A Brief History of (Response) Time

Marcus Scott, Caveon Test Security

Sarah Toton, Caveon Test Security

Dennis Maynes, Caveon Test Security

caveon™
Test Security

Utility of Response Time Data

Analysis of response time data is an important aspect of test security. These data can be used to detect:

- Examinees with pre-knowledge
- Compromised items
- Proxy test takers
- Item harvesters
- Other threats

Response Time Research

Response time research for testing purposes focuses on two main areas:

- Modeling response time data
- Using response time data to detect security threats

Over 30 years of research in these areas

Outline

- Approaches to modeling response time data
- Detection methods based on response times
  - Type of security threat detected
  - Computation
  - Strengths and weaknesses
- Future research

Approaches to Modeling Response Time Data

caveon™
Test Security

Nature of Response Time Data

- Response time data can be fit to known statistical distributions
- For a given item, the response times are:
  - Non-negative
  - Unimodal
  - Positively-skewed
- Several distributions meet these criteria

Gamma Distribution

- Used by Rasch (1960) to model the amount of time required for an examinee to read *N* words
- Used by Verhelst, Verstralen, and Jansen (1997) to model response times for tests with a time limit
  - Accounted for response time data when computing examinee ability
  - The gamma distribution was used so the ability distribution would have a logistic probability density function (pdf), which leads to a logistic item response function

Gamma PDF

- $f(t) = \left(\frac{t}{\beta}\right)^{\alpha-1} \frac{e^{(-t/\beta)}}{\beta\Gamma(\alpha)}$
- $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$
- Shape parameter α and scale parameter β can be estimated from sample statistics:
  - Let $\bar{t}$ be the mean response time and s_t be the sample standard deviation
  - $\alpha \approx \left(\frac{\bar{t}}{s_t}\right)^2$
  - $\beta \approx \frac{s_t^2}{\bar{t}}$

Weibull Distribution

- Used in system reliability theory
  - Models the time required for a system to fail
  - A test item is analogous to the system
  - The examinee's efforts to respond to the item are "attacks" on the system
  - Response to the item is analogous to failure of the system
- Used by Tatsuoka & Tatsuoka (1980) to build a scoring model that incorporates response time to account for teaching effect
- Roskam (1997) considered the whole test to be the system and used the Weibull distribution to predict how long it would take an examinee to finish the test

Weibull PDF

- $f(t) = \frac{\alpha t^{\alpha-1}}{\beta^{\alpha}} e^{-(t/\beta)^{\alpha}}, t > 0$
- The shape parameter α and scale parameter β are poorly estimated by sample statistics
- Can be estimated by a least squares fit to the cumulative distribution function (CDF)
- Magnitude of the shape parameter is related to the conditional probability that an examinee who has not responded as of time t will respond shortly after

Lognormal Distribution

- Used by Thissen (1983) to model a speed-accuracy relationship by regressing the logarithm of the response time on the IRT logit
- Used by Schnipke and Scrams (1997) to develop a mixture model for determining whether the response time resulted from “solution” behavior or “guessing” behavior

Lognormal PDF

- $f(t) = \frac{1}{t\sigma\sqrt{2\pi}} e^{\frac{-(\ln t - \mu)^2}{2\sigma^2}}$
- The sample mean and standard deviation of the natural logarithms of the response times give estimates of the parameters μ and σ

Which to Use?

- Schnipke and Scrams (1999) fit response time data to a:
  1. Normal distribution (used as a base case)
  2. Lognormal distribution
  3. Gamma distribution
  4. Weibull distribution
- Goodness of fit was compared to determine which distribution was best for modeling response time data

Schnipke & Scrams Experiment

- 30 items with response counts between 1,007 and 7,417
- For each item, used 500 responses (exploratory sample) to fit a model to the response times
- The same models were then applied to the remaining response times (confirmatory sample)
- Root mean square error at every 5^{th} percentile was used to evaluate fit

Schnipke & Scrams Results

	Exploratory RMSE			Confirmatory RMSE		
Distribution	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**
Lognormal	0.016	0.008	0.033	0.020	0.002	0.039
Gamma	0.038	0.020	0.067	0.039	0.019	0.072
Weibull	0.051	0.030	0.076	0.049	0.026	0.075
Normal	0.084	0.065	0.112	0.081	0.055	0.116

