RUNNING HEAD: Examinee Response Times in a CAT

Response Times for Correct and Incorrect Item Responses on Computerized Adaptive Tests

Shu-Ren Chang

Barbara S. Plake

Abdullah A. Ferdous

Department of Educational Psychology

University of Nebraska Lincoln

Paper presented at the 2005 annual meeting of

the American Educational Research Association (AERA), Montréal, Canada

To whom correspondence should be addressed: email changshuren@yahoo.com

Response Times for Correct and Incorrect Item Responses

on Computerized Adaptive Tests

Abstract

This study examined the time different ability level examinees spend taking a CAT on demanding items to these examinees. It was also found that high able examinees spend more time on the pretest items, which are not tailored to the examinees' ability level, than do lower ability examinees. Higher able examinees showed persistence with test questions, regardless of the item's difficulty level on operational and pretest questions they answer correctly or incorrectly. Results showed that more able students spend more time on all items, regardless of whether the items are answered correctly or incorrectly. These results were consistent for male and female examinees and for US and non-US administration sites.

Key words: Computerized Adaptive Testing, Response Time, Fairness

Response Times for Correct and Incorrect Item Responses

on Computerized Adaptive Tests

Introduction

With computerized adaptive testing, examinees are administered test questions that are matched to their performance level. Examinees who are performing well on the test questions (i.e., answering correctly) will receive a more difficult set of test questions than will less able examinees. Proficiency estimates are put on the same scale by use of Item Response Theory (IRT). Therefore, even though these examinees take tests of differing difficulty, their performances are made comparable through use of IRT scaling.

Although not necessarily part of the theory, in practice most operational computerized adaptive tests have a fixed amount of time for the examinees to complete the test and a fixed total number of operational items. Within the operational items, embedded pretest items are also routinely administered. Even though examinees are informed that pretest items will be administered (but not counted toward their final proficiency estimates), examinees do not know which items are operational and which are for pretest purposes. Therefore, the time that examinees spend answering the pretest items deducts from the time available for responding to the operational items. Typically, examinees are administered the same total number of pretest items and operational items. Unlike the operational items, these pretest items usually are not tailored to the ability level of the examinee.

Purpose statement

The purpose of this study is to examine the amount of time examinees spend on operational and pretest items as a function of their ability levels and background characteristics. Hornke (2000) found that examinees spent more time in a CAT on items they answered incorrectly than on items they answered correctly. However, their analyses did not consider examinee characteristics such as gender or USA citizenship. In addition, Hornke did not consider whether this result was dependent on the ability level of the examinees. Another purpose of the study is to examine the impact on the overall test time allocation due to embedded pretest items.

Research Questions

This study focused on three major research questions:

1. Do different ability level examinees spent equal amounts of time on items they answer correctly and incorrectly?
2. Do different ability level examinees, as a function of gender and USA citizenship, spend equal amounts of time on items they answer correctly and incorrectly?
3. Do examinees, regardless of ability level, spend equal amounts of time on pretest items and the other blocks of items that they answer correctly and incorrectly?

Data Source

Data were from an extant dataset provided to us by an international CAT admissions test administered worldwide in 2002. Over 10,000 examinee records are available. Two major components are assessed, verbal and quantitative, providing a total overall score. For each of the 28 operational items administered to the examinees, the data set contains examinee's item performances, response time per item, examinee's final ability estimates (on theta scale) and item difficulty estimates for the operational items administered to the examinee. For each component (verbal and quantitative) there are two parallel pools: Pool A and Pool B. For each examinee, items were drawn from one of these pools to comprise his or her operational test. In addition nine pretest items were embedded in the sequence of administered test items.

This study focuses only on the quantitative item pools. Future research will consider whether the results found for the quantitative item pools are similar to those found for the verbal section. In order to address research questions 1 and 2, examinees were grouped into 6 performance categories based on their overall proficiency estimation on the theta scale, from very low (e.g., <-2.0) to very high (>2.0). Examinee performance on the test questions was determined to be correct or incorrect. Average response times by ability groups were then computed and compared for correct and incorrect responses. This was done for the nine pretest items and for four sequential 7-item blocks of the 28 operational items. Because a CAT generally takes a few items to focus in on the examinee's ability level, the performance times for the early items may not be as relevant as the response times for items in the remaining blocks of operational items (number 8 – 28). Although the pretest items are embedded throughout the

administration, they are not tailored to examinee ability. Therefore for the purposes of this study, pretest items were grouped into a separate block consisting of 9 items.

These same analyses will be performed for male and female examinees and for USA and non-USA examinees to examine whether the time needed to respond to the items differs as a function of these background characteristics (Research Question 2).

Results

The results will be initially reported by research question and then synthesized for trends across the studies.

Research Question #1: Do different ability level examinees spend equal amounts of time on items they answer correctly and incorrectly?

This research question was analyzed across the 5 item blocks (9-item pretest items; items 1 – 7; items 8 – 14; items 15 – 21; items 22-28) by the six ability levels and item score (correct or incorrect) using a MANOVA analysis (See Figures 1-5). Using Wilks' Lambda as the significance test, a significant theta level by item score interaction was found (Wilks' Lambda = 0.007, $p < .001$). Follow up analyses revealed that there were both significant simple main effects for item score (Wilks' Lambda = 0.001, $p < .001$) for all the six-theta levels (Wilks' Lambda = 0.000, $p < .001$) for the correct and incorrect responses. Across item blocks, lower ability candidates systematically spent nearly equivalent time on items that they answered correctly and incorrectly. Higher ability candidates, in general, spent more time averaged across items they answered incorrectly than ones they answered correctly. Across the six ability levels, higher ability examinees spent significantly more average time on test questions than did their lower

ability counterparts, regardless of whether they answered the items correctly or incorrectly. Means and standard deviations of time spent on blocks of items by different ability examinees for items they answered correctly and incorrectly are displayed in Table 1 for quantitative pool A. Parallel results were found for quantitative pool B and are therefore not reported.

