Differential Speededness in the Context of Randomized Item Positioning

Brad Ching-Chao Wu
brad.wu@pearson.com
Pearson VUE
Keyword: differential speededness, test timing and fairness

Abstract

This study examines differential speededness when items in a test are presented to examinees in a random order. In computerized testing, item order is often randomized for the purpose of security and minimization of positioning effect on item parameter estimation. However, randomized item positioning might lead to undesirable differential speededness. When a test starts with time-intense items, examinees are likely to spend more time on them and therefore less likely to reach items at the end. Examinees under such condition could be unfairly disadvantage compared to those who start with less time-intense items. Empirical data from a large scale computerized test was analyzed to investigate the degree of differential speededness. Evaluation methods included proportion of items unreached conditioned on ability and average conditional differences in not reaching (Lawrence, 1993). The results indicated that in a completely randomized positioning design, differential speededness could be a potential threat to test fairness. The degree of differential speededness was especially prominent for lower ability candidates. Possible remedies to the undesirable differential speededness were briefly discussed.

Background and Introduction

Test speededness has been extensively investigated in the literature, especially with regards to its definition and threat to test validity. A less treaded path is the fairness aspect of test speededness. This area of reach has a fundamental hypothesis that speededness is not a universal phenomenon, or at least does not manifest to the same degree, within a test population. Researches in this area attempt to examine diverse patterns of speededness across various groups of examinees with the belief that some groups are more vulnerable to test speededness. The majority of differential speededness research so far has focused on the comparisons between gender, ethnics and language groups. Findings from these researches support the notion that degrees of speededness vary across subgroups. This study shares the same hypothesis of differential speededness but varies in group classification. The interest of this study is not in the natural differences between examinees, but the inevitable variation caused by the test design.

In computer-based testing, item positions are often randomized for security and validity purposes. While random-positioning has the advantage of preventing content memorization and eliminating the bias in parameter estimation caused by fatigue, it could also pose a threat to test fairness. For example, it is not unreasonable to hypothesize that examinees who receive the most time-intense items up front would spend more time on those items, therefore are less likely to finish the test. Likewise, examinees getting the least time-intense items at the beginning could be in a much more advantageous position. This downside of the randomized item positioning is worth investigating because it could affect the validity and comparability of the test score. The purpose of this study is to evaluate differential patterns of speededness as a result of randomized item positioning.

Methods

Instruments: Data of a large-scale college entrance examination were examined and analyzed. The examination is designed to measure candidates' general aptitude. It contains four sections: verbal reasoning, quantitative reasoning, abstract reasoning and decision analysis. All sections of the examination are timed and administered through computers. Previous analysis of the response time indicated that verbal reasoning was the least speeded and quantitative reasoning was the most speeded. That is, the majority of candidates (98%) were able to respond to all items in the verbal reasoning section within the allotted timeframe without random guessing. On the other hand, many candidates (64%) were unable to respond to all items in the quantitative reasoning section or had engaged in random guessing for additional credits. The other two sections (abstract reasoning and decision analysis) showed moderate degree of speededness and were not considered in this study. Data of the verbal section was analyzed to show the level of differential speededness when a test itself is not speeded, while the quantitative section was analyzed to show the level of differential speededness when a test is highly speeded. Three forms were created for both verbal and quantitative sections. Each verbal form comprised 11 testlets with 4 multiple-choice items following each testlet. Each quantitative form comprised 10 testlets with 4 multiple-choice items following each testlet. Twenty-one minutes were allowed for each section. Data from a total of 22,187 candidates were collected and analyzed.

Defining Focal and Reference Groups: Prior to data analysis, focal and reference groups were defined based on time intensity of the first two testlets. There are two focal groups and one reference group defined in this study. The focal groups refer to candidates who received either the most or the least time-intense testlets at the beginning of the test. Time intensity of a testlet was determined by the average response time (in seconds) on the four items attached to the testlet. Appendices 1 and 2 show the most and the least time-intense testlets for all forms in verbal and quantitative sections. The time intensity data clearly show the variation of the average item response time among the testlets within each form. In addition, quantitative testlets are generally more time-intense than verbal testlets, which explains the significant speededness observation in the quantitative section. Candidates who received two of the three most time-intense testlets at the beginning of the section were defined as the high-time-intensity focal group. On the other hand, candidates who received two of the three least time-intense testlets at the beginning of the section were defined as the low-time-intensity focal group. The rest of the candidates were categorized as the reference group.

