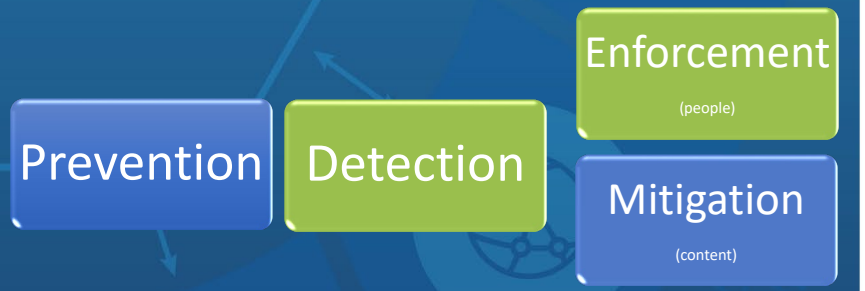


option	p-value	correlation	avg. time	0-20	21-40	41-60	61-80	81-100
A	0.145	-0.073	66	16	13	13	5	10
B	0.361	-0.052	59	28	35	32	31	16
C	0.450	0.238	54	26	34	23	48	46
D	0.028	-0.158	48	9	1			1





Score similarity index (SSI)
Person Residual Correlations (B3)



		Perfect Fit		Reasonable		Unexpected	
		PersonA	PersonB	PersonC	PersonD	PersonE	PersonF
Item	Measure	1	0.5	1	0.5	1	0.5
Item001	-1	1	1	1	1	0	0
Item002	-0.75	1	1	1	1	0	0
Item003	-0.5	1	1	1	1	1	0
Item004	-0.25	1	1	1	1	1	0
Item005	-0.1	1	1	0	0	1	1
Item006	0.1	1	1	1	1	1	1
Item007	0.25	1	0	1	0	1	1
Item008	0.5	1	0	1	1	1	1
Item009	0.75	0	0	0	0	1	1
Item010	1	0	0	1	0	1	1
	Score	8	6	8	6	8	6
	Number matching	8		8		8	
	Person Residual Correlation/B3	0.3789		0.5274		0.7468	
	SSI Binomial Prob	0.1389		0.1389		0.1389	

Score similarity index (SSI)
Person Residual Correlations (B3)

$$P(\theta) = \frac{e^{(\theta - b)}}{1 + e^{(\theta - b)}}$$

Alpine Testing Solutions

		Perfect Fit				Perfect Fit	
		PersonA	PersonB			PersonE	PersonF
Item	Measure	1	0.5	Expected (1)	Expected (.5)	Residual (1)	Residual (.5)
Item001	-1	1	1	0.881	0.818	-0.119	-0.182
Item002	-0.75	1	1	0.852	0.777	-0.148	-0.223
Item003	-0.5	1	1	0.818	0.731	-0.182	-0.269
Item004	-0.25	1	1	0.777	0.679	-0.223	-0.321
Item005	-0.1	1	1	0.750	0.646	-0.250	-0.354
Item006	0.1	1	1	0.711	0.599	-0.289	-0.401
Item007	0.25	1	0	0.679	0.562	-0.321	0.562
Item008	0.5	1	0	0.622	0.500	-0.378	0.500
Item009	0.75	0	0	0.562	0.438	0.562	0.438
Item010	1	0	0	0.500	0.378	0.500	0.378
	Score	8	6				
						Correlation	
						0.3789	

Prevention
Detection
Enforcement (people)
Mitigation (content)

Score similarity index (SSI)
Person Residual Correlations (B3)

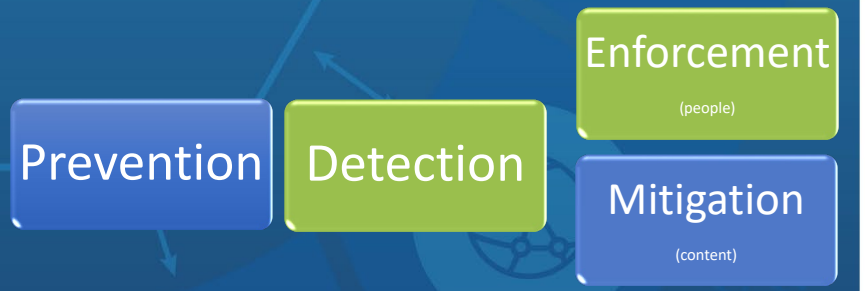
$$P(\theta) = \frac{e^{(\theta - b)}}{1 + e^{(\theta - b)}}$$

