The Impact of Test Characteristics on Kullback-Leibler Divergence Index to Identify Examinees with Aberrant Responses

Jaehoon Seol, Ph. D.

Jonathan D. Rubright, Ph. D.

American Institute of CPAs

Abstract

This article analyzes the impact of test characteristics on Belov and Armstrong's (2009) two-stage algorithm to identify aberrant candidate responses. The two-stage algorithm developed by Belov et al. (2007) and Belov and Armstrong (2009) is based on Kullback-Leibler Divergence (KLD) and the K-index to detect answer copying by comparing the posterior distributions of candidate ability between the operational and pretest parts of an examination. Because the two-stage algorithm compares these two parts, the accuracy of the procedure is sensitive to the psychometric characteristics and structure of the individual components. However, in many licensure and certification examinations that are administered via CAT, MST, and LOFT, the structural differences between these two parts is not strictly defined. In this study, we analyze how different lengths and difficulties of pretest portions, along with the amount of copying, affect the performance of the two-stage algorithm using Type I and Type II error rates. It is found that Type I error is consistently low across conditions, yet Type II error is very sensitive to pretest length, pretest item difficulty, and the amount of copying simulated.

Introduction

Before the introduction of the two-stage Kullback-Leibler Divergence (KLD) method by Belov et al. (2009) to detect answer copying, many different statistical methods have been developed to detect aberrant candidate responses. These include the K-Index method by Holland (1996), person fit statistics, and cumulative sum statistics (CUSUM), among others. However, most of these methods were designed to detect more general aberrant candidate responses. In contrast, Belov et al. (2009)'s two-stage KLD method was specifically designed to detect the type of answer copying that could happen in a large scale high-stakes test, such as the Law School Admission Council (LSAC) exams. The key idea behind the algorithm is to first filter aberrant candidate responses by using the KLD index, and then compare these flagged responses with all possible source candidates by using the K-index.

As explained in Cover and Thomas (1991), the KLD is defined by

$$D(R||S) = \int_{-\infty}^{\infty} R(\theta) log \frac{R(\theta)}{S(\theta)} d\theta \quad (1)$$

In here, $R(\theta_i)$ is the posterior probability for the operational portion of an exam and $S(\theta_i)$ is the posterior probability for the pretest portion of an exam, further defined by

$$R(\theta_i) = \frac{\prod_{j=1}^{m} P(r_j|\theta_i)}{\sum_{k=1}^{n} \prod_{j=1}^{m} P(r_j|\theta_k)}, \quad (2)$$

and

$$S(\theta_i) = \frac{\prod_{j=1}^{n} P(s|\theta_i)}{\sum_{k=1}^{n} \prod_{j=1}^{n} P(s_j|\theta_k)}. \quad (3)$$

The KLD, denoted by $D(R||S)$, is widely used in information sciences to measure entropy differences between two different signals (Cover and Thomas, 1991). In general, a large KLD value indicates a divergence in the examinee's performance between the two components of the exam. Belov et al. (2009) shows that the two-stage KDL algorithm provides superior performance in detecting answer copying over the K-Index method. Yet, quality performance of this method is based on two preconditions:

- The operational parts for test takers sitting in close proximity are generally identical. This helps find the asymptotic/experimental distribution of the KLD-index in advance.
- The operational and pretest parts of the exam should have statistical characteristics similar to each other to ensure the compatibility of an examinee's performance on the two parts.

However, examinations vary in the extent to which they satisfy these conditions listed above. The operational and pretest portions may have notably different psychometric properties, especially in exam formats such as CAT, CBT and LOFT. Additionally, the statistical properties of pretest items are generally unknown in advance, making it difficult to build a form to satisfy the second condition.

This simulation study considers several factors (the percentage of pretest items in the exam, the difficulty level of the pretest items, the percentage of copying items), and evaluates the impact of these factors on the performance of the two-stage KLD method to detect answer copying. The results of this study will be important in identifying test characteristics where the two-step KLD algorithm may be appropriately applied to identify answer copying and other aberrant candidate response behavior.