Measures of Speededness: In this study, the number and proportion of unreached items were used as measures of speededness (Dorans, 1988; Lawrence, 1993). The measures were examined conditioned on candidate's ability, namely the total raw score. The average number of unreached items across the entire raw score range were calculated and plotted separately for the two focal groups and the reference group. Additionally, an index of Average Conditional Differences in Not Reaching (ACDNR) was generated for each comparison between the focal and the reference groups. The mathematical definition of ACDNR is shown in Appendix 3. This index is in the metric of the number of items in the test, thus the index indicates the average difference in the number of items not reached by the focal group relative to the reference group after matching the groups on the raw score (Lawrence, 1993).

Raw Scores Equating: A major concern of using un-equated raw scores as estimates of candidates' abilities is that raw scores themselves could possibly be affected by randomized item positioning. For example, a candidate's low total score could be a result of ability and the fact that he/she received the high-time-intense testlets at the beginning of the section. To eliminate the variance due to item positioning and get a more precise estimate of candidate's ability,

equipercentile equating (Kolen & Brennan, 2004) between the focal and the reference groups were performed. Specifically, all raw scores for the high-time-intensity focal group were equated to the reference group based on percentiles and the converted scores were used as estimates of candidates' ability. The same procedure was implemented between the low-time-intensity focal group and the reference group.

Results

Number of Unreached Items Conditioned on Ability: The plots in Appendices 4-6 illustrate the average number of unreached items across all raw score points on the verbal reasoning forms for the two focal groups and the reference group. The plots in Appendices 7-9 illustrate the same results on the three quantitative forms. Intuitively, the number of unreached items decreases as candidate ability increases. Focusing on the low-ability candidates where unreached items were mostly observed, one can see that there is virtually no difference in the number of unreached items among the three groups on the verbal reasoning test forms. However, differences in the number of unreached items were more prominent on the quantitative test forms, where significant speededness was observed. Looking closely in the low raw score range of the quantitative test, the average numbers of unreached items were higher for the high-time-intensity group than for the reference group. On the other hand, the average numbers of unreached items were lower for the low-time-intensity group than for the reference group. The high-time-intensity group was clearly disadvantaged compared to the other candidates.

ACDNR observation: Appendices 10 and 11 show the ACDNR between the focal group and the reference group on the verbal and quantitative reasoning tests respectively. The results are consistent with the number-of-unreached-items plots. The ACDNRs were low on the verbal test forms, which indicate that the differences in the average number of unreached items were negligible. The ACDNRs were larger on the quantitative test. The positive ACDNRs for the high-time-intensity group on the quantitative test indicate that candidates in that group missed 1.32 to 1.87 more items on average compared to the reference group. Whereas the negative ACDNRs for the low-time-intensity group show that candidates, on average, had 1.18 to 1.69 less unreached items when compared to the reference group.

Discussion

Results from this study support the notion that speededness could be a fairness concern in the context of randomized item positioning. Test takers who are assigned time-intense items at the beginning are more likely to be discouraged and frustrated by such design, therefore are unfairly disadvantaged and underestimated in terms of ability. Such fairness concern would naturally disappear when all or most of the test takers have sufficient time to respond to all items within the allotted timeframe. It is also reasonable to hypothesize that the larger the variation in time-intensity among testlet/items, the more disadvantaged a test taker is when he/she is presented with the most time-intense testlets/items at the beginning of the test. Under an unlikely condition where all testlets/items are similar in terms of time-intensity, differential speededness should not be observed.

Differential speededness does not exist when item positions are fixed for all test takers. Unfortunately, fixing item positions is often not recommendable due to the security and validity concerns (e.g, fatigue effect). The quick and perhaps the best fix to differential speededness under randomized item positioning is to eliminate speededness so all test takers have enough time to

respond and provide their best answers to all items. However, speededness is usually inevitable or even intended in real test settings. One possible remedy of differential speededness when speededness is inevitable or intended is to partially randomize item positions. That is, items can be ordered in a fashion where all test takers receive the same items (preferably with medium time-intensity) for the first 20-30% of the test and are assigned items randomly afterwards. The other option is to stratify the testlets/items by time-intensity and randomly assign testlets/items of the same time-intensity in each position to all test takers.