Lognormal distribution seems to be the best for modeling response times

Advancements to the Lognormal Model

- van der Linden & van Krimpen-Stoop (2003) parameterized the lognormal model as a loglinear model to quantify item complexity and test taker working rate
- Meijer & Sotaridona (2006) used this model to estimate an “effective response time,” which is the time required by an able examinee to answer correctly
- Van der Linden (2006) simplified the 2003 model

Loglinear Model (van der Linden & van Krimpen-Stoop)

- For item i and examinee j
  - $\ln(T_{i,j}) = \mu + \delta_i + \tau_j + \varepsilon_{i,j}$
  - μ is a general response time for all items and examinees
  - δ_i is the time required to respond to the item
  - τ_j is the examinee's slowness
  - $\varepsilon_{i,j}$ is a residual term; $\varepsilon_{i,j} \sim N(0, \sigma^2)$
- $\ln(T_{i,j}) \sim N(\mu + \delta_i + \tau_j, \sigma^2)$

caveon™ Test Security

Parameter Estimation

- With M items and N examinees, $\mu = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\ln\left(T_{i,j}\right)$
- Presence of the population average, μ, requires:
  - $\frac{1}{M}\sum_{i=1}^{M}\delta_i = 0$
  - $\frac{1}{N}\sum_{j=1}^{N}\tau_j = 0$
- Therefore,
  - $\delta_i = \left(\frac{1}{N}\sum_{j=1}^{N}\ln T_{i,j}\right) - \mu$
  - $\tau_j = \left(\frac{1}{M}\sum_{i=1}^{M}\ln T_{i,j}\right) - \mu$
- Bayesian parameter estimation may also be used

Why Use This Method?

- Parameterization gives more information about the items and examinees
  - For items, difficulty ≠ time intensity
  - For examinees, faster ≠ higher ability
- Parameters τ_j and δ_i are used in Bayesian analysis of the response times

Effective Response Time

- Effective Response Time (ERT) is the time used by an able examinee to answer an item correctly
- $\ln T_{i,j} = \beta_0 + \beta_1 \theta_j + \beta_2 \tau_j + \varepsilon_j$
  - θ_j is computed according to the IRT model chosen (Meijer & Sotaridona used 3PL)
  - τ_j is computed as was shown previously
  - $\varepsilon_j \sim N(0, \sigma_i^2)$
- $\widehat{\ln T_{i,j}} = \hat{\beta}_0 + \hat{\beta}_1 \theta_j + \hat{\beta}_2 \tau_j$

Response Time Restrictions

- Not all response times are used for the regression:
  - Only response times for correct responses are considered, and
  - $P(\text{correct}) > \gamma$
- These restrictions reduce variability in the response times:
  - Guessing by less-able examinees, whether correct or incorrect, leads to misleadingly short response times
  - Long response times by less-able examinees, regardless of a correct or incorrect response, are misleading

van der Linden Lognormal Model (2006)

- $f(t_{i,j}) = \frac{\alpha_i}{t_{i,j}\sqrt{2\pi}} e^{-\frac{1}{2}\left[\alpha_i\left(\ln t_{i,j} - (\beta_i - \tau_j)\right)\right]^2}$
- Treatment of response times that is similar to IRT
- $\alpha_i = \frac{1}{\sigma_i}$ is a time discrimination parameter for the item
- β_i is the time intensity for the item
- τ_j is the examinee's test-taking speed (not slowness)

Parameter Estimation

- The quantity $\beta_i - \tau_j$ is not identifiable, like in IRT
- Require $E(\tau_j) = 0$
- Bayesian procedures are used to estimate the parameters

Detection Methods Based on Response Times

caveon™
Test Security

Van der Linden & van Krimpen-Stoop Loglinear Model (2003)

- Detects pre-knowledge by identifying unexpectedly short response times with correct responses
- Detects memorization by identifying unexpectedly long response times with random or incorrect responses
- Probability of response time and correct response are computed and analyzed with classical and Bayesian checks
- Their method had very low power and a high type-I error rate; Bayesian methods had a higher type-I error rate
- Was not recommended for use