Therefore, for the full examinee group, in general, more able candidates spent more time on questions they answered incorrectly regardless of which pool was used (A or B). This is especially true for the most able candidate groups (theta >2.0). When we looked at the trends across the 7-item blocks, again most frequently less able students spent less time on average on operational items they answered correctly than they did on the items they answered incorrectly. In particular, less able students, who are receiving less difficult items than their more able counterparts, are spending less time on both the items they answer correctly and the ones they answer incorrectly. This is particularly noticeable in the middle two 7-item blocks, where the test is likely optimally functioning in tailoring the items to the examinees. Similar trends were found for the pretest items where the more able examinees spent on average a full minute longer on the items they answered incorrectly than did their less able counterparts. It is interesting to note that these pretest items are not tailored to the candidates' ability, so the fact that more able examinees are spending more time on these items is not necessarily related to the overall difficulty of these questions. Instead it may be that the more able candidates are more likely to persevere on test questions, whereas less able students may recognize the difficulty level of the question exceeds their knowledge and more quickly respond and move on to the next test question.

Research Question #2: Do different ability level examinees, as a function of gender and USA citizenship, spend equal amounts of time on items they answer correctly and incorrectly?

A similar MANOVA analysis was conducted for time devoted to answering the test question as a function of ability level (six theta levels) and item score (correct or incorrect) for male and female examinees and examinees who took the test in the USA or abroad (See Figures 6-7). MANOVA results were not significant for gender (Gender: Theta by gender interaction Wilks' Lambda = .363, p = .538) and were significant for administration locations either in USA or abroad (USA vs. Non-USA administration: Theta by location interaction Wilks' Lambda = .168, $p < .05$). The follow-up significant simple main effects were found (USA administration: Wilks' Lambda = .031, $p < .001$; Non-USA administration: Wilks' Lambda = .026, $p < .001$). Table 2 - 3 display means and standard deviations for response time for males and females whereas Table 4 - 5 display these results for US and non-US test takers. Results for males and females and US and non-US candidates mirror those reported above for the full candidate group.

Research Question #3: Do examinees, regardless of ability level, spend equal amounts of time on pretest items and other item blocks that they answer correctly and incorrectly?

One-way ANOVA test was used to address this question. Overall, across ability levels on the pretest items, candidates spent significantly more time on pretest items that they answered incorrectly than ones that they answered correctly ($Mean_{incorrect}$ = 128.61 seconds; $Mean_{correct}$ = 104.98) (See Figure 8). This difference was significant ($F_{(1,22)}$ =

11.619, p <.05) at alpha .05 level. This significant difference in time devoted to answering questions correctly and incorrectly was found consistently across the first and second operational item blocks as well ($Mean_{items\ 1\text{-}7,\ incorrect}$ = 138 vs. $Mean_{items\ 1\text{-}7,\ correct}$ = 111.70, $F_{(1,22)}$ = 28.38 , p <.001; $Mean_{items\ 8\ –14,\ incorrect}$ = 132.64 vs. $Mean_{items\ 8\text{-}14,\ correct}$ = 117.03, $F_{(1,22)}$ = 6.98 , p <.05). However, the difference for the third and fourth item blocks was non-significant ($Mean_{items\ 15\text{-}21,\ incorrect}$ = 117.18 vs. $Mean_{items\ 15\text{-}21\ correct}$ = 105.87, $F_{(1,22)}$ = 2.20 , p =.15; $Mean_{items\ 22\text{-}28,\ incorrect}$ = 90.62 vs. $Mean_{items\ 22\text{-}28,\ correct}$ = 88.81, $F_{(1,22)}$ = .10, p =.75).

Conclusion and Implications

This study examined the amount of time different ability level examinees spend while taking a fixed-length, time restricted CAT on questions they answer correctly or incorrectly. This was evaluated for sequential 7-block item sets, to mirror better the tailoring feature of a CAT for different ability candidates. The study considered quantitative components through the use of two parallel pools. In addition, the study compared the time spent of operational and pretest items as a function of item correctness. Results were also presented by examinee background characteristics, such as gender and USA or international status.

Overall, this study showed that more able students tend to spend more average time on items in the CAT, regardless of whether the questions are answered correctly or incorrectly, and whether the questions were operational or pretest items. Some people have expressed concerns that a time restricted, fixed-length CAT may be differentially speeded for high able candidates due to the item selection algorithm's delivery of

cognitively complex, more time demanding items to these examinees (Bridgeman and Cline, 2004). The results of this study are consistent with this concern. However, it was also found that these high able examinees spend more average time on the pretest items, which are not tailored to the examinees' ability level than do lower ability examinees. Therefore, this result may be more of an indication of higher able examinees' persistence with test questions, regardless of the item's difficulty level.

