



***Development of an Internet-based,
Multidimensional Computer Adaptive Testing
Platform***

Adaptive Assessment Services, Inc.

Damon Bryant and Kenneth James, Jr.



Overview

- *Introduction*
- *Research Questions*
- *Method*
- *Results*
- *Conclusion*
- *Questions*
- *Abstract and Reference*



Introduction

- *More and more organizations are moving assessment systems to the Internet:*
 - *Military*
 - *Employment*
 - *Certification*
 - *Education*
- *There are a variety of benefits that can be gained from Internet-based testing:*
 - *Administration Time*
 - *Scoring*
 - *Perception of Being Innovative*
 - *Faster Decision Support Information*

Introduction

- *There are also a variety of threats to test score interpretation*
 - *Verification of examinee identity (International Test Commission, 2000)*
 - *Organized item theft rings and cheating*
 - *ETS*
 - *IBM*
 - *Atlanta, Detroit, DC, Los Angeles*
 - *Tests measuring multiple relevant dimensions with only one reported score (Reckase, 1985, 2009)*
 - *Tests measuring irrelevant dimensions on a test with only one reported score (Bryant 2007)*

Introduction

- *There are specific threats associated with computer adaptive testing (CAT) despite claims of increased test security*
 - *Item overexposure (Way, 1998)*
 - *Between-test overlap (Chen, Ackenmann & Spray, 2003)*
 - *Item underexposure (low discrimination and high guessing)*
 - *Item pool utilization (proportion of test items used in operational pool)*
- *Multidimensional Item Response Theory, Computer Adaptive Testing (MIRT CAT) can address some concerns about validity of test score interpretation*
 - *Multiple scores and composite scores can be reported in a direction (Bryant, 2005; Bryant & Davis, 2011; Reckase, 2009)*
 - *Irrelevant or nuisance dimensions can be measured and weighted accordingly*

Introduction

- *However, the aforementioned security threats from a psychometric viewpoint remain despite research focused on accuracy of estimating theta vector using MIRT CAT algorithms (Chang, 2011a; Chang, 2011b; Reckase, 2009; van der Linden, 2010)*
- *There should be a focus on psychometric test security in MIRT CAT.*
- *Little published work has been done to address the issue of exposure control, overlap rates, and item pool utilization in MIRT CAT*
 - Generalized Stocking-Lewis Method (Finkelman, Nering, & Roussos, 2009)



Introduction

- *Purpose: To develop a Internet-based, MIRT CAT system and evaluate the extent to which the algorithm reduces threats to test security and test score interpretation.*

Introduction

- *Simulation study uses Smart Test Technology[®] (STT) MIRT CAT engine*
 - *Multidimensional Item Response Theory (Multidimensional 3-PLM)*
 - *Probability Model [Reckase (1985, 2009), Reckase & McKinley (1991)]*
 - *Directional Information Function (Bryant, 2005)*
 - *Theta Maximum in a Direction (Bryant, 2005; Bryant & Davis, 2011)*
 - *D-Optimality (see Segall, 1996 for a detailed description)*

Model

*Multidimensional IRT (Bryant, 2005; Bryant & Davis, 2011
Reckase, 1985, 2009)*

M3-PL

Item response function $P_i(\boldsymbol{\theta}_j) = c_i + (1 - c_i) \{1 + \text{Exp}[-D(\mathbf{a}_i' \boldsymbol{\theta} + d_i)]\}^{-1}$

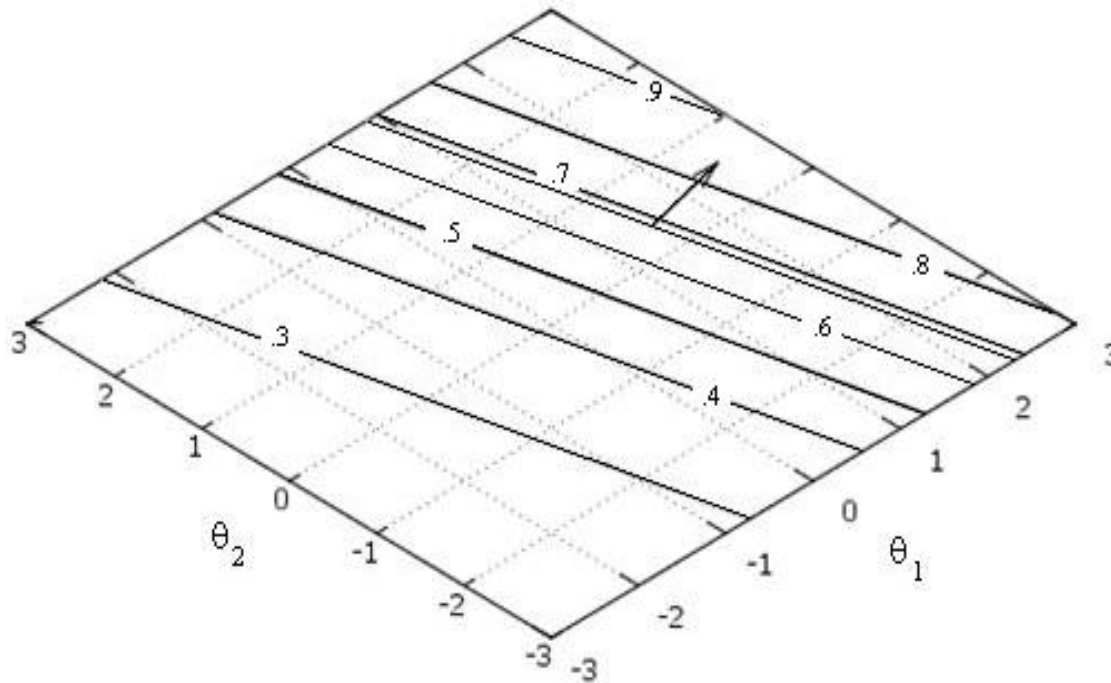
Item information $I_{ii}(\boldsymbol{\theta}) = D^2 (\mathbf{a}_i' \mathbf{u}_i)^2 Q_i(\boldsymbol{\theta}) \{P_i(\boldsymbol{\theta}) [1 + \text{Exp}(-L)]^2\}^{-1}$