Alpine Testing Solutions

Item	Measure	Expected (1)	Expected (.5)	Unexpected PersonE Residual (1)	Unexpected PersonF Residual (.5)	Unexpected PersonE 1	Unexpected PersonF 0.5
Item001	-1	0.881	0.818	0.881	0.818	0	0
Item002	-0.75	0.852	0.777	0.852	0.777	0	0
Item003	-0.5	0.818	0.731	-0.182	0.731	1	0
Item004	-0.25	0.777	0.679	-0.223	0.679	1	0
Item005	-0.1	0.750	0.646	-0.250	-0.354	1	1
Item006	0.1	0.711	0.599	-0.289	-0.401	1	1
Item007	0.25	0.679	0.562	-0.321	-0.438	1	1
Item008	0.5	0.622	0.500	-0.378	-0.500	1	1
Item009	0.75	0.562	0.438	-0.438	-0.562	1	1
Item010	1	0.500	0.378	-0.500	-0.622	1	1
				Correlation			
				0.7468			

Prevention
Detection
Enforcement (people)
Mitigation (content)

Score similarity index (SSI)
Person Residual Correlations (B3)



		Perfect Fit		Reasonable		Unexpected	
		PersonA	PersonB	PersonC	PersonD	PersonE	PersonF
Item	Measure	1	0.5	1	0.5	1	0.5
Item001	-1	1	1	1	1	0	0
Item002	-0.75	1	1	1	1	0	0
Item003	-0.5	1	1	1	1	1	0
Item004	-0.25	1	1	1	1	1	0
Item005	-0.1	1	1	0	0	1	1
Item006	0.1	1	1	1	1	1	1
Item007	0.25	1	0	1	0	1	1
Item008	0.5	1	0	1	1	1	1
Item009	0.75	0	0	0	0	1	1
Item010	1	0	0	1	0	1	1
	Score	8	6	8	6	8	6
	Number matching	8		8		8	
	Person Residual Correlation/B3	0.3789		0.5274		0.7468	
	SSI Binomial Prob	0.1389		0.1389		0.1389	

SSI with cluster analysis

P-values	Examinee 1	Examinee 2	Examinee 3	Examinee 4
Examinee 2	0.000001			
Examinee 3	0.000001	0.000001		
Examinee 4	0.000001	0.000001	0.000001	
Examinee 5	0.000001	0.000001	0.000001	0.000001