Purpose of the Study

As a first step to expand the applicability of the two-stage KLD algorithm to various exam structures such as CAT, MST, and LOFT that are commonly used in licensure and certification exams, this study evaluates the stability of the two-stage KLD algorithm for one of the two preconditions described above. If the operational and pretest parts of the exam have different statistical characteristics, what would be the impact of this difference on the performance of the two-stage algorithm? This study provides an answer to this question by analyzing, via simulation, the performance of the two-step algorithm for exams with mixed total form lengths having different operational-to-pretest length ratios, and for exams with varying difficulty levels of the pretest items in comparison to the operational items. We also manipulate the percent of items copied by copying examinee pairs. More specifically, this study answers the following questions:

- First, how does the ratio of pretest to operational items affect the performance of the two-stage KLD algorithm to detect answer copying? Even if most high-stakes linear exams have a relatively well-defined ratio of pretest to operational items, this structure can be changed very easily during the post-administration review process. Moreover, in many CAT, MST, and LOFT exams that are administered continuously, the pretest items are inserted into the item bank and tested depending on need, making it hard to keep a fixed ratio between operational and pretest items. So, it is important to understand how the two-stage algorithm works when applied to exams with different numbers of pretest items.

- Second, how does the difficulty level of pretest items affect the performance of the algorithm to detect answer copying? Pretest items are by nature items being tested on the real test population. Even if content specialists may have some intuition on difficulty levels of the pretest items, most of the time it cannot be accurately predicted. Since most testing organizations, especially those interested in using CAT, MST, and LOFT, insert several pretest item blocks into the operational pool simultaneously to save cost, it is important to understand how the two-stage algorithm works when applied to exams with different pretest item difficulty levels.
- Third, how does the percentage of answer copying affect the performance of the detection algorithm? In most licensure and certification exams administered through CAT, MST and LOFT, both the percentage of candidates who do the answer copying and the percentage of items whose answers are copied are limited. Belov and colleagues (2009) provide a partial answer to this question when a test has 100 operational and 25 pretest items. They reported an almost 47% increase in Type II error when the percentage of answer copying is reduced from 100% to 60%. In this study, we investigate how different percentages of answer copying affect the performance of the detection algorithm under different exam structures between operational and pretest items.

Methods

The KLD two-stage algorithm is based on two fundamental statistical concepts: Kullback-Leibler Divergence (KLD) (Cover & Thomas, 1991; Kullback & Leibler, 1951) and the K-Index probability (Holland, 1996). Given two posterior distributions $R(\theta)$ and $S(\theta)$ of candidate abilities over operational and pretest parts of the exam, the KLD is defined by Eq. (1). The KLD is a non-equivalent measure of the relative entropy difference between the two posterior distributions. The KLD is transitive, but it does not satisfy the symmetric relationship. Using the same terminology and notation used in Holland (1996), the K-Index is defined as

$$\mathrm{P}\{\mathrm{w}(\mathrm{X},\mathrm{Y}) \geq \mathrm{w} | \mathrm{w}(\mathrm{X}) = \mathrm{w_e}, \mathrm{Y} = \mathrm{X(f)}\} \tag{4}$$

where

e, f	e is the subject and f is the source.
X, Y	Response arrays.
$w(X, Y)$	Number of matching incorrect responses shared by two response arrays X and Y.
$w(X)$	Number of incorrect responses in X.
$X(f)$	Response array by the source f.

It is a conditional agreement probability that measures the proportion of examinee pairs in the population with w or more matching incorrect answers. A detailed rationale of the definition and two equivalent interpretations of the K-Index are described in Holland (1996). Let T represent the total number of items in the exam, w_f the number of incorrect responses by the source, w_e the number of incorrect responses by the subject, and w the number of matching incorrect responses between the source and the subject. Then, the K-Index can be approximated by a binomial distribution (Holland, 1996):

$$K(w_f, w_e, w, T) = \sum_{i=w}^{w_f} \binom{w_f}{i} p\left(\frac{w_e}{T}\right)^i \left(1 - p\left(\frac{w_e}{T}\right)\right)^{w_f - i} \quad (5)$$

In here, the probability $p(x)$ is defined by

$$p(x) = \begin{cases} 0.085 + bx, & 0 < x \leq 0.3 \\ 0.085 + 0.18b + 0.4b, & 0.3 < x \leq 1.0 \end{cases} \quad (6)$$

The probability $\mathrm{p(x)}$ is called the Kling function originally developed by F. Kling and used by Holland (1996) to estimate K-index. The Kling function is a monotonically increasing piecewise linear function. The slope parameter b can be estimated from the empirical data as described in detail by Belov et al. (2009), and can differ from one administration to another. In this study, $b = 0.48$ is used to ensure a conservative estimate for the detection of answer copying.