A point worth mentioning is that speededness can be and have been defined differently in literature. This study follows the classical definition of speededness through the examination of unreached items and average response time. While the classical definition of speededness suffices to serve as a general indicator of time intensity and speed behavior, it also assumes validity of all responses and ignores random guessing behavior. Test takers attempt random guessing at the end of a test in order to gain additional credits by chance are considered un-speeded under the classical definition because they did not miss an item. This could lead to bias in the evaluation of speededdness because random guessing is theoretically and practically invalid and should be regarded as unreached. Definition and identification of random guessing behavior call for a different method that involves modeling of item response time (Schnipke & Scrams, 2006; van der Linden, 2006). Investigation of differential speededness through response-time modeling may provide additional insight to the fairness issue of randomized item positioning.

Reference

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. Journal of Educational Measurement, 29, 309-319.
Kolen, M. J., Brennan, R. L. (2004). Test Equating, Scaling, and Linking. Springer Science, New York, NY, USA.
Lawrence, I. M. (1993). The effect of test speededness on subgroup performance. (ETS-RR-93-49). Princeton, NJ: Educational Testing Service.
Schmitt, A. P., Dorans, N. J., Crone, C. R., & Maneckshana, B. T. (1991). Differential speededness and item omit patterns on the SAT. (ETS-RR-91-50). Princeton, NJ: Educational Testing Service.
Schnipke, D. L., Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. Potenza, J. J. Fremer, & W. Ward (Eds.), Computer-based testing: Building the foundation for future assessments (pp. 237-266). Hillsdale, NJ: Lawrence Erlbaum Associates.
van der Linden, W. J. (2006). A lognormal model for response times on test items. Journal of Educational and Behavioral Statistics, 31, 181-204.

Appendix 1. Average Item Response Time for the Most and the Least Time-Intense Testlets in the Verbal Section

	Form 1	Form 2	Form 3
Most Time-intense Testlets	34.61	33.31	33.31
	31.34	31.97	31.97
	28.17	27.43	27.43
Least Time-intense Testlets	22.46	21.07	21.07
	20.08	19.38	19.38
	17.51	17.17	17.17

Appendix 2. Average Item Response Time for the Most and the Least Time-Intense Testlets in the Quantitative Section

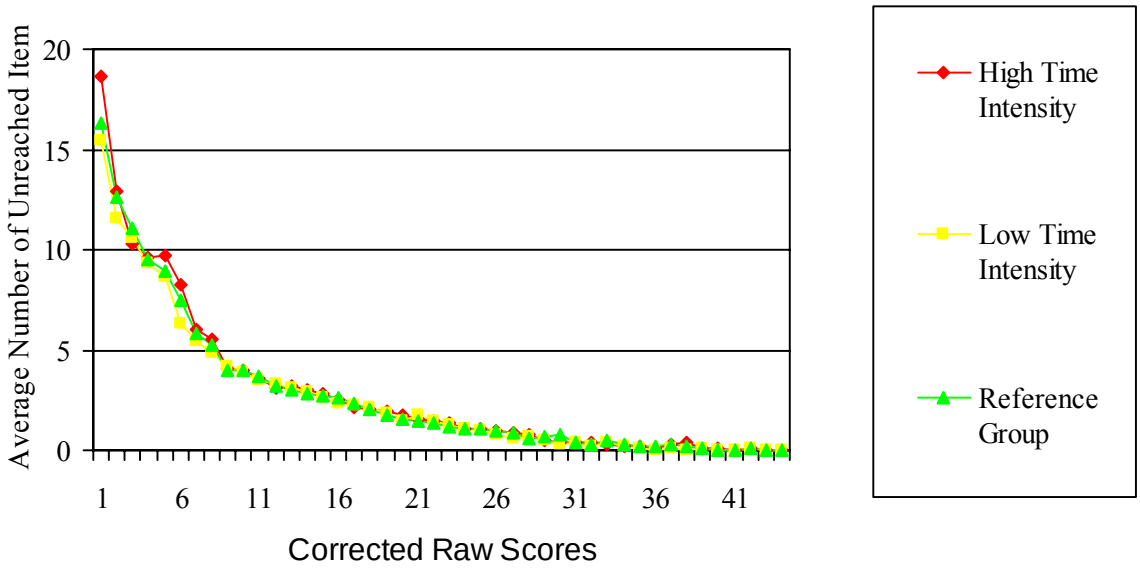
	Form 1	Form 2	Form 3
Most Time-intense Testlets	48.59	51.37	44.58
	46.38	45.47	43.19
	41.68	42.19	40.97
Least Time-intense Testlets	26.34	27.16	29.17
	22.84	25.46	24.37
	19.73	22.15	21.93