Detecting Pre-knowledge with ERT

- Pre-knowledge is assumed to manifest as short response times with correct responses
- For an examinee suspected of pre-knowledge, compute
  - $\mathrm{X} = \sum_{i=1}^{M} \left(\frac{\ln T_{i,j} - \widehat{\ln T_{i,j}}}{\sigma_i} \right)^2$
  - $\sigma_i^2 = (J_i - 1)^{-1} \sum_{j=1}^{J_i} \left(\ln T_{i,j} - \widehat{\ln T_{i,j}} \right)^2$
- The variable X is modeled by a chi-squared distribution with M degrees of freedom
- $p(\mathrm{X} \geq \chi)$ can be compared to a significance level α

Meijer & Sotaridona Experiment

- Data for 528 examinees who took a CAT test
- 100 were randomly selected and $p(\mathrm{X} \geq \chi)$was computed
  - Repeated 100 times
  - Compared to $\alpha = 0.01$ and 0.05
  - Average is estimate of type I error
- Power investigation
  1. Select random sample of examinees
  2. For ½ or ¾ of the items, change response time to ½ or ¼ of the value
  3. Repeat 1000 times

Meijer & Sotaridona Results

- Approach 1: compute ERT using the regression
- Approach 2: compute ERT using the average of $\ln T_{i,j}$
- Type I error rates:

Alpha	Approach 1	Approach 2
0.05	0.022	0.038
0.01	0.009	0.011

Meijer & Sotaridona Results

- Detection Rates:

		Approach 1		Approach 2	
Pre-knowledge Proportion	**Time Decrease**	**α = 0.05**	**α = 0.01**	**α = 0.05**	**α = 0.01**
0.50	50%	0.345	0.216	0.475	0.346
0.50	75%	0.938	0.878	0.944	0.895
0.75	50%	0.482	0.311	0.585	0.447
0.75	75%	0.976	0.949	0.985	0.965

caveon™ Test Security

Meijer & Sotaridona Conclusions

- Approach 1 is more conservative than Approach 2
- Approach 2 is more powerful than Approach 1
- Time decrease had a stronger effect on power than pre-knowledge proportion
- $\ln T_{i,j}$ was nearly constant across θ, so the average is a suitable estimate of ERT

Liu, Primoli, and Plackner (2013)

- Studied ERT with grade 4 mathematics state test with multiple choice (MC) and constructed response (CR) items
- Used wrong-to-right answer changes and item visit counts to cross-validate aberrant response times
- Used a 3PL model for items
- Pseudo-guessing parameter was used as γ

Liu, Primoli, and Plackner Findings

- MC items were more likely to have aberrant response times than CR items
- For both types, difficult items were more associated with aberrant response times
- High WTR rates were consistent with aberrant response times
- High visit counts were not consistent with aberrant response times

van der Linden & Guo (2008)

- Studied efficacy of van der Linden's lognormal model (2006)
- Used to detect pre-knowledge and item memorization
- Compute response time residuals: $e_{i,j} = \alpha_i\left[\ln t_{i,j} - \left(\beta_i - \tau_j\right)\right]$
- $e_{i,j} \sim N(0,1)$; flag as aberrant if $\left|e_{i,j}\right| > 1.96$

van der Linden & Guo Case Study

- 110,562 response times from GMAT data
  - 2,487 (2.25%) were longer than expected
  - 1,863 (1.69%) were shorter than expected
- Model tends to over-represent long response times and under-represent short response times
- Detection rate was close to the nominal rate, so a power study was conducted

van der Linden & Guo Power Study

- Simulated pre-knowledge of one item by setting the response time (δ) to 10, 20, or 30 seconds
- $m \in [2, 30]$ additional items had normal response times
- 800 replications for each combination of pre-knowledge time and m
- Method did not perform well with $m = 2$
- Detection rate generally decreased with δ

van der Linden & Guo Detection Rates

	α = 0.05			α = 0.01		
Regular Items	**δ = 10**	**δ = 20**	**δ = 30**	**δ = 10**	**δ = 20**	**δ = 30**
2	0.26	0.03	0.00	0.02	0.00	0.00
4	0.85	0.50	0.26	0.64	0.25	0.08
6	0.77	0.36	0.15	0.54	0.14	0.03
8	0.84	0.45	0.22	0.60	0.22	0.05
10	0.87	0.48	0.26	0.68	0.24	0.08
20	0.87	0.45	0.33	0.70	0.34	0.16
30	0.87	0.36	0.31	0.60	0.31	0.17