The KLD two-stage algorithm proposed by Belov et al. (2009) to detect answer copying can be summarized as follows:

Algorithm

Step 1: Given threshold value T_{KL}, create a list of candidates whose KLD value is greater than T_{KL}.

Step 2: For each candidate detected in Step 1, compare the K-index of the candidate with other candidates who belong to the same group as the candidate. If the K-index is smaller than a given threshold value T_K, report the pair of candidates and manually review their seating and test booklets.

Belov et al. (2009) describe the procedures to calibrate the threshold value T_{KL} by approximating cumulative distributions of empirical KLDs using the lognormal distribution. In this simulation study, the threshold value was determined by following a similar procedure, but using the

simulated data set instead of empirical data set and choosing the T_{KL} to be equal to the 5% significance level.

Simulation Design

Together, the study involves three design factors: (1) percentage of pretest items: 5%, 10%, 20%, and 30%; (2) difficulty level of pretest items: easy, medium, and hard; (3) percentage of answer copying: 60%, 70%, 80%, 90%, and 100%. Fully-crossing these design factors leads to $4 \times 3 \times 5 = 60$ different conditions being examined (see Table 1).

Table 1 Simulation Conditions

Design Factor	Design Level	Number of Levels
Percentage of pretest items	5%, 10%, 20%, 30%	4
Difficulty level of pretest items	Easy, Medium, Hard	3
Percentage of answer copying	60%, 70%, 80%, 90%, 100%	5
	Total	60

For each of these 60 conditions, 10,000 person ability estimates are sampled from a normal distribution with mean 0 and standard deviation 1 (i.e. $\theta_i \sim N(0,1)$), and then the 10,000 simulated candidates are randomly split into 100 groups of 100 candidates. These groups represent the group of candidates taking the test at the same test center. All candidate responses are generated using the three-parameter logistic function $P_i(\theta)$. To simulate answer copying, we add 100 aberrant pairs, one pair in each of the 100 groups. The ability level θ_f of the source follows the uniform distribution $U(0,3)$, and the ability level θ_e of the subject is chosen so that $\theta_f = \theta_e + 2.5$. This is done to ensure a meaningful ability level difference between the source and the subject regardless of the difficulty level of the administered exam.

Table 2 Difficulty Parameter Distributions by Condition

		Operational Items	Pretest Items 5	10	20	30
Easy	**Mean**	-0.5142	-2.15633	-2.1689	-2.18265	-2.12252
	Std	0.892716	0.356177	0.574203	0.528753	0.950627
Medium	**Mean**	-0.5142	-0.47093	-0.5353	-0.50591	-0.55289
	Std	0.892716	0.402618	0.682701	0.584604	0.970799
Difficult	**Mean**	-0.5142	0.982048	0.995776	1.044194	0.966271
	Std	0.892716	0.966494	0.974198	0.946799	1.197682

For the simulation study, 12 different forms are generated in total. All forms have 100 operational items so that the percentage of pretest items matches the number of pretest items in each form. Table 2 shows means and standard deviations of item difficulties in these forms. All forms had the same operational part, and the operational items have mean difficulty value -0.5142 and standard deviation 0.892716. The first four forms have relatively easier pretest items compared to the operational part. Even if they have a different number of pretest items, the mean values of these pretest items are close to -2.15. The next four forms have pretest items with almost the same difficulty as the operational part. The final four forms have relatively harder pretest items compared to the operational part, with mean difficulty levels close to 1.0.

Results

All algorithms used in this study are implemented in MATLAB because of its high accuracy, which is especially important when computing and comparing posterior distributions requiring high levels of precision.

As explained above and shown in Table 1, the three main factors manipulated in this simulation study are (1) the percentage of pretest items (5%, 10%, 20%, and 30%), (2) the difficulty level of pretest items (easy, medium, and hard), and (3) the percentage of answer copying (60%, 70%, 80%, and 90%, and 100%). The results of these analyses are shown in Table 3 through Table 5 for the easy pretest items (Table 3), medium pretest items (Table 4), and hard pretest items (Table 5) respectively. All tables show Type I and Type II error rates, along with the number of correctly and incorrectly flagged examinee pairs broken out by number of pretest items included on the exam and the proportion of items that were copied. The Type I error rate shows the proportion of examinee pairs that were incorrectly classified as copying answers. The Type II error rate shows the proportion of examinee pairs who were actually simulated to be copying, yet were not flagged by the KLD two-stage algorithm.