Appendix 3. Definition of Average Conditional Differences in Not Reaching (Lawrence, 1993)

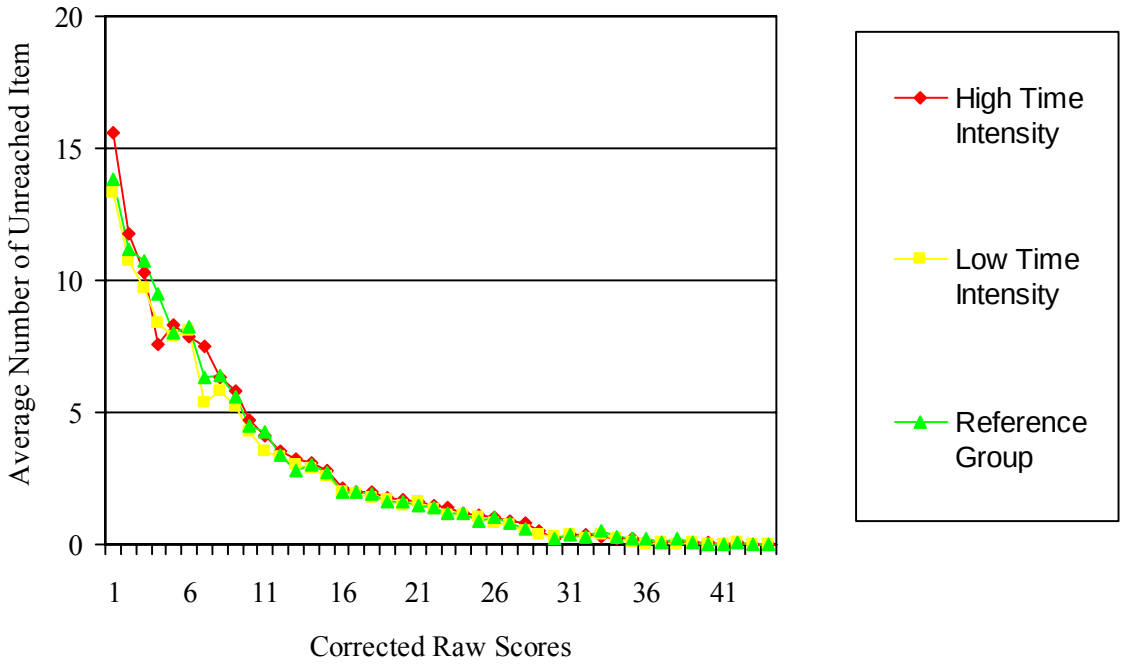
- Differences in Not Reaching Conditioned on Score M (CDNRM)
  $CDNRM=\Sigma i(PNRfim-PNRrim)*NI$
- Average Conditional Differences in Not Reaching (ACDNR)
  $ACDNR=\Sigma[(Nm/\Sigma Nm)*CDNRM]$

where
PNRfim and PNRrim = proportions not reaching item "i" in the focal and reference groups at the score level "m",
NI is the number of items in the test, and
Nm is the number of candidates at raw score level "m".