Qian, Staniewska, Reckase, & Woo (2016)

- Applied van der Linden lognormal model to real data
  - CBT data from 2010 (assumed uncompromised) and 2012 (possibly compromised)
  - CAT data from beginning and end of item pool operating time (3-month period)
- Detected some possibly compromised items and some examinees who potentially had pre-knowledge
- Simulation study showed 67% power for detecting examinees with pre-knowledge of < 10% of the items

Caveon Response Time Statistic

- Computes “Robust Time,” which is an average response time per item
  - Censor 15% of the response times from both tails to remove the effect of long or short times
  - Robust Time is the exponentiated average of log response times, which is equivalent to the geometric mean of the response times
- $T_R = e^{\frac{1}{N}\sum_{i=1}^{N} \ln T_{i,j}} = \left(e^{\sum \ln T_{i,j}}\right)^{\frac{1}{N}} = \left(\prod_{i=1}^{N} e^{\ln T_{i,j}}\right)^{\frac{1}{N}} = \sqrt[N]{\prod T_{i,j}}$
- Get a standardized value by comparing to expected Robust Time

caveon™ Test Security

Caveon Response Time Statistic

- Can detect examinees with very fast response times
- Clients have used this to invalidate tests
  - Very conservative threshold of 12 seconds
  - A sample of invalidated tests had Robust Times of 8.4 seconds or less
  - Session times for this sample ranged between 6:41 and 16:18 for 60-item test
- Invalidation is not accusation of cheating—the score is not trustworthy

caveon™
Test Security

Caveon Fast-Erratic Statistic

- Detects pre-knowledge by identifying examinees with
  1. Rapid response times
  2. Response times that do not correlate with those of the population
- Combination of 2 statistics:
  1. Working rate – Median standardized logarithm of the response time
  2. Kendall's tau – Non-parametric correlation between the examinee's response times and the population's response times, adjusted for correct or incorrect response
- Non-parametric statistics are used for robustness