One of the features of a CAT is that the items are tailored to the ability level of the examinee, thereby providing a more precise assessment of their proficiency level. However, if some examinees are administered more time-demanding questions, while still within the same fixed time period, the fairness of the CAT to these students may be brought into question. The results of this study suggest that, for whatever reason, more able candidates are taking more time to complete their operational and pretest questions, resulting in more time pressure on these examinees to complete the test in the allotted time. However, even under the time constraints, on average, higher ability candidates are completing their tests within the 75-minute allocated time (Mean total time for higher ability candidates = 70.00 minutes). Therefore, even though these candidates use more time to complete their tests, it appears the time allocation, on average, is sufficient for them to complete the test. It is also interesting to note that the lower ability candidates, on average, complete the test in 55 minutes, 15 minutes sooner than their higher ability counterparts. It is not clear whether this time difference is related to the difficulty level of the items administered to the lower ability candidates or a lack of dedicated effort on the part of these candidates to answer test questions. On the other hand, the difference in total time spent on the test by the higher ability candidates may reflect the level of

difficulty and complexity of the questions administered through the item selection algorithm or simply by the level of dedicated effort by these examinees to test questions in general. Because these higher ability candidates tend to spend more time on all questions, those tailored to their ability level (operational items) and those that are not (pretest items), it could be argued that the increased time usage by these higher ability candidates could be due to their test taken strategy than simply due to the difficulty of the tasks administered to them through the item selection algorithm.

References

Bridgeman, B., & Cline, F. (2004). Effects of differentially time consuming tests on computer-adaptive test scores. *Journal of Educational Measurement, 41,* 137-148.

Hornke, L.F. (2000). Item response time in computerized adaptive testing. *Psychologia - Revista de Metodologia y Psycologia Experimental, 21*, 175-189.

Acknowledgements

We would like to express our appreciation to Lawrence Rudner, Chief Psychometrician of Graduate Management Admission Council (GMAC) for providing the data used in this study.

About the Authors:

Shu-Ren Chang is a PhD candidate majoring in Qualitative and Quantitative Methods in Education (QQME) in Department of Educational Psychology at the University of Nebraska-Lincoln, 21 Teachers College Hall, Lincoln, NE 68588; Email: changshuren@yahoo.com His research interests include applied educational measurement.

Barbara S. Plake, Ph.D., Univ. of Nebraska-Lincoln; W.C. Meierhenry Distinguished Univ. Professor, Director of the Oscar and Luella Buros Center for Testing, and Director of the Buros Institute of Mental Measurements at the University of Nebraska-Lincoln, 21 Teachers College Hall, Lincoln, NE 68588. Her expertise is primarily in the areas of computerized testing, including adaptive testing methods, and licensure/certification testing, including setting of performance standards or cutscores.

Ferdous Abdullah is a PhD candidate majoring in Qualitative and Quantitative Methods in Education (QQME) in Department of Educational Psychology at the University of Nebraska-Lincoln, 21 Teachers College Hall, Lincoln, NE 68588. His research interests include applied educational measurement.

Table 1. Comparison of response time when answering correctly and incorrectly by ability estimate group [Pool: Quant A]

	Ability Estimate Group	Sample Size (n)	Response Time									
			Pretest items 1-9		items 1-7		Items 8-14		Items 15-21		Items 22-28	
			Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
Correct Response	-3.0 to -2.0	145	98.31	50.64	115.12	64.43	99.97	55.00	96.24	52.35	81.97	58.97
	-2.0 to -1.0	450	108.17	50.80	114.64	53.91	111.38	53.24	103.29	48.60	84.46	50.42
	-1.0 to 0.0	1607	111.87	41.17	114.74	49.66	121.64	47.07	104.36	41.79	85.99	46.05
	0.0 to 1.0	2777	108.12	31.80	115.66	43.35	125.24	40.58	106.72	33.59	89.95	43.49
	1.0 to 2.0	2157	103.76	26.51	114.63	37.65	117.84	32.86	111.23	31.14	95.45	36.44
	2.0 to 3.0	805	101.31	23.98	113.80	33.08	111.75	26.25	113.05	29.53	104.71	33.38
	Total	7941	105.26	37.48	114.77	47.01	114.64	42.50	105.82	39.50	90.42	44.79
	Ability Estimate Group	Sample Size (n)	Response Time									
			Pretest items 1-9		items 1-7		Items 8-14		Items 15-21		Items 22-28	
			Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
Incorrect Response	-3.0 to -2.0	145	99.49	41.07	125.93	69.35	106.08	51.05	87.88	39.61	68.41	40.33
	-2.0 to -1.0	450	108.90	35.45	128.39	59.85	112.29	48.89	97.34	44.13	74.77	46.00
	-1.0 to 0.0	1607	119.63	38.26	136.85	63.11	130.90	57.25	105.57	45.82	84.43	51.18
	0.0 to 1.0	2777	135.03	52.91	142.17	66.66	136.90	62.91	117.18	53.38	90.04	56.16
	1.0 to 2.0	2157	149.89	68.68	156.32	82.13	148.53	72.95	138.36	72.38	100.71	66.22
	2.0 to 3.0	805	168.68	88.55	163.58	95.38	150.70	91.35	163.42	104.96	127.72	90.27
	Total	7941	130.27	54.15	142.21	72.75	130.90	64.07	118.29	60.05	91.01	58.36

Table 2. Comparison of response time when answering correctly and incorrectly by ability estimate group [Pool: Quant A- Male]

