A Parametric Approach to Detect a Disproportionate Number of Identical Item Responses on a Test

Leonardo S. Sotaridona, Arianto Wibowo, & Irene Hendrawan

Measurement Inc.

Abstract

Routine statistical screening and detection of systematic test irregularities, e.g., cheating orchestrated by school administrators, is an important aspect of test security procedures to ensure that examinees are treated fairly and the assessment results are valid and reliable. Not only that the use of routine statistical screening serves as additional layer of test score integrity check, it also serves as a deterring factor and provides useful information that can be used by state education agency to gauge the effectiveness of test irregularities prevention protocols currently in place. Two enhancements to the statistical procedure proposed by Sotaridona & Choi (2007) that can be used to detect possible occurrence of systematic cheating on statewide assessment program were presented in this paper. Empirical and simulation studies were conducted to investigate the usefulness of the new method for varying class sizes, and number and difficulty of items copied. The first enhancement utilized item response theory model to estimate the item response probability instead of using a non-parametric approach presented in the previous paper. This enhancement capitalizes on the fact that when the IRT model holds, parameter estimates are invariant of the examinee population and therefore circumvent the population-dependency problem by the previous method. The second enhancement is an adjustment to the normalization procedure aimed to bring the distribution of the test statistic consistent with the assumed distribution. Results from empirical and simulation studies showed that the proposed normalization made the distribution of the test statistic practically identical to standard normal. Consequently, significant improvements in the error rates was noted, in particular, the error rates are consistently below the nominal level for medium to large class sizes. Results further showed that the parametric method exhibited promising detection rates and are consistently more powerful than the non-parametric method.

Key words**:** *cheating on test, response similarity analysis, routine screening of systematic cheating*

1. Introduction

Preventing test irregularities such as systematic cheating on statewide assessment program is one of two important aspects of test security procedures – the second aspect is screening or monitoring for possible occurrence of test irregularities. Manually checking students' responses in each testing room for irregularities is a daunting task given a very large number of testing rooms throughout the states and a short turn-around time required for reporting students' scores. Statistical methods for detecting cheating on test can be broadly categorized as individual-level (Angoff, 1974; Frary, Tideman, & Watts, 1977; Bay, 1994; Holland, 1996; Wollack, 1997; Lewis & Thayer, 1998; Sotaridona, van der Linden & Meijer, 2006; van der Linden & Sotaridona, 2006) and group-level. For individual-level, the unit of analysis is on pair of examinees, one being the source and the other the copier. Group-level on the other hand focuses on group of examinees, for example, examinees in an entire classroom. The focus of this paper is on group-level cheating detection method.

In 2007, Sotaridona & Choi presented at the NCME conference a routine statistical procedure designed to screen or monitor possible occurrence of systematic cheating in statewide assessment settings by analysing the similarity of item responses of all unique examinee pairs in each testing room. The end-product of this routine statistical procedure is a list of testing rooms, with their corresponding school and district information, flagged for possible test irregularities. The number of flagged testing rooms is often a very small subset of testing rooms in the entire state that is manageable enough for further investigation. Although the routine screening procedure proposed by Sotaridona & Choi has been shown to have good statistical properties and have been used by a number of states, its detection rate is not known. The purpose of this paper is twofold – (1) to introduce two enhancements to the previous procedure, and (2) to investigate the statistical properties of both procedures under

varying conditions such as class size and number and type of items copied. The outline of the paper is as follows. Section 2 revisits the statistical test of group-level cheating proposed by Sotaridona & Choi and also introduce the enhancements suggested in this paper. Section 3 presents the research methodologies. The results are presented in Section 4 followed by a discussions presented in Section 5.

2. Statistical Test of Group-Level Cheating

2.1. Non-Parametric Method

Let $i=1,2,\ldots,N$ denote multiple-choice items with options $k=1,2,\ldots K_i$, $M_{jj'}$ the number of match item responses by an examinee pair (j, j'), $j' \neq j$ and $P_{i_k j}$ denotes the response probability to option k of item i by examinee j. The response probability $P_{i_k j}$ is estimated using nonparametric approach, e.g., proportion of examinees that chose option k. The expected probability that (j, j') will match on their response to item i is $P_{jj'i} = \sum_{k=1}^{K_i} P_{i_k j} P_{i_k j'}$.