caveon™ Test Security

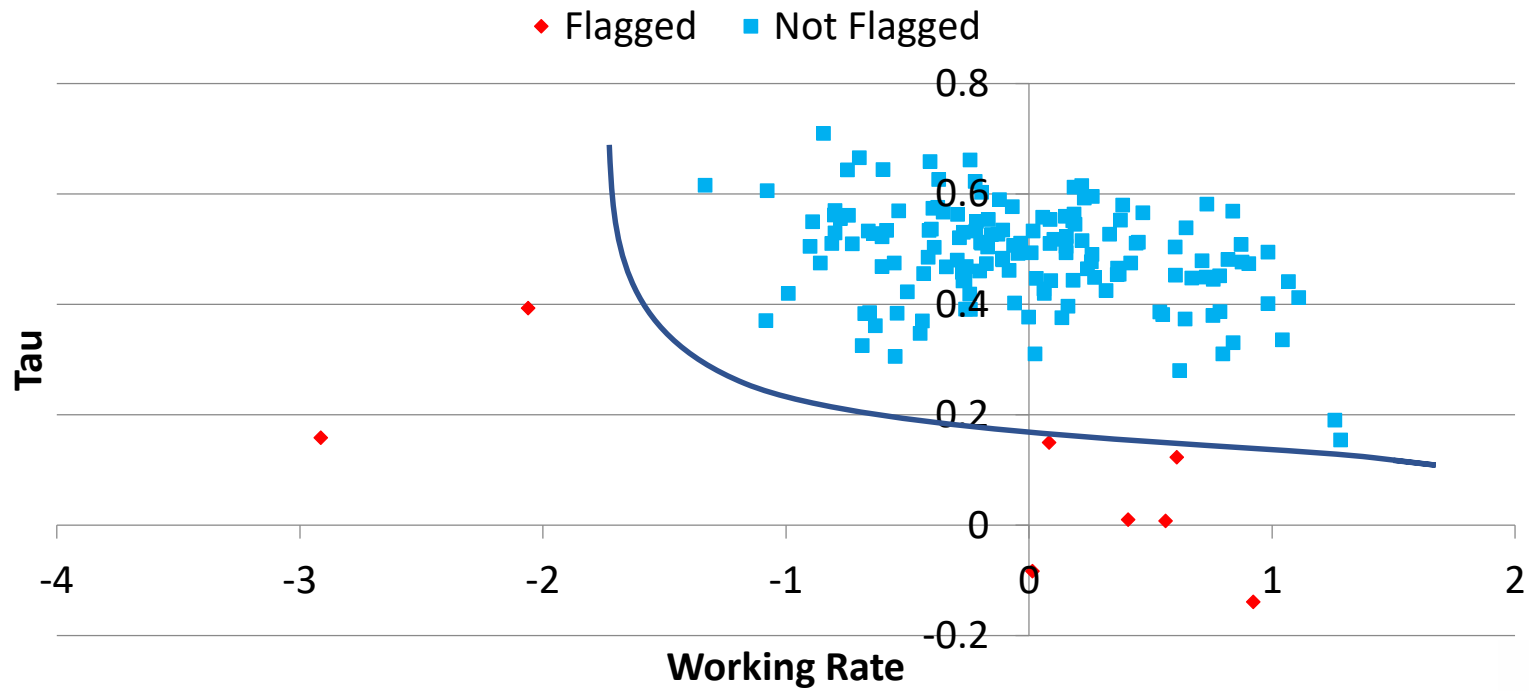
Caveon Fast-Erratic Statistic

- To combine statistics, assume:
  1. Data represent the entire population,
  2. Working rate is distributed normally,
  3. Kendall's tau is distributed normally, and
  4. Working rate and tau are independent
- The transformation $y = -2 \ln u$ gives a Chi-Square variate with 2 degrees of freedom
- Due to independence, the sum of transformed working rate and tau probabilities is Chi-Square with 4 degrees of freedom