	Ability Estimate Group	Response Time									
		Pretest items 1-9		Items 1-7		Items 8-14		Items 15-21		Items 22-28	
		Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
Correct Response	-3.0 to -2.0	96.74	42.12	115.07	70.51	101.48	59.16	100.22	56.30	78.95	52.74
	-2.0 to -1.0	104.36	45.57	112.42	52.33	115.19	58.30	105.89	52.94	83.01	47.98
	-1.0 to 0.0	114.54	40.39	116.25	49.19	125.93	47.53	106.63	42.86	85.82	47.76
	0.0 to 1.0	109.17	31.41	114.82	42.43	126.43	40.67	106.71	33.92	90.21	42.89
	1.0 to 2.0	104.16	26.07	115.01	37.72	118.66	32.92	111.06	30.96	95.43	35.59
	2.0 to 3.0	102.14	23.72	115.09	33.80	111.77	25.94	113.54	29.66	103.59	31.14
	Total	105.19	34.88	114.77	47.66	116.58	44.09	107.34	41.11	89.50	43.02
	Ability Estimate Group	Response Time									
		Pretest items 1-9		Items 1-7		Items 8-14		Items 15-21		Items 22-28	
		Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
Incorrect Response	-3.0 to -2.0	98.24	41.92	120.63	66.82	108.21	46.25	90.54	38.56	68.83	38.77
	-2.0 to -1.0	113.42	35.56	125.66	57.83	116.79	52.47	96.95	46.73	75.29	47.07
	-1.0 to 0.0	121.44	38.49	137.66	62.88	132.37	57.71	106.59	46.43	82.95	51.25
	0.0 to 1.0	136.21	55.33	142.52	62.94	137.40	62.09	118.07	51.46	89.88	54.22
	1.0 to 2.0	149.75	65.02	154.29	80.39	148.54	72.77	138.31	73.08	101.44	65.96
	2.0 to 3.0	166.38	86.29	162.56	97.43	155.69	97.28	162.49	104.64	123.22	85.18
	Total	130.91	53.77	140.56	71.38	133.17	64.76	118.82	60.15	90.27	57.07

Table 3. Comparison of response time when answering correctly and incorrectly by ability estimate group [Pool: Quant A-Female]

	Ability Estimate Group	Response Time									
		Pretest items 1-9		Items 1-7		Items 8-14		Items 15-21		Items 22-28	
		Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
Correct Response	-3.0 to -2.0	97.60	54.61	110.41	56.78	96.59	50.64	91.53	46.74	85.37	64.90
	-2.0 to -1.0	111.20	54.77	116.53	55.30	109.04	48.82	101.35	45.68	85.45	52.76
	-1.0 to 0.0	108.03	41.00	112.87	50.58	116.51	45.84	102.23	40.71	86.57	43.84
	0.0 to 1.0	105.69	32.14	116.57	44.83	122.98	40.78	106.62	33.01	89.56	44.18
	1.0 to 2.0	101.58	27.26	112.26	37.70	115.84	33.50	112.82	31.58	96.59	39.08
	2.0 to 3.0	99.14	24.37	109.53	29.26	111.91	26.37	113.56	29.67	107.03	38.98
	Total	103.87	39.03	113.03	45.74	112.15	40.99	104.69	37.90	91.76	47.29
	Ability Estimate Group	Response Time									
		Pretest items 1-9		Items 1-7		Items 8-14		Items 15-21		Items 22-28	
		Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
Incorrect Response	-3.0 to -2.0	97.85	39.32	127.62	72.30	104.44	54.97	85.35	40.88	68.45	42.51
	-2.0 to -1.0	105.35	35.09	131.36	61.89	109.14	46.01	97.93	42.45	74.52	45.82
	-1.0 to 0.0	117.27	37.45	135.13	62.58	128.53	56.26	104.49	44.85	86.26	50.50
	0.0 to 1.0	132.15	48.70	139.01	71.75	134.92	64.27	116.35	57.21	91.19	59.26
	1.0 to 2.0	149.22	76.04	158.52	84.81	146.99	75.67	135.53	67.63	100.04	67.08
	2.0 to 3.0	173.85	93.71	171.95	92.98	138.41	72.99	171.98	110.26	142.43	100.49
	Total	129.28	55.05	143.93	74.39	127.07	61.69	118.61	60.55	93.81	60.94

Table 4. Comparison of response time when answering correctly and incorrectly by ability estimate group [Pool: Quant A- USA]