Looking at the results across Table 3 through Table 5, four patterns emerge. First, Type I error rates are consistently low and almost close to 0, regardless of condition. This pattern is similar to the results shown in Belov et al. (2009). This tells us that the procedure rarely inappropriately flags examinee pairs. Second, Type II error rates appear to be related to the number of pretest items included on the exam. As the number of pretest items increases, Type II error decreases. Thus, the procedure appears to gain accuracy in copying identification as the pretest portion lengthens. Third, Type II error rates appear to be affected by the difficulty level of

the pretest items. When the pretest items have a medium difficulty level, similar to the difficulty level of the operational items, the procedures appears to have higher accuracy in detecting answer copying. Fourth, Type II error also appears to be related to the percentage of items copied: as the percentage of copying increases, Type II error decreases. Again, the procedure gains accuracy with a higher percentage of copied items. Together, the difficulty level and the number of pretest items included on an exam, along with the percentage of answers actually copied, significantly impacts the sensitivity of this procedure. Graphing these Type II errors may make these relationships clearer; since the Type I error rates are so consistently low, they are not further explored.

Figure 1 through Figure 3 graph Type II error against the percentage of items copied for all pretest lengths for the easy items (Figure 1), the medium items (Figure 2), and the hard items (Figure 3). These graphs clearly show the trends noted in the previous paragraph from reviewing the Tables. First, length is consistently ordered in all three Figures: higher numbers of pretest items show consistently lower Type II error. Second, the lines consistently show a decrease from left to right, visualizing how Type II error decreases as the percentage of answer copying increases. Together, the Figures and the Tables show that both the pretest length and percent of answer copying are important design factors. The next Figures attempt to shed light on the final design factor, that is, the impact of the difficulty of the pretest items compared to the operational test portion.

Figure 4 through Figure 7 show the Type II error across the different difficulty levels of the pretest items, holding the other factors constant. The Figures are repeated for the 5 item pretest length (Figure 4), the 10 item pretest length (Figure 5), the 20 item pretest length (Figure 6), and the 30 item pretest length (Figure 7). Graphing these values allows a final pattern to

emerge: across all four Figures, the easy item pretest portions show the worst Type II error performance, and the medium difficulty performs best, closely followed by the hard pretest portion.

Together, the Tables and Figures tell a consistent story that the performance of the two-stage KLD procedure under study is rather dramatically impacted by the characteristics of the pretest portion included in an exam. Specifically, the procedure's performance is worse when the pretest portion is shorter, easier, and has less copying behavior. The procedure performs best when the pretest portion is longer and with a difficulty level matched to the difficulty level of the operational portion. Still, the Type I error is relatively low and unchanged by these factors.

Discussion

Recent scandals across a range of high-stakes tests have generated a renewed interest in statistical methods for identifying inappropriate examinee behavior. This has led to a variety of statistical methods being proposed, and heavily researched, for this purpose. This article focuses on one of these methods: the two-stage KLD procedure. Although this procedure has shown promise for identifying pairs of examinees likely sharing answers, it depends on strong preconditions, including that the operational and pretest portions of an exam need share similar characteristics. However, depending on the type of examination being implemented, this precondition may either (1) not be known in advance, or (2) not be possible at all. Thus, this study aimed at looking at the applicability of this procedure to different examination structures by varying the amount of copying behavior, the length of the pretest portion of the examination, and the difficulty level of the pretest portion of the examination.

This procedure relies on a comparison between the posterior ability distributions from the operational and pretest portions of an examination. If they differ significantly, we may posit that cheating behavior is present. By examining the way the procedure works, we can hypothesize that the factors considered here may impact its performance. Theoretically, we may expect that longer pretest portions may lead to better performance of the procedure because a longer form should lead to higher "reliability" for that portion of the exam, leading to a more consistent posterior distribution for the pretest posterior. Similarly, higher rates of answer copying should also translate into a greater distinction between posteriors, leading to higher rates of correct identification and lower rates of Type II error. Thus, if the pretest distribution should truly be different from that of the operational portion, both longer pretest portions and higher levels of cheating should lead to a higher likelihood of determining that the posteriors are, indeed, different.