caveon™
Test Security

Fast-Erratic Example

Curve marks the boundary of p = 0.001 for Chi-Square with 4 df

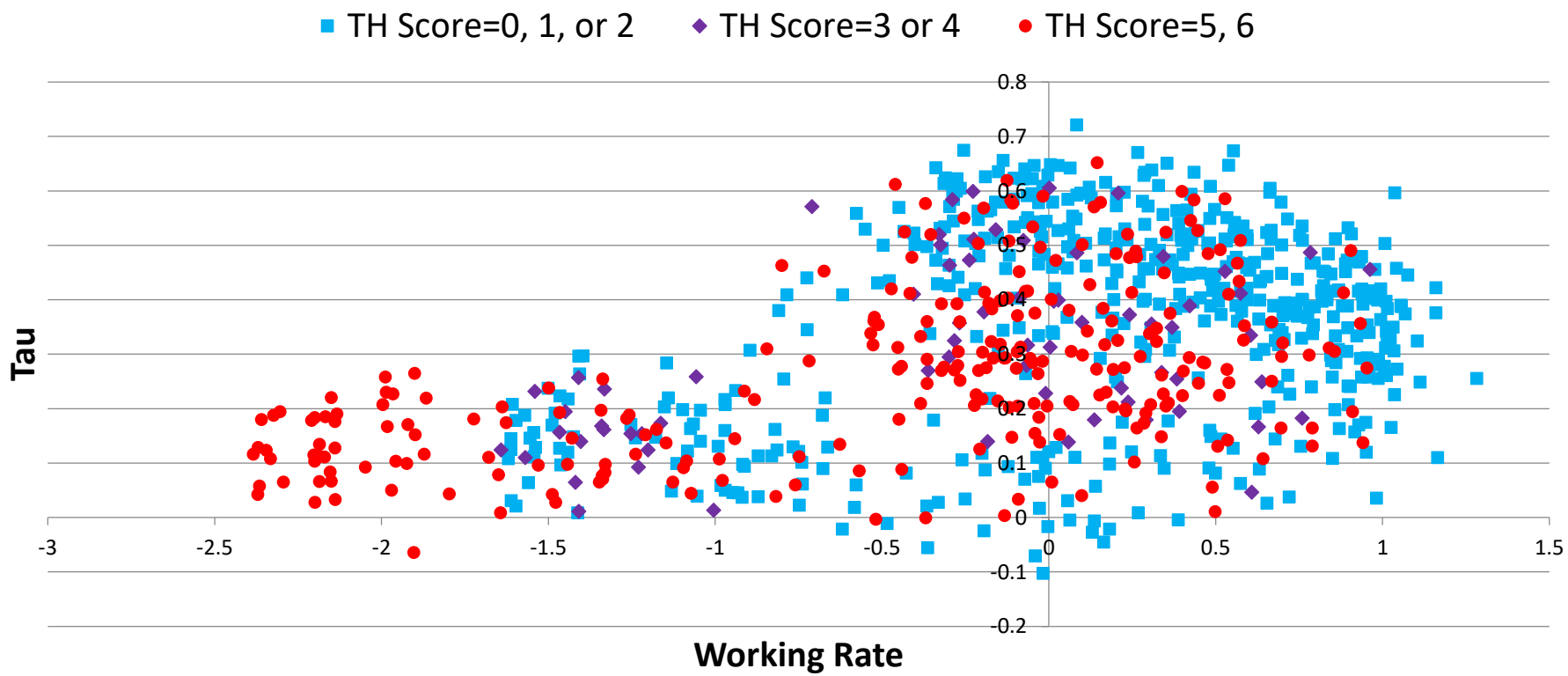


caveon™ Test Security

Fast-Erratic Validation

- 831 test instances for exam with 6 Trojan Horse items
- Tests with Trojan Horse scores of 5 or 6 were assumed to have pre-knowledge (265 out of 831, or 32%)
- Compute Fast-Erratic Statistic and compare with Trojan Horse score
- Chi-Square test found that Fast-Erratic detections are not random, but are associated with pre-knowledge

Fast-Erratic Validation



caveon™
Test Security

Conditional zRTs

- Computes a conditional scaling of response times
- Compares response times for each examinee on each item to a sample of examinees without pre-knowledge, based on their item score
- Exploratory clustering analyses on these values show very good separation between compromised and uncompromised items (96% accurate) and between examinees with and without pre-knowledge (97% accuracy) in experimental data

Potential Future Research

Analysis of Reading Rates

- Knowing item word counts will allow us to estimate how long it should take to read the item
- Adults typically read between 250-300 words per minute
  - For technical content, that decreases to 50-75 w.p.m.
  - Rates of 1,000-1,200 w.p.m. are speed-reading competitor level
- One testing program found reading rates of 60-70 w.p.m. were the fastest for SMEs (20 SMEs studied)

Analysis of Reading Rates

- Estimating reading rate can help determine whether the examinee likely read the question
- Word counts can help validate methods that compute item intensity parameters, ERTs, etc.
- These data would not be difficult to obtain

Additional ERT Research

- Robustness of ERT to selection of γ
  - Original Meijer & Sotaridona paper may have used $\gamma = 0.25$, but this is not explicitly stated
  - Liu, Primoli, & Plackner paper used the pseudo-guessing parameter from the 3PL
  - How much does choice of γ affect the ERT?
- Would be interesting to see how ERT is affected when a test taker speed parameter from a different model is used

Additional Fast-Erratic Research

- Caveon is developing a version that accounts for correct/incorrect when computing working rate, but it's computationally slow
- Would also like to develop a version that computes likelihood ratio of two hypotheses:
  1. Null hypothesis – Examinee has response times that change with item complexity
  2. Alternative hypothesis – Examinee has response times that are unaffected by item complexity (potential pre-knowledge)

caveon™
Test Security

Questions?

caveon™
Test Security

Thank You!

Marcus Scott

marcus.scott@caveon.com

Sarah Toton

sarah.toton@caveon.com

Dennis Maynes

dennis.maynes@caveon.com

caveon™
Test Security