	Ability Estimate Group	Response Time									
		Pretest items 1-9		Items 1-7		Items 8-14		Items 15-21		Items 22-28	
		Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
Correct Response	-3.0 to -2.0	98.10	50.91	114.91	65.52	97.73	53.85	93.00	50.28	81.20	60.24
	-2.0 to -1.0	109.23	51.48	113.58	53.86	111.08	52.03	104.73	48.67	84.53	50.39
	-1.0 to 0.0	110.76	40.84	113.94	50.16	120.67	45.51	104.73	41.76	85.91	45.24
	0.0 to 1.0	107.86	32.09	115.25	43.68	125.06	41.10	106.29	33.93	90.67	42.66
	1.0 to 2.0	103.44	26.41	114.42	37.03	119.47	33.75	111.97	32.01	96.38	35.77
	2.0 to 3.0	98.79	24.43	115.77	33.61	111.73	27.83	112.48	29.87	104.02	32.97
	Total	104.69	37.70	114.65	47.31	114.29	42.35	105.53	39.42	90.45	44.54
	Ability Estimate Group	Response Time									
		Pretest items 1-9		Items 1-7		Items 8-14		Items 15-21		Items 22-28	
		Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
Incorrect Response	-3.0 to -2.0	99.62	39.95	128.07	71.34	106.78	51.03	86.76	39.52	67.30	39.75
	-2.0 to -1.0	109.28	35.62	127.47	59.12	111.61	49.28	98.08	44.57	74.27	44.12
	-1.0 to 0.0	118.62	37.23	134.38	61.40	129.85	56.42	106.18	45.96	85.41	51.98
	0.0 to 1.0	133.48	53.18	140.73	66.05	135.05	63.77	116.92	54.37	90.92	56.21
	1.0 to 2.0	147.12	65.15	153.25	79.50	145.16	71.72	133.41	68.88	101.92	66.82
	2.0 to 3.0	168.29	91.42	157.66	89.18	154.03	98.64	171.03	113.61	123.55	89.16
	Total	129.40	53.76	140.26	71.10	130.41	65.14	118.73	61.15	90.56	58.01

Table 5. Comparison of response time when answering correctly and incorrectly by ability estimate group [Pool: Quant A-International]

	Ability Estimate Group	Response Time									
		Pretest items 1-9		Items 1-7		Items 8-14		Items 15-21		Items 22-28	
		Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
Correct Response	-3.0 to -2.0	97.58	50.70	116.96	62.00	111.78	61.36	107.31	59.48	86.61	55.01
	-2.0 to -1.0	101.54	45.98	124.09	53.82	114.70	62.04	93.35	48.20	83.73	52.13
	-1.0 to 0.0	118.31	42.89	119.54	46.68	127.26	55.29	102.42	42.33	86.31	50.68
	0.0 to 1.0	109.27	30.99	117.60	42.35	125.87	38.85	107.91	32.20	87.51	45.90
	1.0 to 2.0	103.97	26.63	114.83	38.54	116.18	31.55	110.46	30.29	94.33	37.06
	2.0 to 3.0	102.65	23.59	112.66	33.16	111.80	25.33	113.39	29.47	105.31	33.71
	Total	105.55	36.80	117.61	46.09	117.93	45.74	105.81	40.33	90.63	45.75
	Ability Estimate Group	Response Time									
		Pretest items 1-9		Items 1-7		Items 8-14		Items 15-21		Items 22-28	
		Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
Incorrect Response	-3.0 to -2.0	99.72	46.95	119.30	60.52	103.53	52.87	93.03	41.09	72.77	43.91
	-2.0 to -1.0	107.15	34.74	135.67	65.30	118.51	46.05	92.26	41.62	76.61	58.11
	-1.0 to 0.0	125.44	43.58	151.02	71.13	137.03	61.46	101.54	44.89	78.61	46.43
	0.0 to 1.0	140.15	51.93	147.05	68.69	143.03	59.98	117.99	50.37	86.93	55.92
	1.0 to 2.0	153.43	72.62	160.29	85.76	152.53	74.79	144.57	76.14	99.64	65.47
	2.0 to 3.0	169.83	87.78	168.37	100.08	147.88	87.65	159.77	98.44	130.35	90.13
	Total	132.62	56.27	146.95	75.25	133.75	63.80	118.19	58.76	90.82	60.00

Figure 1: Averaged Response Time for Pretest Items by Candidates' Six Ability Levels