Theta maximum $\boldsymbol{\theta}_{\max} = [\ln\{.5 [1 + (8c_i + 1)^{1/2}]\} (D \cdot MDISC_i)^{-1} + MDIFF_i] \mathbf{u}_i$

Note. k = number of dimensions, $\mathbf{1}$ is a $k \times 1$ vector of ones. $D = 1.7$, \mathbf{a}_i is a $k \times 1$ vector of discrimination parameters for item i , $[a_{1i}, \dots, a_{ki}]'$, $\boldsymbol{\theta}$ is a vector of k ability parameters, $[\theta_{1j}, \dots, \theta_{kj}]'$, d_i is a scalar related to difficulty, $L = D(\mathbf{a}_i' \boldsymbol{\theta} + d_i)$, $MDISC_i = \|\mathbf{a}_i\|$, $MDIFF_i = -d_i / \|\mathbf{a}_i\|$, and \mathbf{u} is a vector of directional cosines, $\mathbf{a}_i / \|\mathbf{a}_i\|$ or $[\cos \alpha_{1i}, \dots, \cos \alpha_{ki}]'$.

Model

Multidimensional Representation of Directional θ max
 Item i: $a_1 = .97$, $a_2 = .23$, $d = -1.25$, $c = .25$, $\theta_{\max} = [1.52, .36]$



Research Questions



- *How do the different item selection methods compare in terms of*
 - (1) Bias of Theta*
 - (2) Error in Estimating Theta Vector (Euclidian Distance)*
 - (3) Item Exposure*
 - (4) Item Pool Utility*
 - (5) Overlap Rate?*

Method



- *STT Web-based Framework*
 - *Operating System: Linux*
 - *Web Tier: Apache and Glassfish Web Server*
 - *Database: MySQL*
 - *Item Selection Algorithms and Objects: Python and Java (with Java Beans)*
 - *Web-based User Interface: HTML and Java Server Faces*
- *Computer Simulation of Software Performance*
 - *Created using object-oriented, programming languages Python and Java*

Method

Independent Variables (3x4x5)

- *Item Pool Size*
 - *300, 400, and 500 items*
- *Item Selection Criteria*
 - *Random Item Selection*
 - *Maximum Directional Information (45 degrees, 45 degrees)*
 - *Directional Theta Maximum (45 degrees, 45, degrees)*
 - *D-Optimality (Maximize Determinant of Fisher's Information, Segall, 1996)*
- *Length of Adaptive Test*
 - *20, 25, 30, 35, and 40 items*

Method

Dependent Variables

(1) Bias of θ_x :

$$(\theta_{Estimated\ i} - \theta_{True\ i})$$

(2) Error in Estimating Theta Vector (Euclidian Distance) :

$$||\theta_{Estimated} - \theta_{True}||$$

(3) Item Exposure Rate:

$Exposure_i = \text{Total \# of Exposures for Item } i / \text{Total \# of Test Administrations}$

$Max[Exposure_i]$

(4) Item Pool Utility

$\text{Total \# of Exposed Items} / \text{Total \# of Items in Item Pool}$

(5) Item Overlap Rate (Chen, Ackenmann, Spray, 2003)

$p = \# \text{ CATs administered}$

$k = \text{CAT length}$

$r = \text{Item exposure rate}$

$$\frac{p \sum_{i=1}^n r_i^2}{k(p-1)} - \frac{1}{p-1}$$

Method



Procedures

Examinee Population

MVN, Mean vector $[0,0]^T$ and variance-covariance matrix assuming .5 in off-diagonals

1000 true theta vectors created

Theta Vector (θ) estimated using Expected A Posteriori (EAP) approach

25 Equally spaced quadrature nodes $[-2, 2]$ along each dimension (625 total)

Prior distribution matched generated.