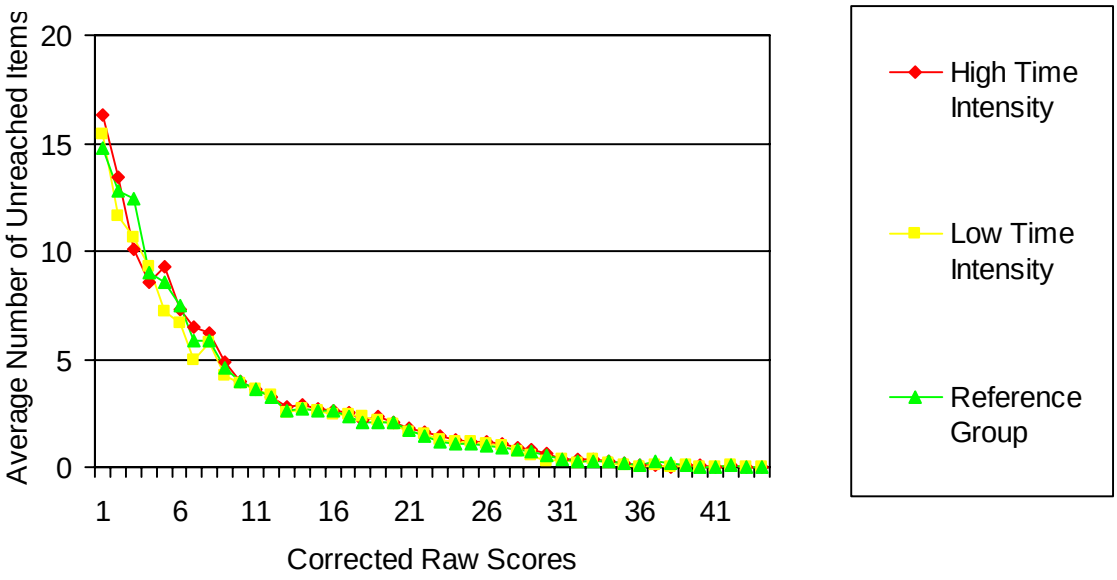
Appendix 4. Average Number of Unreached Items for Verbal Form 1.



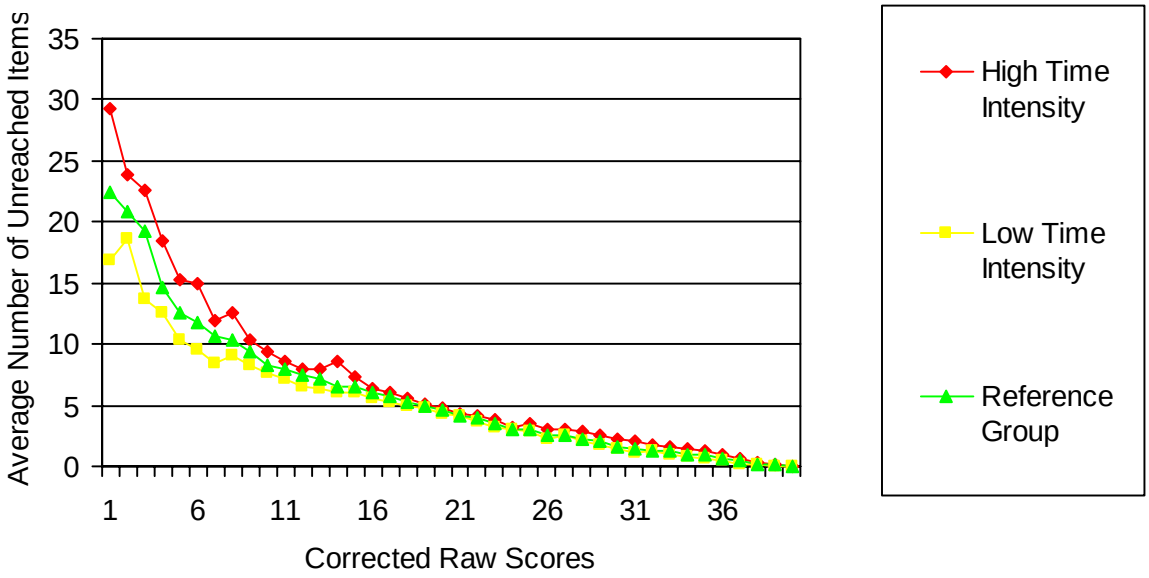
Appendix 5. Average Number of Unreached Items for Verbal Form 2.