The standardized value

$$Z_{jj'} = \frac{M_{jj'} - \sum_{i=1}^{N} P_{jj'i}}{\sqrt{\sum_{i=1}^{N} P_{jj'i}(1 - P_{jj'i})}}, \qquad (1)$$

is asymptotically standard normal (Sotaridona & Choi, 2007) where $\sum_{i=1}^{N} P_{jj'i}$ and $\sum_{i=1}^{N} P_{jj'i}(1 - P_{jj'i})$ are expectation and variance of $M_{jj'}$ respectively. Let (μ_u, σ_u^2) denotes the mean and variance of $Z_{jj'}$ within class u. Classes could be testing rooms. When there is

copying in a class, μ_u would deviate from its expectation μ, e.g., larger than 0. Hence, the null hypothesis $H_0: \mu_u - \mu = 0$ is tested against alternative hypothesis $H_1: \mu_u - \mu > 0$. A class is flagged for cheating if $T_u > z^*$, where for a level of significance α, $\Pr(T_u \geq z^*) = \alpha$,

$$T_u = \frac{\mu_u - \mu}{\sigma_u}, \tag{2}$$

where $\sigma_u = \frac{\sigma}{\sqrt{n_u}}$ and T_u is asymptotically standard normal.

2.2. Parametric Method

The first enhancement presented in this paper uses the nominal response model (Bock, 1972) to estimate the probability that examinee *j* with ability level θ_j selecting option *k*, $P_{i_k}(\theta_j)$ which is given by

$$P_{i_k}(\theta_j) = \frac{\exp(\xi_{i_k} + \lambda_{i_k}\theta_j)}{\sum_{v=1}^{K_i} \exp(\xi_{i_k} + \lambda_{i_v}\theta_j)}, \tag{3}$$

where ς_{i_k} and λ_{i_k} are the intercept and slope parameters. Alternatively, other polytomous item response model can also be used instead of the nominal response model (see for example van der Linden & Hambleton, 1997). The second enhancement replaced (μ_u, σ_u) in (2) by the mean and standard deviation of the class means to normalize the class mean, that is

$$T_u^* = \frac{\mu_u - \overline{\mu}}{\sigma_{\overline{u}}}, \tag{4}$$

where $(\overline{u}, \sigma_{\overline{\mu}})$ are the mean and standard deviation of the class means. As with the non-parametric method, a class is flagged for cheating if $T_u^* > z^*$.

3. Methods

3.1. Data and Analysis Plan

The data we used is a fifth grade mathematics achievement test from a statewide assessment program. It is a 4-option multiple-choice test consisting of 33 items. Information for each examinee, in addition to the item responses, includes district, school, and classroom code information. The classroom codes allow us to track which examinees belongs to which class, a crucial information needed to conduct a statistical test of cheating using classroom as the unit of analysis. The class code is also needed when doing the simulation of class-level cheating. The analysis plan includes (a) checking the distributional assumptions of the test statistic, (b) compare the Type I error rates, (c) compare the detection rates, and (d) evaluate the decision consistency of the two detection methods.

3.2. Factors, Level of Significance, and Calibration Software

The statistical properties of the group-level cheating detection method presented in this paper were investigated under three varying condition, namely, class size, number of items copied, and type of items copied. Class size is categorized as small, medium, large representing the lower third, the middle third, and the upper third of class sizes in the population. The number of items copied could be 3, 7, 10, or 13 that represent 10%, 20%, 30%, or 40% copying. The type of items copied could be easy items (upper 55^{th} percentile of the p-values), difficult items (lower 55^{th} percentile of the p-values), or random. The levels of significance are in the range 0.0005 to 0.05 with increment 0.005. The item parameters of Bock's nominal response model were estimated using Multilog Version 7.03 (Thissen, 1991).

3.3. Data Simulation and Cheating

The data that we used to investigate the Type I error and detection rates consisted of real a data sets that has been re-shuffled, that is, examinee class-level information such as class id is retain while the item response strings are replaced by response strings randomly selected from other classes in the population. This re-shuffling approach has the advantage that the generated data has the feel of a real dataset, e.g., it contains some data-noise that are most often cannot be captured through real simulation. Classes with less than 10 observations were excluded. For each iteration, 2% of the classes were exposed to cheating. The simulations were carried out as follows: (i) identify the classes that will be exposed to cheating, (ii) select which type of item to simulate, e.g., easy, difficult or random, (iii) identify how many items will be copied, (iv) for the item identified in steps (ii)-(iii), change to correct response the responses of examinees in step (i), (v) estimate the item paramaters using multilog; for the

non-parametric method, compute the conditional p-values, (vi) compute the test statistics, (vii) repeat steps (i)-(vi) to complete 500 iterations, (viii) compute the Type I error rates, detection rates, and decision consistency, (ix) repeat steps (i)-(viii) for all other set of factors.