Next, we may even anticipate the trend that the easiest items would have the highest error rates and lowest power. First, the procedure itself assumes consistency between both portions of the examination. So, the medium pretest conditions would be expected to perform best, as the operational portion was also built from medium difficulty items. Next, the hard pretest items should also perform well, as they would make a relatively clear distinction between both posteriors. Thus, the empirical results shown above are entirely consistent with what would be theoretically expected.

One consistent overall result is that Type I error rates are very low, approaching 0, regardless of the conditions manipulated here. This is a quite desirable property of a test security statistic. In contrast, Type II error rates are much more influenced by the manipulated factors. The results show that the power of the procedure is increased by increasing the pretest test length

and by matching the difficulty of this portion to the operational test. As noted, this is quite difficult since, by definition, the pretest portion of the exam has no operational data to determine its difficulty. Still, even when fulfilling the desired properties of the procedure of similar characteristics between test portions, the power is still not as high as may be desired for a test statistic. In the ideal case of medium pretest difficulty, 30 pretest items, and 100% answer copying, 97 out of 100 cheating pairs are correctly identified. This would represent rather organized cheating, and power rates decrease rapidly when moving away from this ideal combination of factors, down to 34 out of 100 when examining 60% copying. However, in a legal world where false positives may be more dangerous to an organization than missing an instance of inappropriate examinee behavior, a very low level of false accusations may be a desirable trade-off for rather fair rates of power.

In conclusion, the performance of the two-stage KLD procedure shows consistently low Type I error. However, the procedure's Type II error is highly contingent upon the psychometric properties of the pretest portion of an exam, including difficulty, length, and extent of cheating. Since the characteristics of the pretest portion are not typically known beforehand, this may limit the procedure's operational use depending on the characteristics of the exams, as its power cannot be readily determined until after an exam is administered. Future research should not only look at factors influencing the procedure's error rates, but also at ways in which power can be increased when considering different pretest characteristics and more moderate levels of examinee cheating behavior.

References

Belov, D. I., & Armstrong, R. D. (2009). *Automatic detection of answer copying via Kullback-Leibler divergence and K-Index.* Newtown, PA.: Law School Admissioin Council.

Belov, D. I., Pashley, P. J., & Armstrong, R. D. (2007). Detecting aberrant responses in Kullback-Leibler distance. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino, *New Trends in Psychometrics* (pp. 7-14). Tokyo: Universal Academic Press.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory.* New York: John Wiley & Sons, Inc.

Holland, P. W. (1996). *Assessing unusual agreement bteween the incorrect answers of two examinees using the K-Index: Statistical theory and empirical support.* Princeton, NJ.: Educational Testing Service.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*, 79-86.

Appendix

Table 3 Comparison Study of Type I and II Errors, Easy Pretest Items

Number of Pretest Items	% of Answers Copied	Type I Error	Type II Error	Number of incorrectly reported pairs	Number of correctly reported pairs
5	60	0.0003	87	3	13
	70	0.0001	79	1	21
	80	0.0002	79	2	21
	90	0.0001	80	1	20
	100	0.0000	68	0	32
10	60	0.0002	90	2	10
	70	0.0002	72	2	28
	80	0.0003	69	3	31
	90	0.0001	68	1	32
	100	0.0004	47	4	53
20	60	0.0001	84	1	16
	70	0.0001	70	1	30
	80	0.0002	53	2	47
	90	0.0001	49	1	51
	100	0.0001	33	1	67
30	60	0.0001	78	1	22
	70	0.0003	71	3	29
	80	0.0001	50	1	50
	90	0.0003	32	3	68
	100	0.0003	30	3	70

Table 4 Comparison Study of Type I and II Errors, Medium Pretest Items

Number of Pretest Items	% of Answers Copied	Type I Error	Type II Error	Number of incorrectly reported pairs	Number of correctly reported pairs
5	60	0.0001	89	1	11
	70	0.0003	76	3	24
	80	0.0000	72	0	28
	90	0.0001	48	1	52
	100	0.0003	43	3	57
10	60	0.0003	71	3	29
	70	0.0001	68	1	32
	80	0.0004	52	4	48
	90	0.0003	34	3	66
	100	0.0003	14	3	86
20	60	0.0006	76	6	24
	70	0.0001	54	1	46
	80	0.0006	30	6	70
	90	0.0003	17	3	83
	100	0.0008	6	8	94
30	60	0.0006	66	6	34
	70	0.0003	48	3	52
	80	0.0005	28	5	72
	90	0.0004	10	4	90
	100	0.0005	3	5	97