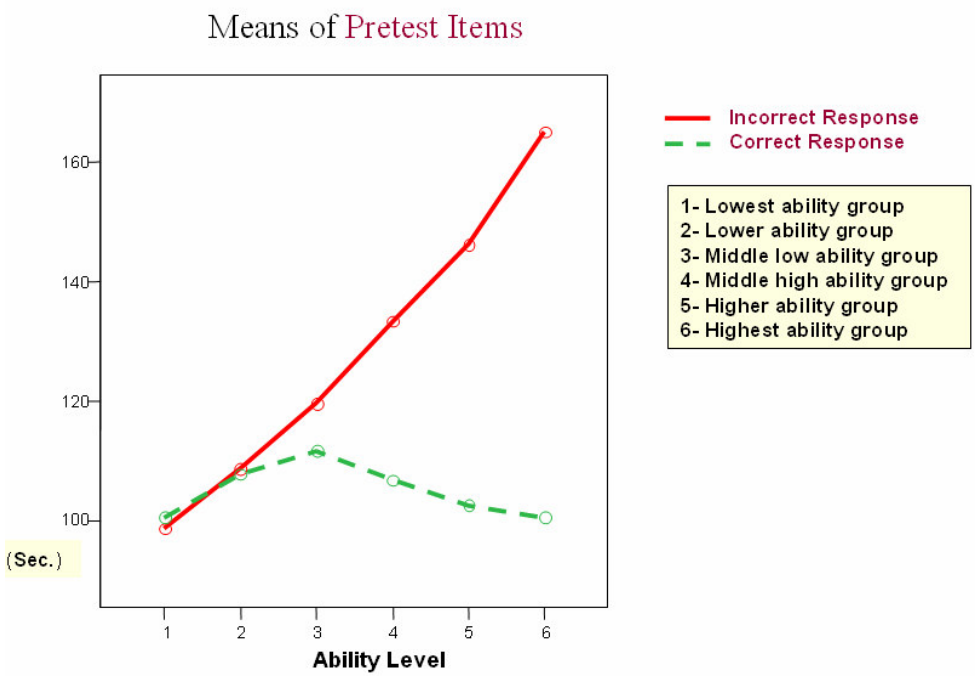

Figure 2: Averaged Response Time for Items 1-7 by Candidates' Six Ability Levels

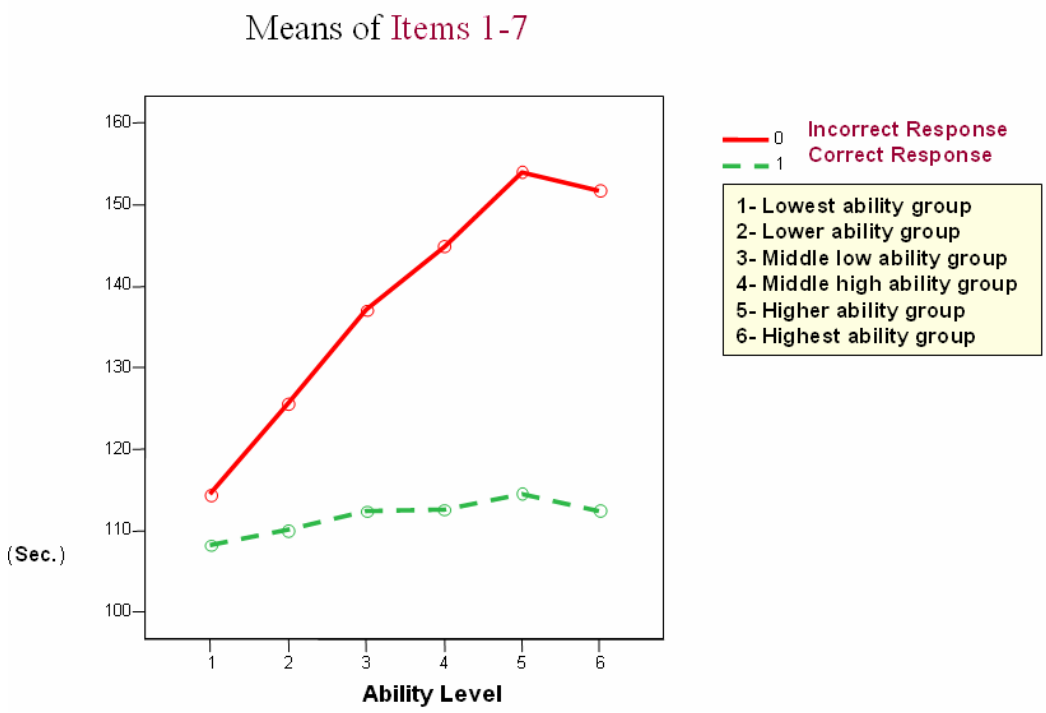

Figure 3: Averaged Response Time for Items 8-14 by Candidates' Six Ability Levels

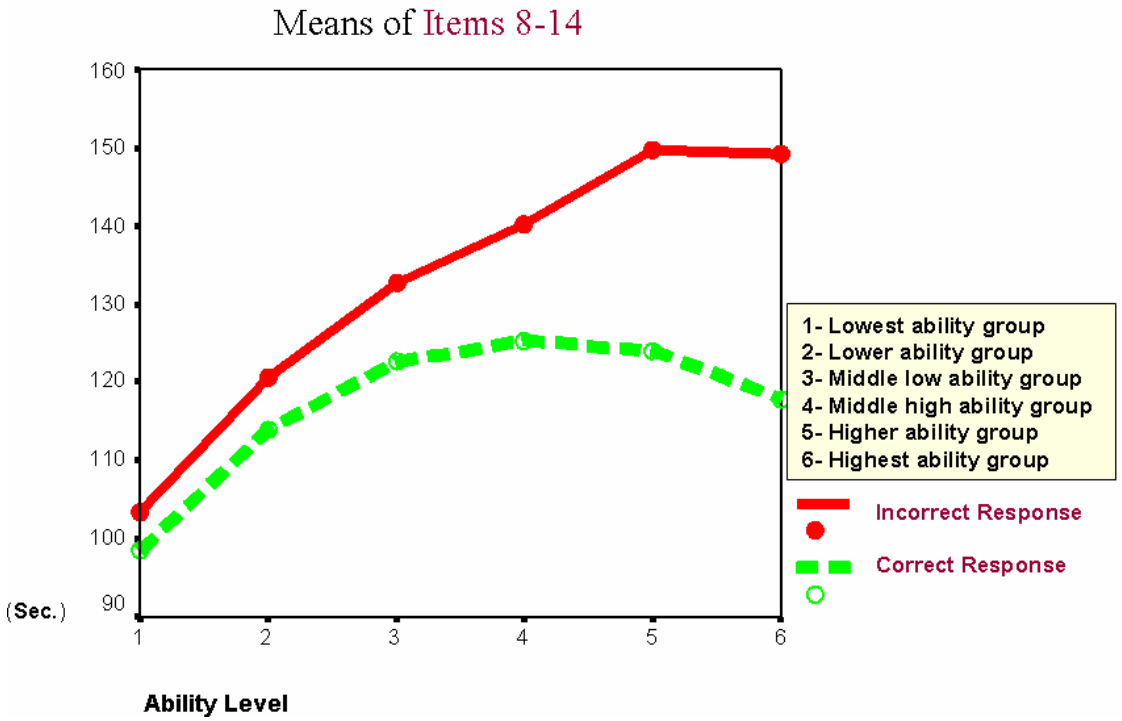

Figure 4: Averaged Response Time for Items 15-21 by Candidates' Six Ability Levels