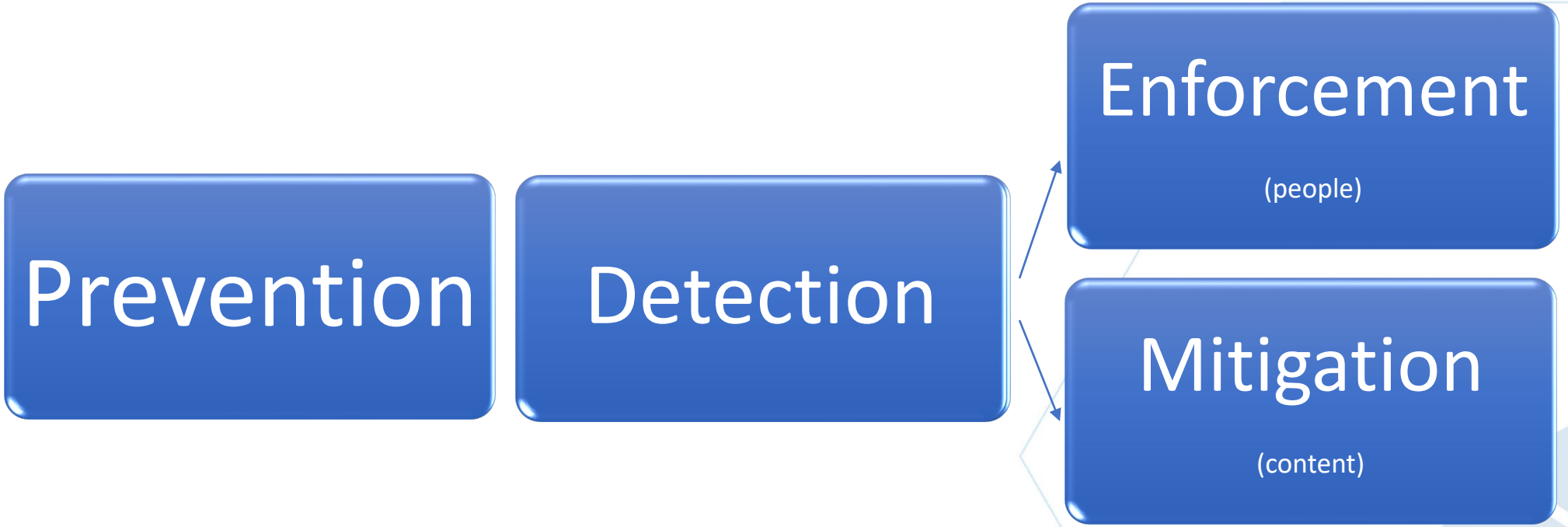
Actual No. Matching	Examinee 1	Examinee 2	Examinee 3	Examinee 4
Examinee 2	60			
Examinee 3	60	60		
Examinee 4	60	60	60	
Examinee 5	60	60	60	60

Expected No. Matching	Examinee 1	Examinee 2	Examinee 3	Examinee 4
Examinee 2	48			
Examinee 3	48	48		
Examinee 4	48	48	48	
Examinee 5	48	48	48	48

Registration No.	Date	Test Center	Score
Examinee 1	2/29/2016	20210	52
Examinee 2	3/2/2016	20210	52
Examinee 3	3/7/2016	20210	52
Examinee 4	2/29/2016	20210	52
Examinee 5	6/13/2016	20195	52

Cluster size	No. of clusters
2 examinees	16
3 examinees	4
4 examinees	4
5 examinees	4
8 examinees	1
9 examinees	2
11 examinees	1
159 examinees	1

1992 examinees -> 1,983,036 pairwise comparisons
Flagging Threshold: 0.000001
Flagging Rate (pairs): 0.0018


Prevention
Detection
Enforcement
(people)
Mitigation
(content)

…you likely have a problem. So… now what?

- Differential Person Functioning (DPF)
- Differential Item Functioning (DIF & Drift)
- DPF & DIF used in conjunction
- Answer Similarity Index (ASI) Analyses
- Score Similarity Index (SSI) Analyses
- Person Residual Correlations (B3)
- Cluster Analyses

Eckerly, C., Smith, R. W., Lee, Y-H. (2018). An Introduction to Item Pre-Knowledge Detection with Real Data Applications. Presented at the 7th Annual Conference on Test Security. Park City, UT.

Foley, B. P. (2019). Collusion Detection Using an Extension of Yen's Q3 Statistic. Presented at the 8th Annual Conference on Test Security. Miami, FL.

O'Leary, L. S., & Smith, R. W. (2017). Detecting candidate preknowledge and compromised content using differential person and item functioning. In G. J. Cizek and J. A. Wollack, Eds., *Handbook of quantitative methods for detecting cheating on tests*. New York, NY: Routledge.

Smith, R. W. (2019). Comparing B3 to Answer Similarity Index for Detecting Collusion. Presented at the 8th Annual Conference on Test Security. Miami, FL.

Wollack, J. A. & Maynes, D. D. (2017). Detection of test collusion using cluster analysis. In G. J. Cizek and J. A. Wollack, Eds., *Handbook of quantitative methods for detecting cheating on tests*. New York, NY: Routledge.

To Summarize

- Security is about Validity!
- Plan for it
- Avoid it (prevention)
- Look at a high level, then (detection)…

- Including candidates (enforcement) and content (mitigation)

Questions?

russell.smith@alpinetesting.com