Table 5 Comparison Study of Type I and II Errors, Hard Pretest Items

Number of Pretest Items	% of Answers Copied	Type I Error	Type II Error	Number of incorrectly reported pairs	Number of correctly reported pairs
5	60	0.0005	97	5	3
	70	0.0003	93	3	7
	80	0.0006	79	6	21
	90	0.0005	63	5	37
	100	0.0003	29	3	71
10	60	0.0006	90	6	10
	70	0.0008	74	8	26
	80	0.0007	56	7	44
	90	0.0007	27	7	73
	100	0.0001	16	1	84
20	60	0.0005	90	5	10
	70	0.0009	70	9	30
	80	0.0010	54	10	46
	90	0.0004	18	4	82
	100	0.0009	10	9	90
30	60	0.0009	81	9	19
	70	0.0011	57	11	43
	80	0.0013	26	13	74
	90	0.0016	13	16	87
	100	0.0014	5	14	95

Figure 1 Comparison of Type II Error for Easy Pretest Items

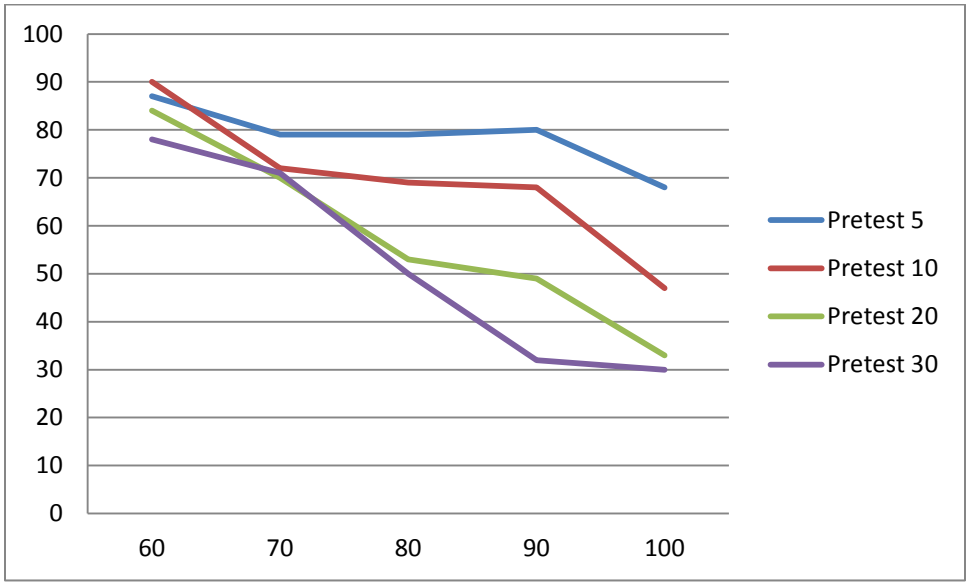

Figure 2 Comparison of Type II Error for Medium Pretest Items

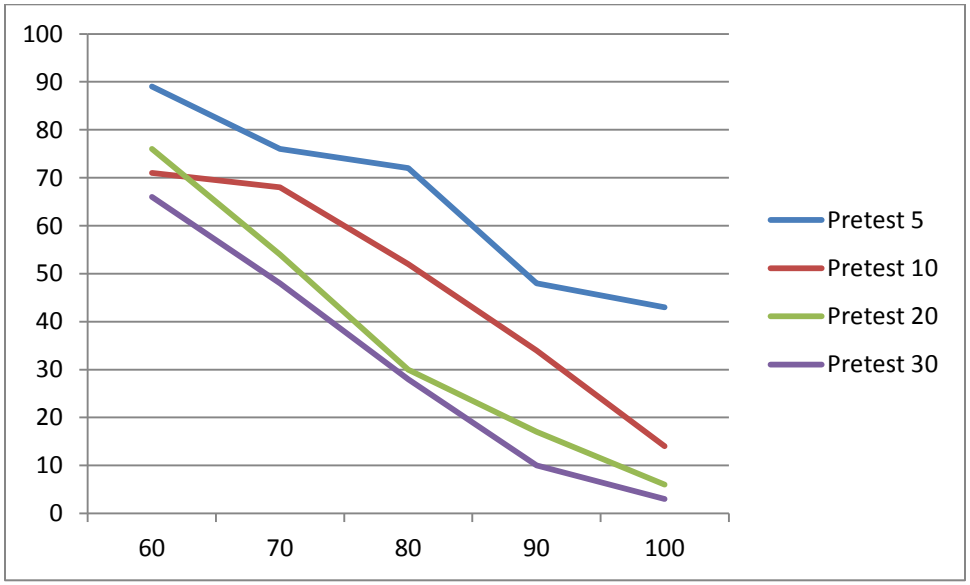

Figure 3 Comparison of Type II Error for Hard Pretest Items

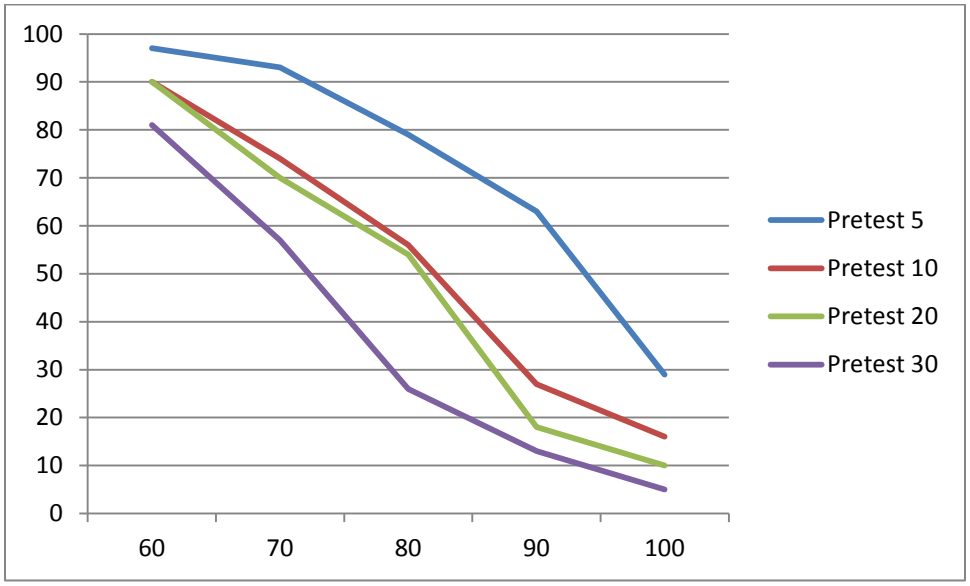

Figure 4 Comparison of Type II Error across Difficulty Levels with 5% Pretest Items

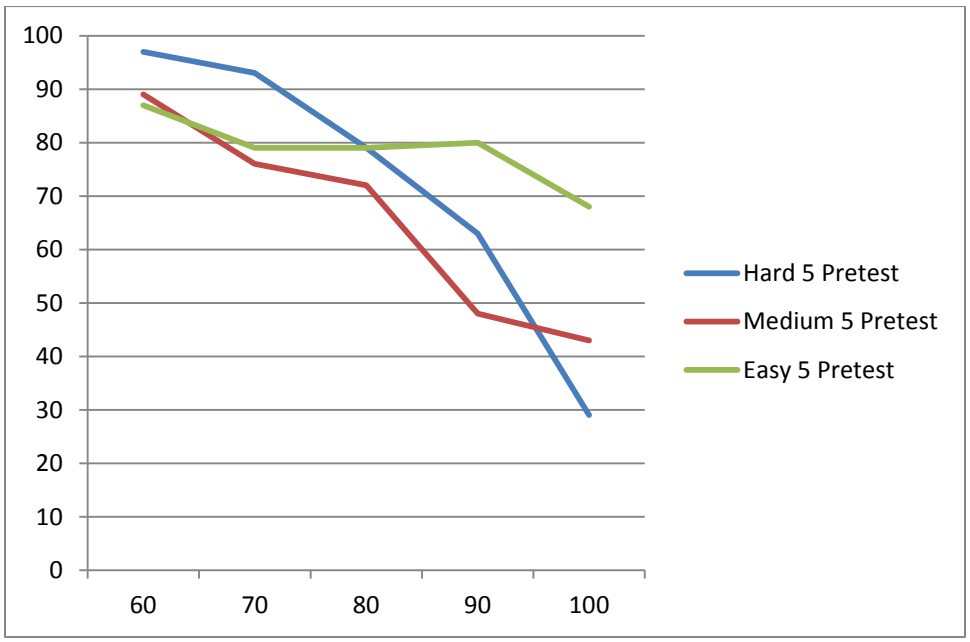

Figure 5 Comparison of Type II Error across Difficulty Levels with 10% Pretest Items