Item Parameters (500) for M3PLM (Reckase, 1985)

\mathbf{a} vector generated: $[U(.5, 1.3), U(.5,1.3)]^T$

Multidimensional difficulty generated: $N(0,1)$

c parameter generated: $U(0,.30)$

$D = 1.7$

Method



Procedures

Number of Simulations

Each condition in the 3x4x5 design was replicated 1000 times (60,000 total simulations)

All 500 item pool conditions were completed first

Randomly deleted 100 items from pool at random to create 400 item pool condition

Repeated step for 300 item pool condition

Standard of evidence ($p < .01$)

Results

- *Bias*: There were no significant differences among the item selection methods in terms of estimation bias of θ_1 or θ_2 (all $p = ns$)
- *Mean Bias* (θ_1, θ_2)
 - *Random* = (-.009, -.005)
 - *Directional Theta Max* = (-.003, -.005)
 - *Directional Information* = (-.004, .000)
 - *D-Optimality* = (.004, .000)

Results

- *Error Estimating Theta Vector:* There was a significant difference among the item selection methods in terms of error in estimating θ .

Tests of Between-Subjects Effects

Dependent Variable: Euclidian Distance

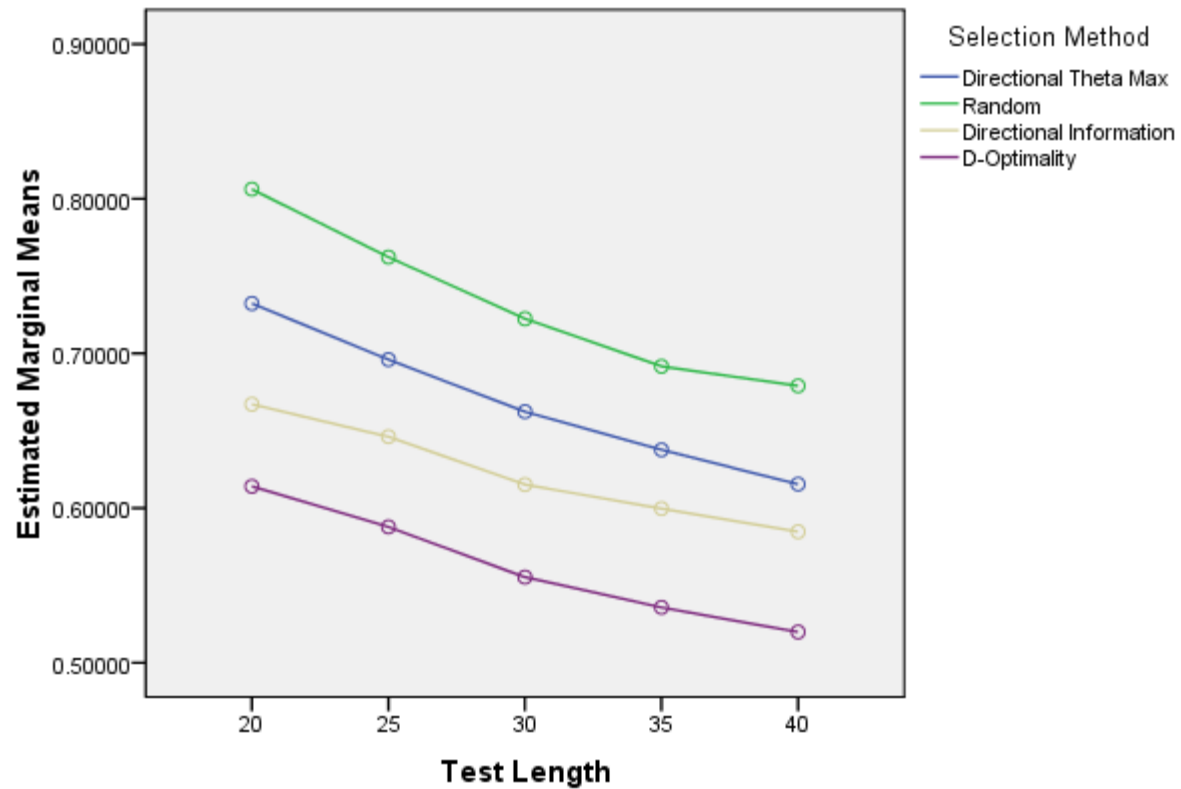
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	326.113 ^a	59	5.527	41.559	.000
Intercept	25081.315	1	25081.315	188580.578	.000
pool	.899	2	.450	3.380	.034
length	87.377	4	21.844	164.243	.000
method	232.197	3	77.399	581.945	.000
pool * length	.754	8	.094	.709	.684
pool * method	.773	6	.129	.969	.445
length * method	2.671	12	.223	1.674	.066
pool * length * method	1.441	24	.060	.451	.990
Error	7972.051	59940	.133		
Total	33379.479	60000			
Corrected Total	8298.164	59999			

a. R Squared = .039 (Adjusted R Squared = .038)

Results



Estimated Marginal Means of Euclidian Distance



Results

- *Exposure Rate*: There were significant differences among the item selection methods in terms of maximum item exposure rates