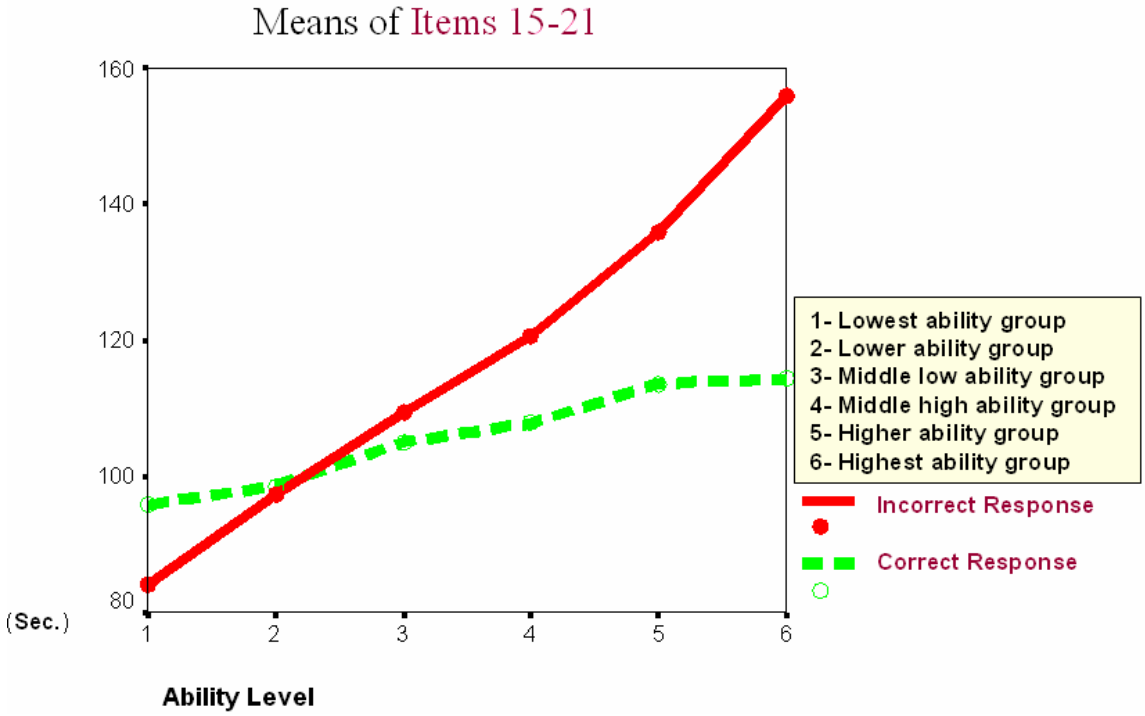

Figure 5: Averaged Response Time for Items 22-28 by Candidates' Six Ability Levels

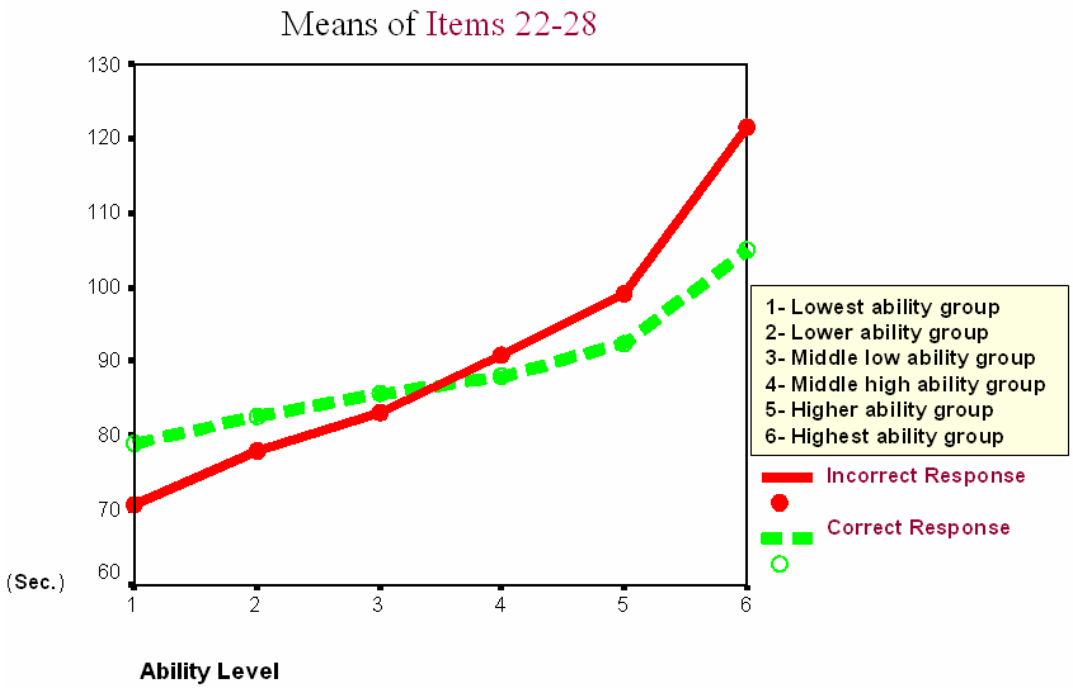

Figure 6: Averaged Response Time for Pretest and Operational Items by Six Ability Levels for Male/Female