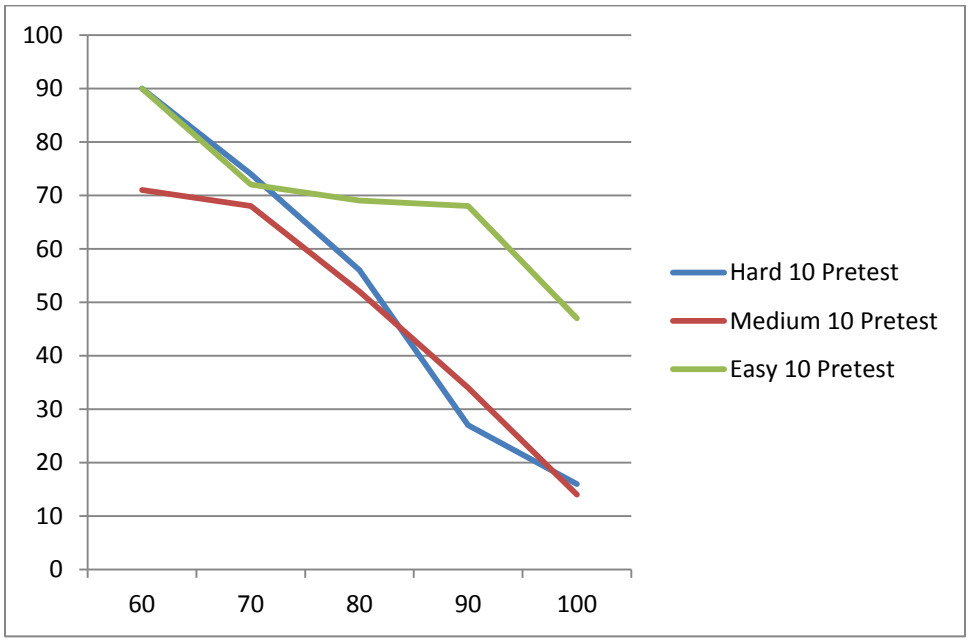

Figure 6 Comparison of Type II Error Across Difficulty Levels with 20% Pretest Items

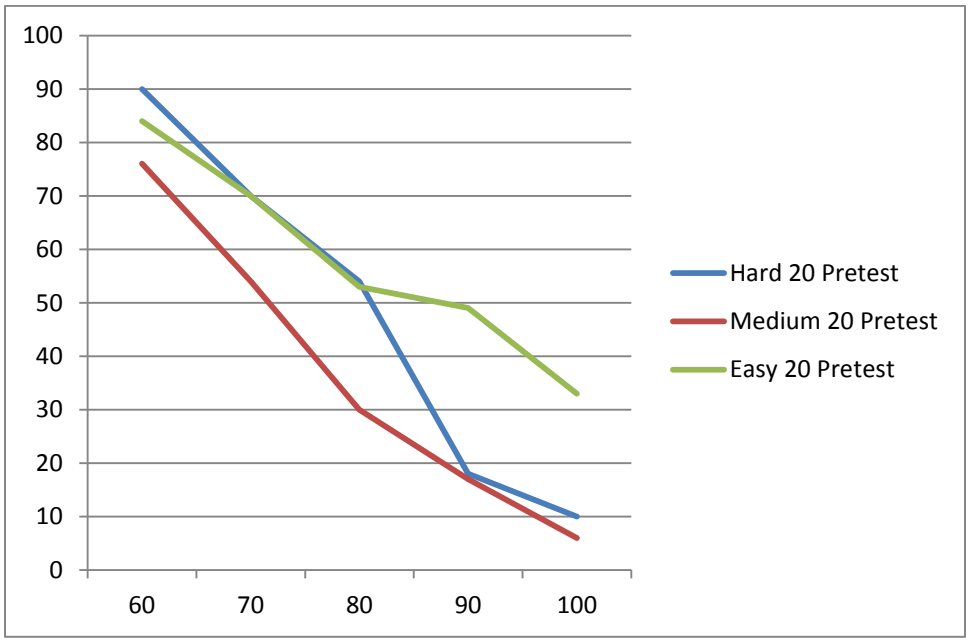

Figure 7 Comparison of Type II Error Across Difficulty Levels with 30% Pretest Items