Tests of Between-Subjects Effects

Dependent Variable: maxExposure

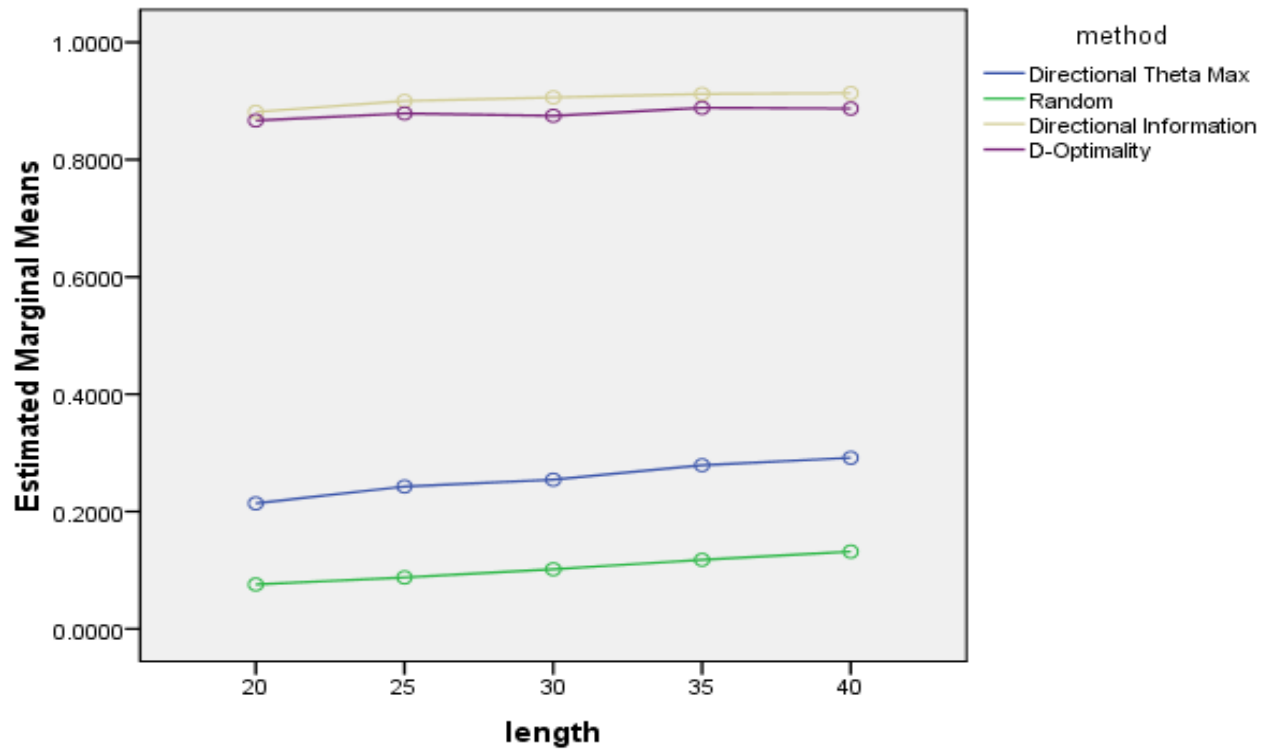
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	7.788 ^a	19	.410	868.835	.000
Intercept	17.186	1	17.186	36428.464	.000
method	7.768	3	2.589	5488.290	.000
length	.016	4	.004	8.620	.000
method * length	.004	12	.000	.710	.733
Error	.019	40	.000		
Total	24.993	60			
Corrected Total	7.807	59			

a. R Squared = .998 (Adjusted R Squared = .996)

Results



Estimated Marginal Means of maxExposure



Results

- *Utility Rate*: There were significant differences among the item selection methods in terms of item pool utility rates