4. Results

4.1. Distribution of Standardized Number of Match Item Responses

The statistical test in Equation 2 assumed that the standardized number of match item responses (Z) is normally distributed. This section briefly revisits and verifies that assumption with empirical data. Visual inspection of a histogram in Figure 1 and a density plot in Figure 2 clearly supports normally of Z. Quantile-quantile plot (QQ-plot) of Z as shown in Figure 3 further supports the normality assumption. QQ-plot is used to assess whether data have a particular distribution, or whether two datasets have the same distribution. If the distributions are the same, e.g., if Z is normally distributed, then the plot will be approximately a straight line (Chambers et.al, 1983). The picture depicted in Figure 4 also shows that the normality assumption holds in the presence of answer copying.

It must be noted however that the previous statistical test (Sotaridona & Choi (2007) assumed standard normal distribution of Z. Although the mean is very close to zero, the standard deviation is around 0.8 (see Figure 1 & Figure 4; see also Table 3 in Sotaridona & Choi (2007)). One of the enhancements presented in this paper takes into account this observation that the standard deviation is slightly off the assumed value of 1. We will revisit this observation in the next section and show how the enhancement brings the Type I error rates at or below the nominal level.

Figure 1. Histogram of Standardized Number of Match Item Responses

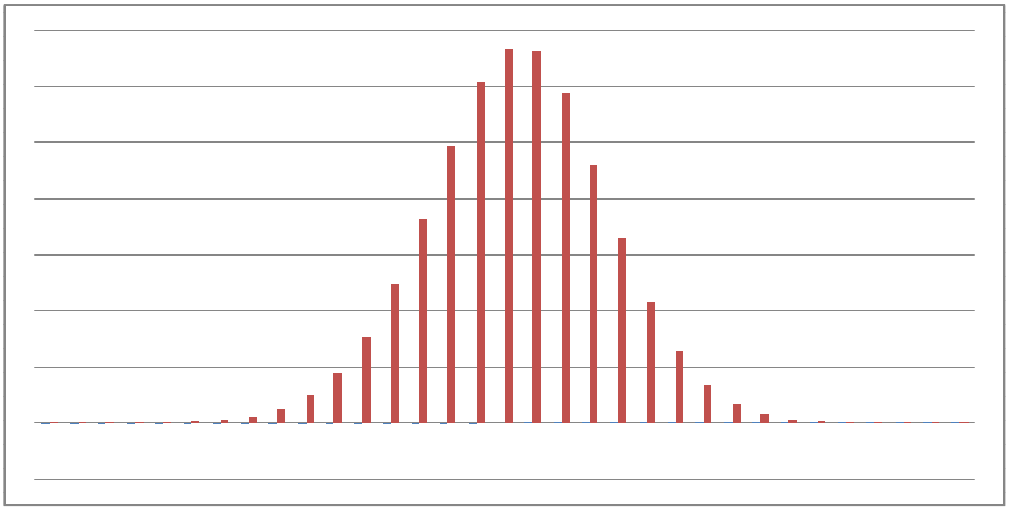


Figure 2. Density Plot of Standardized Number of Match Item Responses

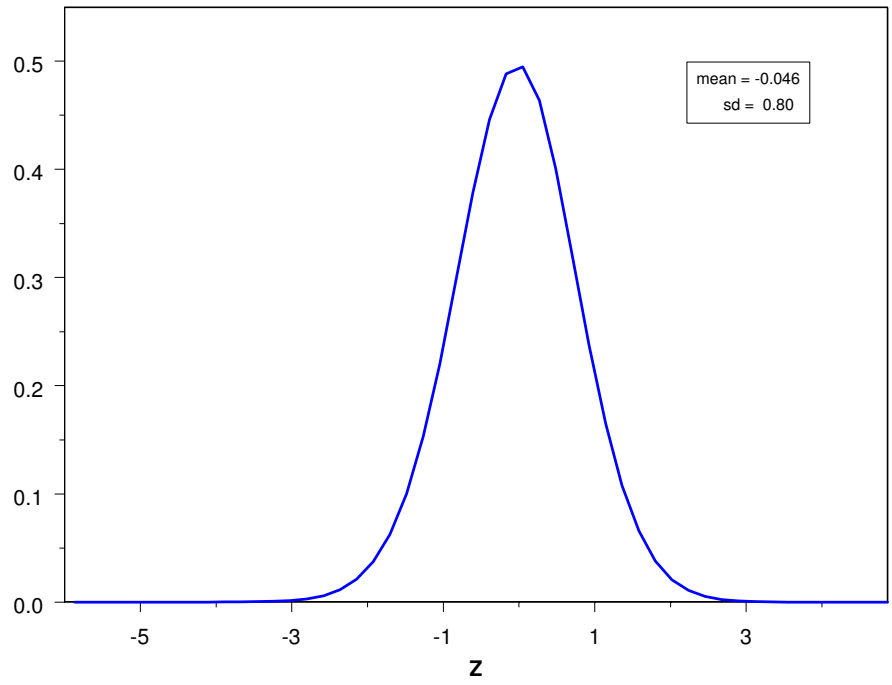


Figure 3. QQ-plot of Standardized Number of Match Item Responses

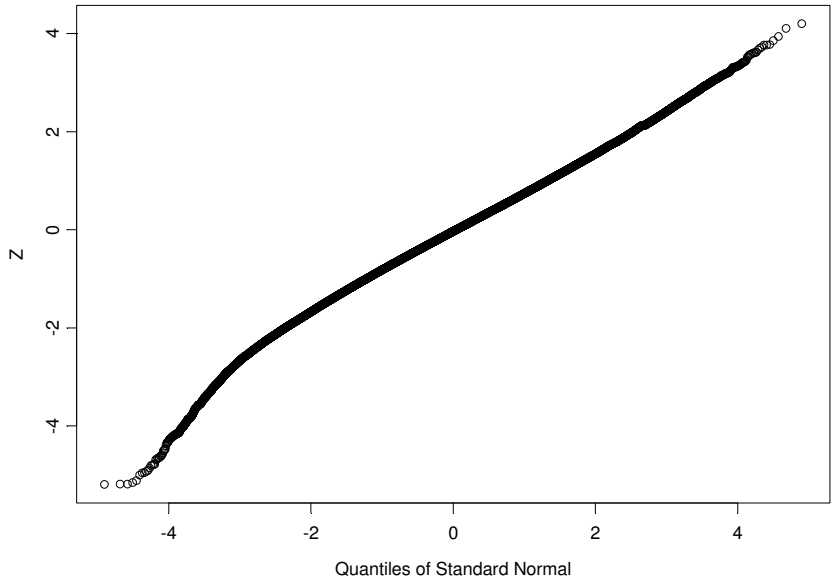
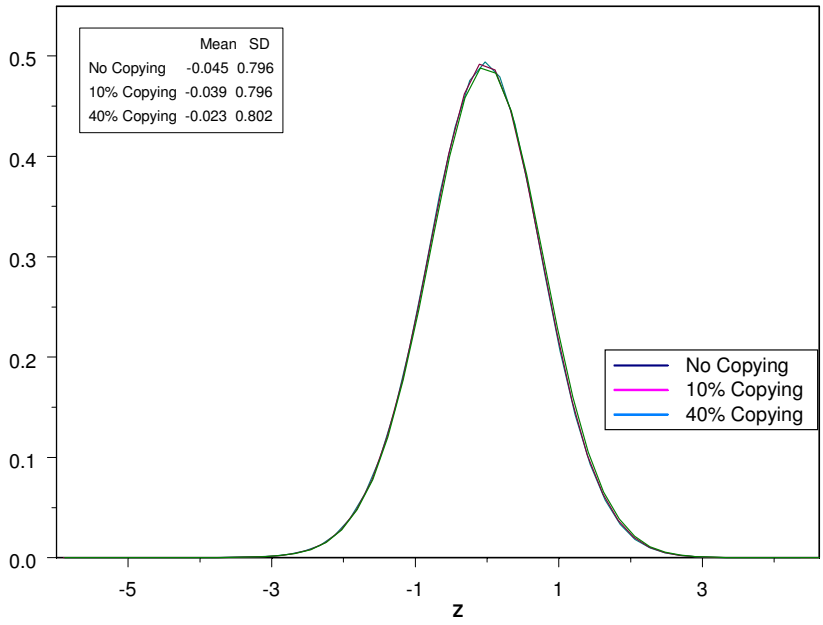