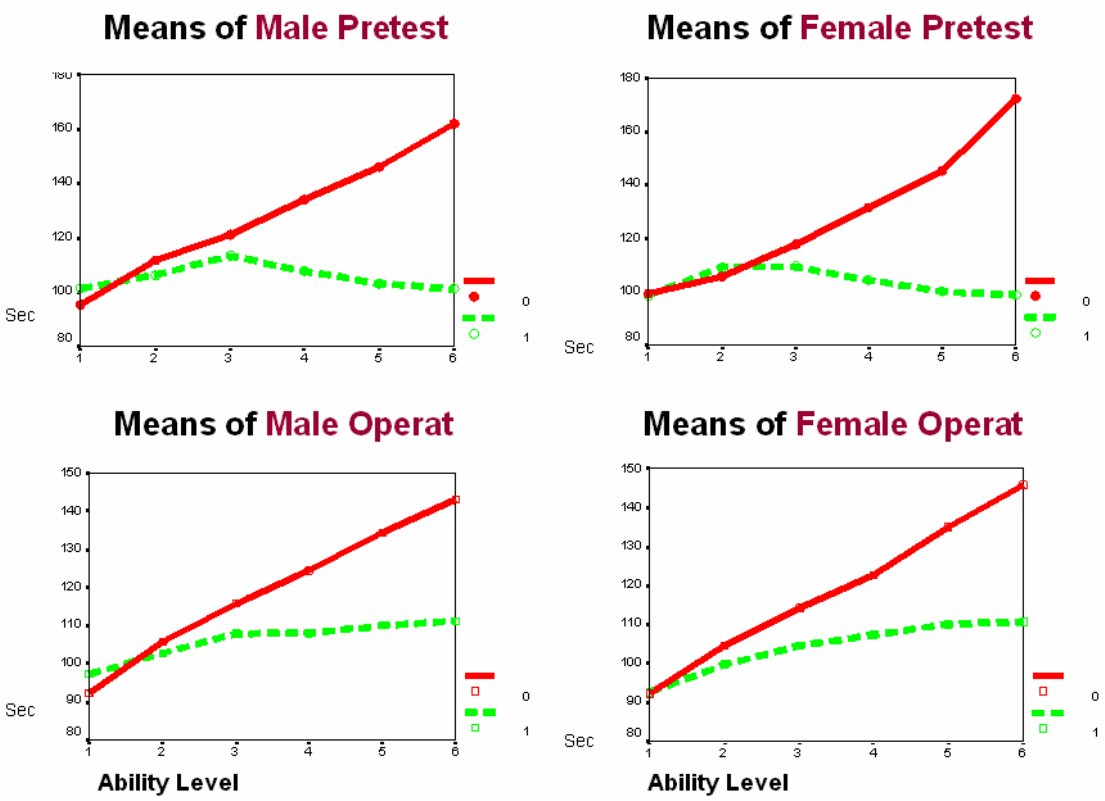

Figure 7: Averaged Response Time for Pretest and Operational Items by Six Ability Levels for USA/Non-USA

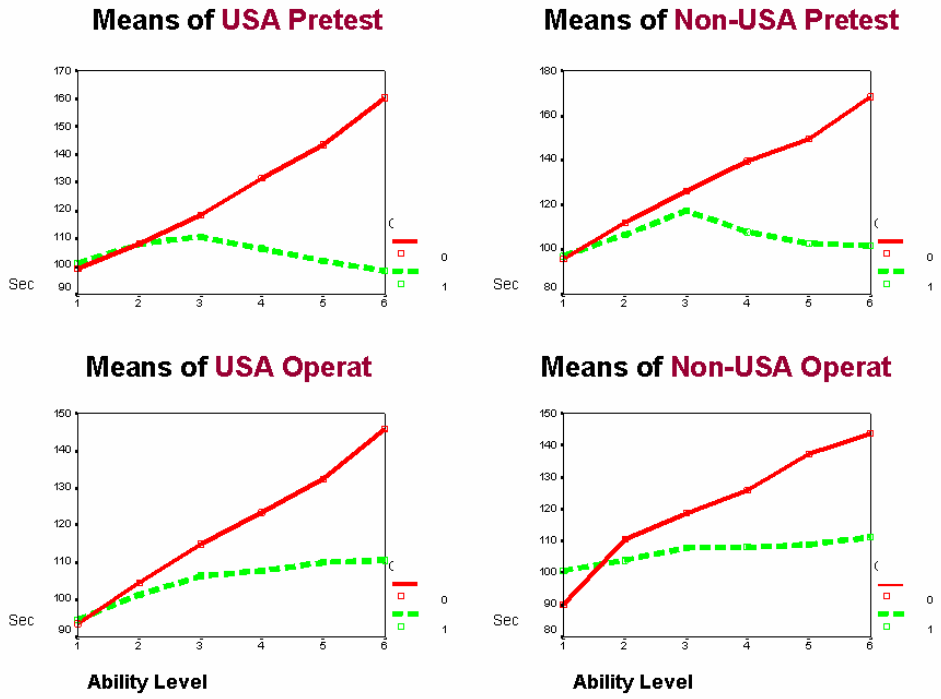

Figure 8: Averaged Response Time for Correct and Incorrect Responses by Five Item Blocks