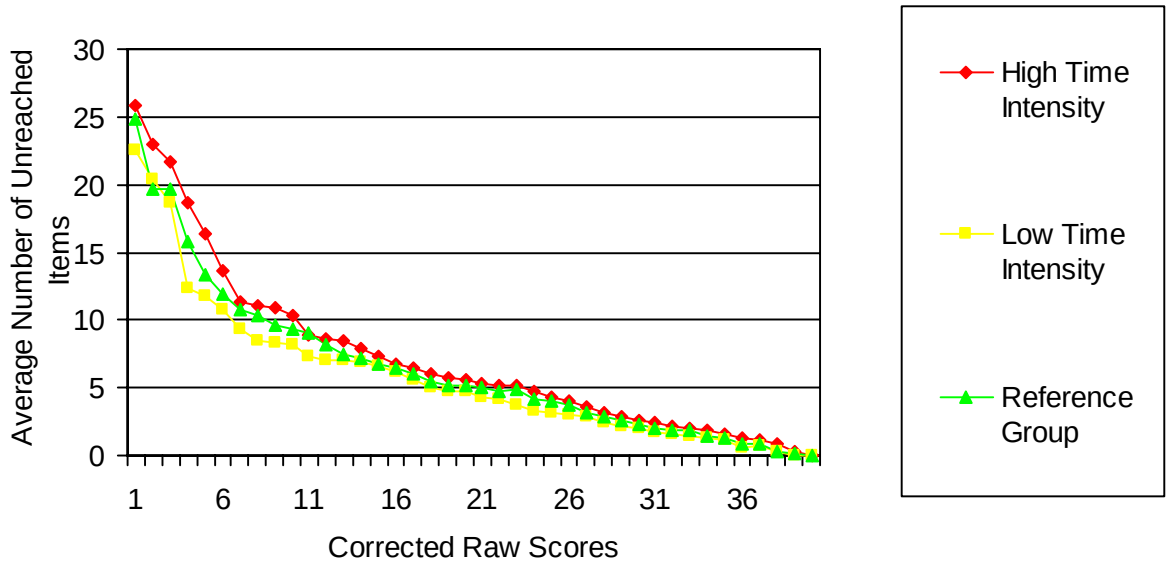
Appendix 6. Average Number of Unreached Items for Verbal Form 3.



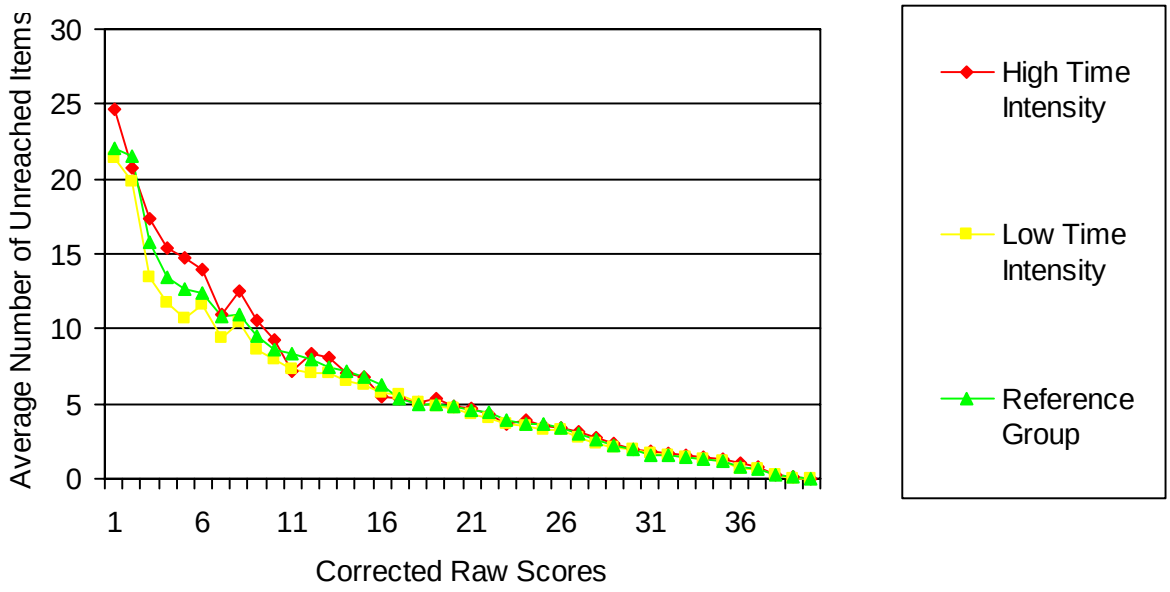
Appendix 7. Average Number of Unreached Items for Quantitative Form 1.



Appendix 8. Average Number of Unreached Items for Quantitative Form 2.



Appendix 9. Average Number of Unreached Items for Quantitative Form 3.



Appendix 10. ACDNR for the Verbal Test Forms

Comparison	Form 1	Form 2	Form 3
High-time-intensity and Reference Groups	0.18	0.37	0.09
Low-time-intensity and Reference Groups	-0.04	-0.29	-0.33

Appendix 11. ACDNR for the Quantitative Test Forms

Comparison	Form 1	Form 2	Form 3
High-time-intensity and Reference Groups	1.56	1.87	1.32
Low-time-intensity and Reference Groups	-1.37	-1.69	-1.18