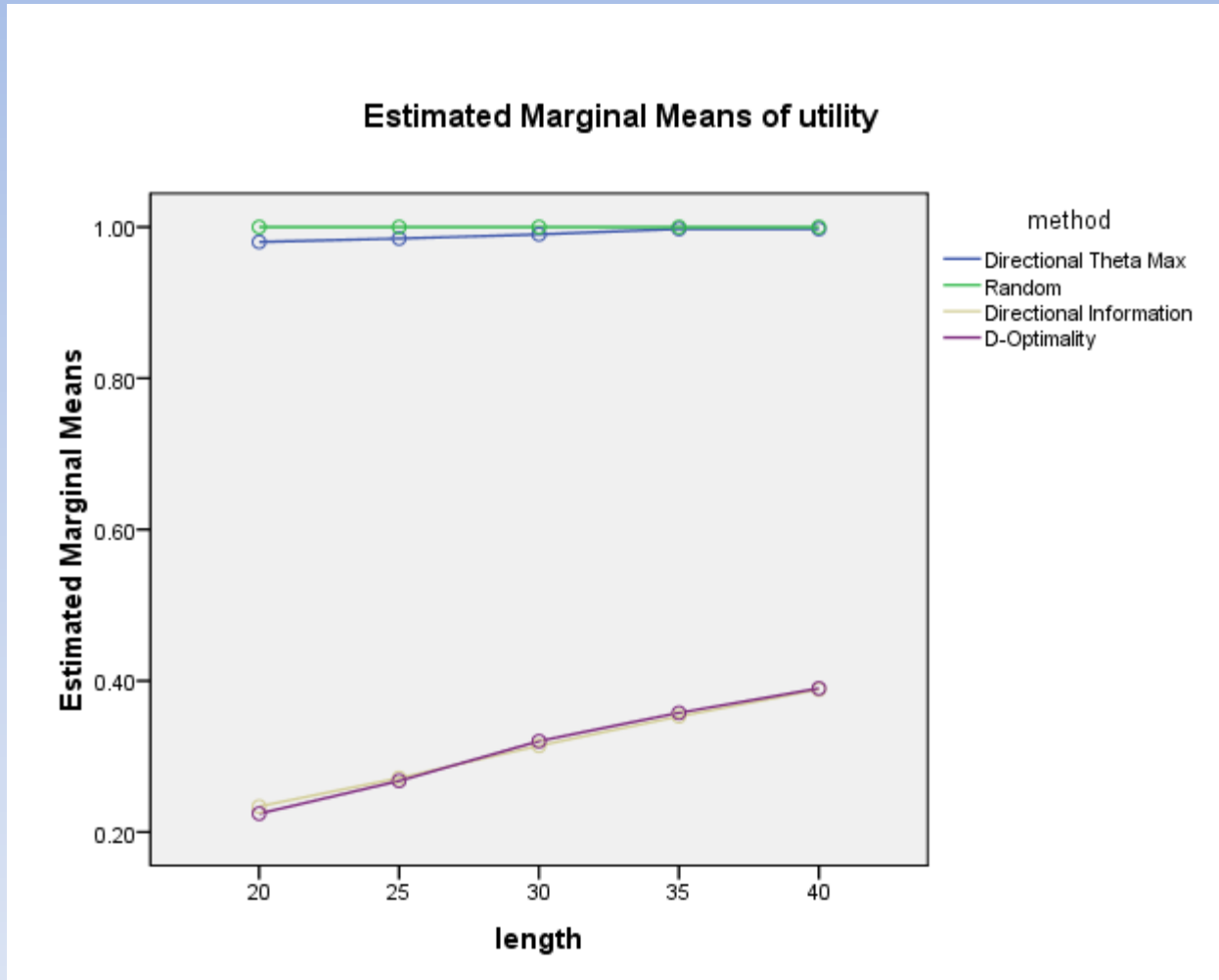
Tests of Between-Subjects Effects

Dependent Variable: utility

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	7.097 ^a	19	.374	210.044	.000
Intercept	25.627	1	25.627	14410.314	.000
method	6.997	3	2.332	1311.504	.000
length	.056	4	.014	7.807	.000
method * length	.045	12	.004	2.091	.040
Error	.071	40	.002		
Total	32.795	60			
Corrected Total	7.168	59			

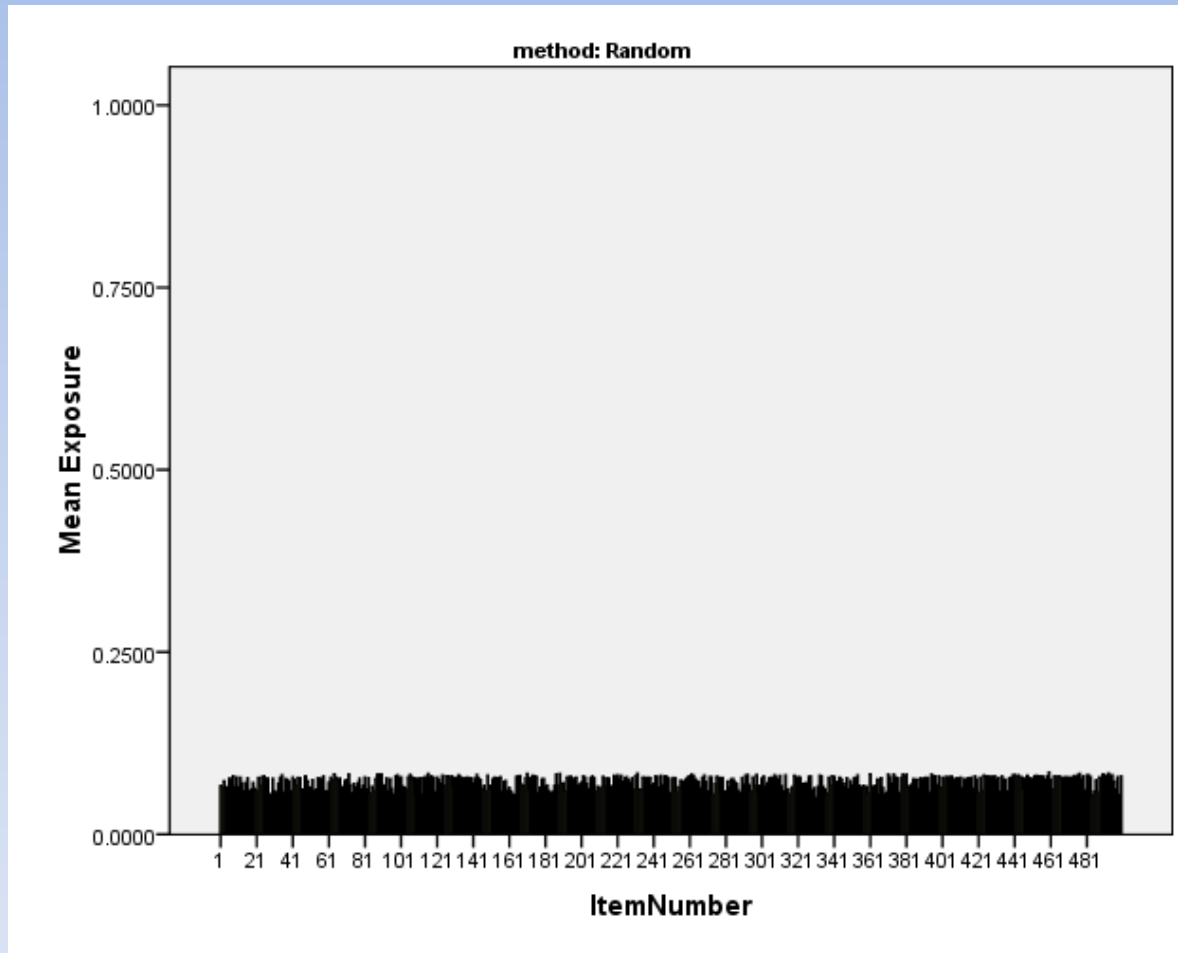
a. R Squared = .990 (Adjusted R Squared = .985)

Results



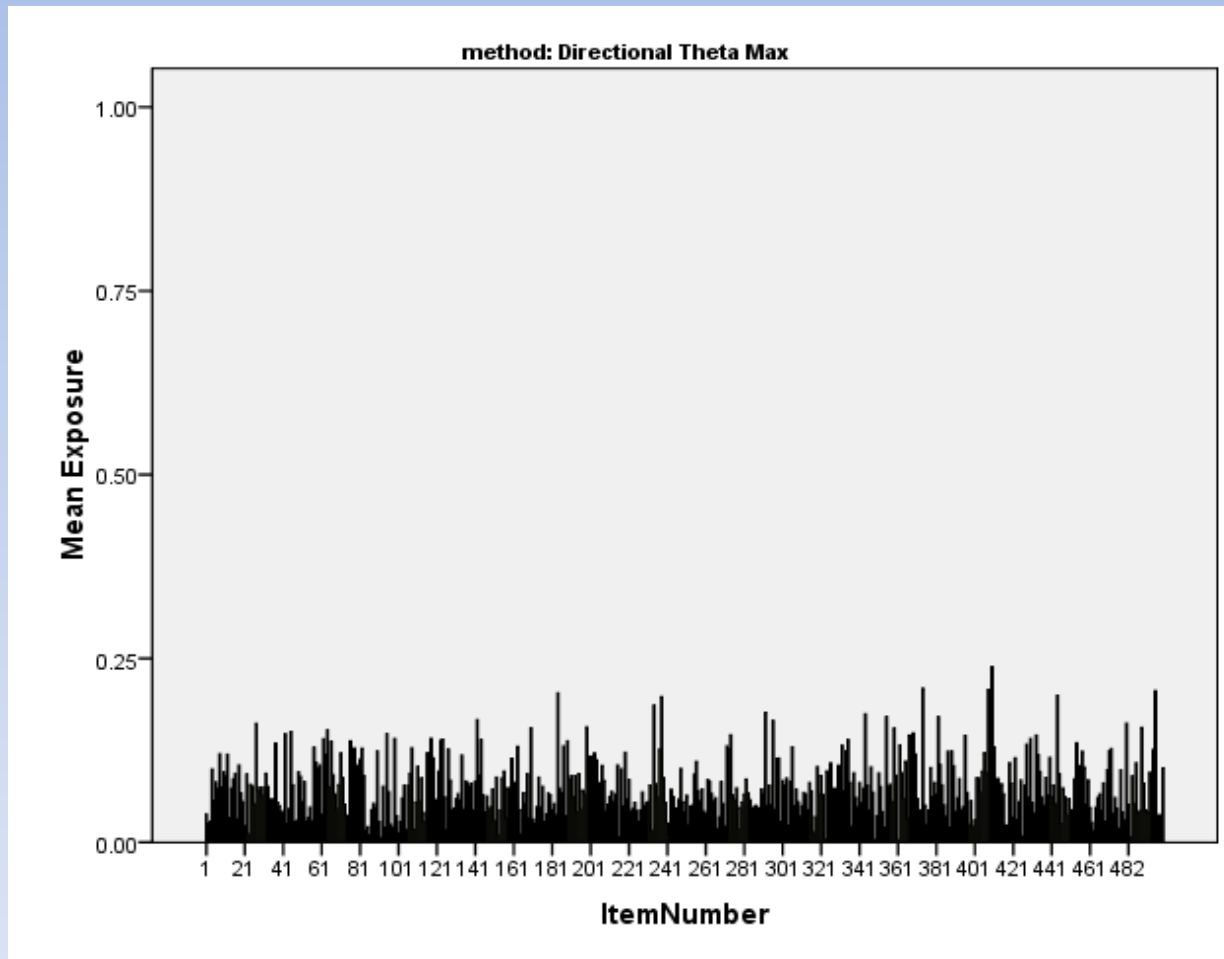
Results

Exposure Rate: Random



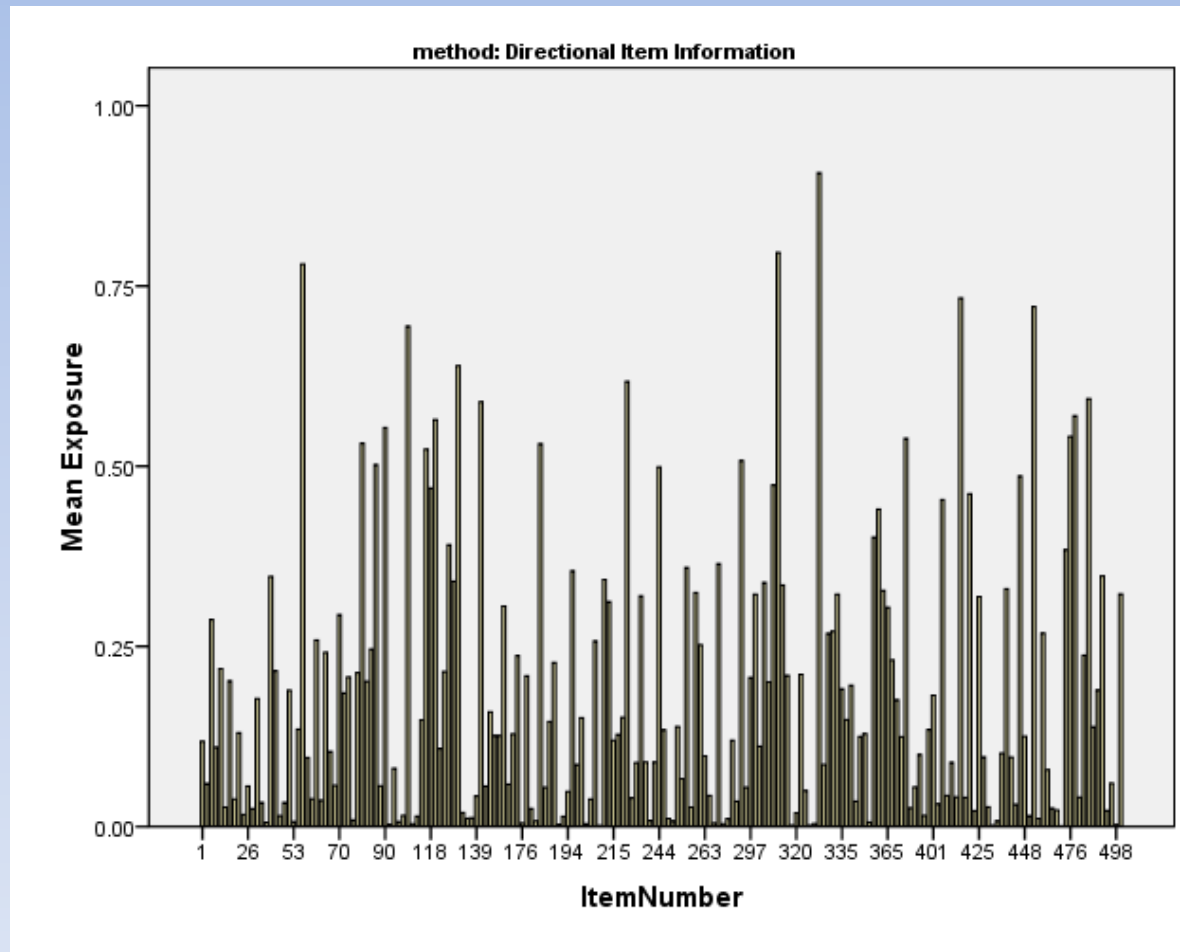
Results

Exposure Rate: Directional Theta Maximum (45, 45)



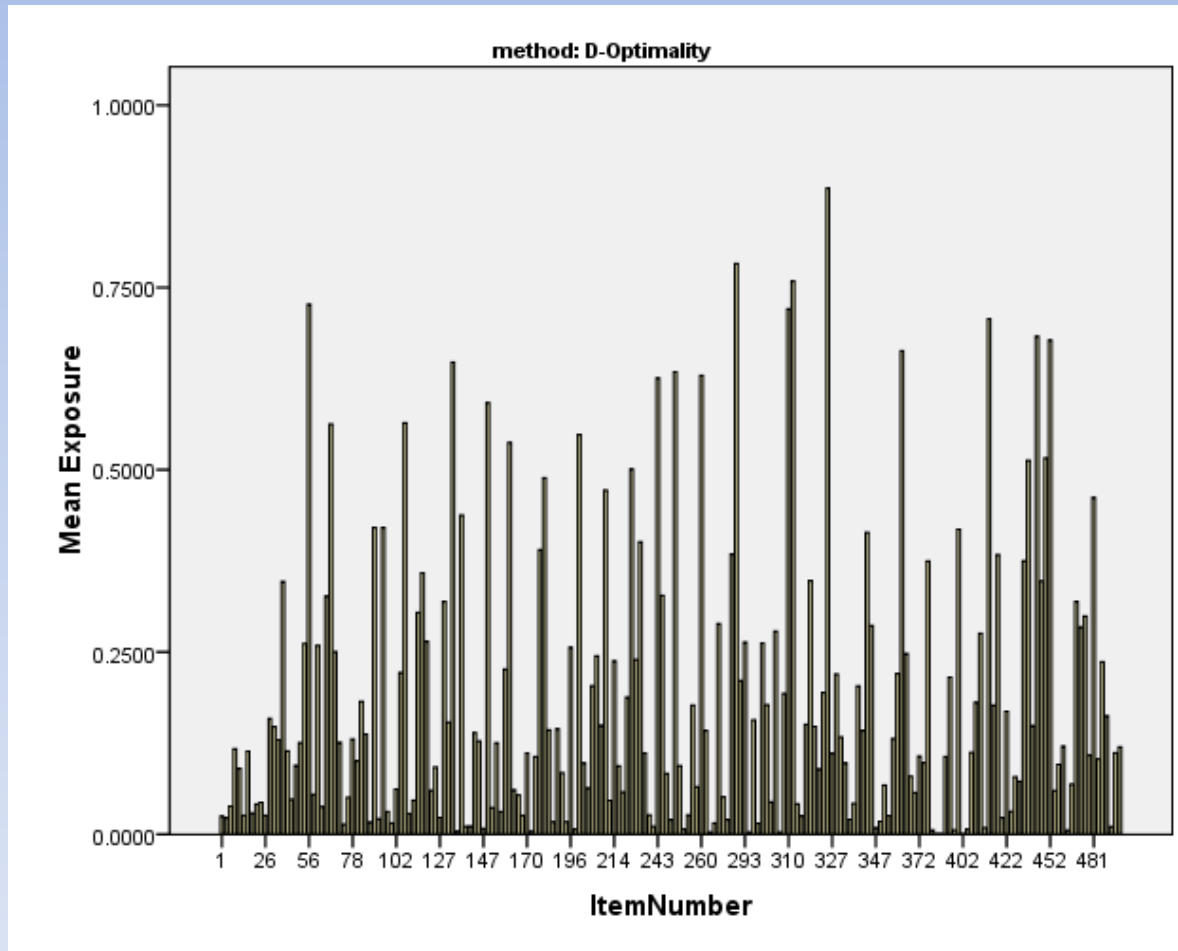
Results

Exposure Rate: Directional Item Information (45, 45)



Results

Exposure Rate: D-Optimality



Results

Conditional Overlap Rate:

Test Length = 20, Pool Size = 500, Equal Weight Composite

Theta Composite	ThetaMax	Random	Directional Information	D-Optimality
	Between Test Overlap Rate	Between Test Overlap Rate	Between Test Overlap Rate	Between Test Overlap Rate
DECILE (n = 100)				
1 (LOWEST)	18%	4%	75%	77%
2	11%	4%	74%	66%
3	8%	4%	76%	70%
4	7%	4%	76%	65%
5	7%	4%	72%	67%
6	7%	4%	75%	71%
7	7%	4%	75%	69%
8	7%	4%	74%	74%
9	7%	4%	71%	73%
10 (HIGHEST)	9%	4%	79%	75%

Results

Conditional Maximum Exposure Rate:

Test Length = 20, Pool Size = 500, Equal Weight Composite

Theta Composite	Directional Theta Max	Random	Directional Information	D-Optimality
DECILES	Maxium Item Exposure Rate	Max Item Exposure Rate	Max Item Exposure Rate	Max Item Exposure Rate
1 (LOWEST)	47%	11%	100%	100%
2	31%	11%	100%	100%
3	22%	12%	100%	100%
4	25%	9%	100%	99%
5	25%	10%	100%	99%
6	24%	10%	100%	100%
7	20%	10%	100%	100%
8	18%	11%	100%	100%
9	18%	10%	100%	100%
10 (HIGHEST)	29%	12%	100%	100%

Conclusions

STT used item selection procedures that

- (1) Had little bias in estimates of theta vector*
- (2) Were more accurate than random item selection*
- (3) Had different rates of item pool utility and max item exposure*
- (4) Had large differences in conditional between test overlap rates and max item exposure rates.*
- (5) Directional Theta Maximum appears to show the greatest promise of providing a balance between test accuracy and test security in a MIRT CAT context.*

Questions



Thank You!

Reference & Abstract



Bryant, D. U., & James, K. (2011, October). *Development of an Internet-based, multidimensional computer adaptive testing platform*. Paper presented at the 2011 International Association for Computer Adaptive Testing Conference, Pacific Grove, CA.

Multidimensional Item Response Theory (MIRT) is becoming a mature framework for the development, administration, and scoring of computer adaptive tests (CAT). Moreover, the Internet is becoming a more viable mode of administering test content for more and more organizations. Evidence suggests that these two trends will continue into the foreseeable future. Despite some theoretical advantages of MIRT over unidimensional IRT, there is no Internet-based, multidimensional computer adaptive testing platform in operation across the world. In addition, there is no evidence to suggest such a system has been created to date. Based upon the theoretical research on MIRT published in *Applied Psychological Measurement* and *Psychometrika* (Ackerman, 1996; Bryant, 2005; Bryant & Davis, 2011, Reckase, 1985), a multidimensional computer adaptive testing platform is developed. This presentation describes the general process used to create the first web-based, multidimensional computer adaptive platform. Results of research on the capabilities of the platform are presented. Implications are discussed.