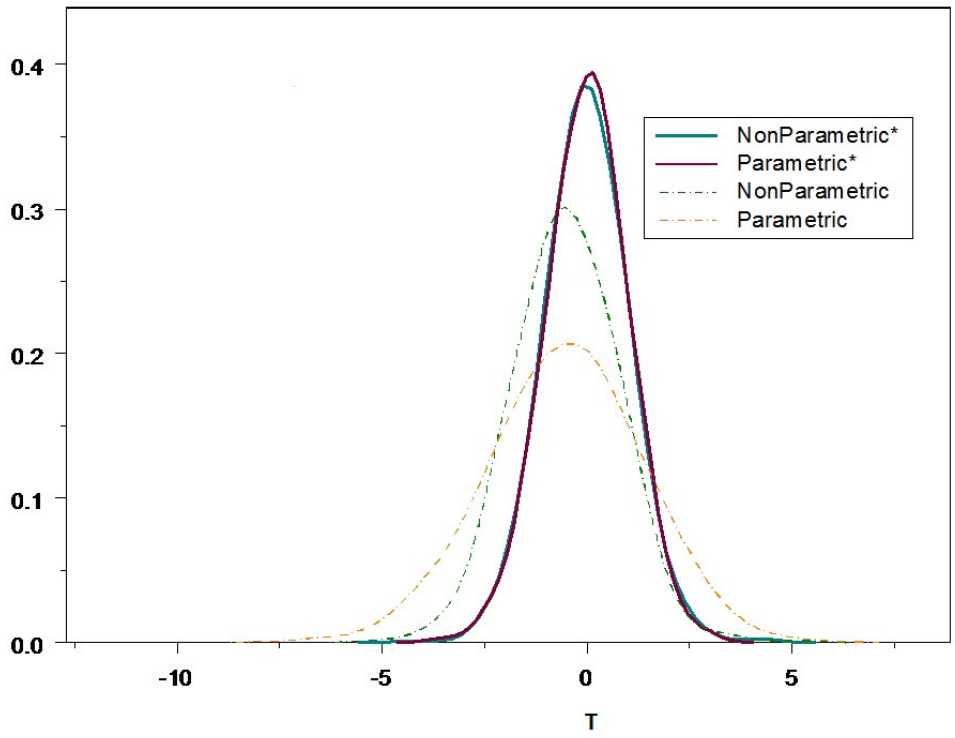
Figure 4. Density Plot of Standardized Number of Match Item Responses When There is Copying.



4.2. Distribution of the Test Statistic

As noted in the earlier section, the distribution of the standardized number of match item responses is standard normal with mean close to 0 but the standard deviation is a bit off from 1, e.g., around 0.8. Consequently, the empirical distribution of the test statistic slightly deviates from standard normal (see Figure 5, dotted lines), e.g., has a wider spread than assumed. Because the critical value of the statistical test is drawn from a standard normal distribution, a test statistic that deviates from standard normal could impact the Type I error rates; in this case, could inflate the error rates. The enhancement suggested in this paper appeared to solve the problem (see Figure 5, solid lines), that is, it brings the distribution of T to standard normal.

Figure 5. Density Plot of Test Statistic.



4.3. Type I Error

The Type I error rates of parametric approach and non-parametric approach as a function of the number of examinees in a class (small, medium, large), number of items copied (10%, 20%, 30%, 40%) and the difficulty of items copied (easy, difficult, random) are shown in Table 1 for level of significance ranging from .0005 to .05 at .005 increment.

A few highlights of the numbers in Table 1: (a) the parametric approach was able to control its error rates better than the nonparametric approach and (b) for medium to large class sizes, the parametric approach showed error rates that are below the nominal level. The highlights in (a) and (b) holds regardless of % copying and item difficulty. Overall, both methods showed excellent control of empirical error rates for practical purpose. For small size classes, setting the level of significance by .005 lower than the actual target when conducting the statistical test will yield a very conservative test. For example, if the analyst aimed for 0.0055 level of significance, (s)he could instead use .0005 when conducting the statistical to ensure that the actual error rates are at or below .00055. Note that the adjustment is not necessary for medium and large class sizes.

Table 1. Empirical Type I Error Rates as a Function of Class Size and Type & Number of Items Copied.

				Easy				Difficult				Random			
Alpha	**Class Size**	**Method**	**Null**	**10%**	**20%**	**30%**	**40%**	**10%**	**20%**	**30%**	**40%**	**10%**	**20%**	**30%**	**40%**
0.0005	Small	Parametric	0.0023	0.0019	0.0012	0.0006	0.0003	0.0017	0.0008	0.0004	0.0002	0.0018	0.0010	0.0005	0.0002
		Non-parametric	0.0056	0.0048	0.0031	0.0019	0.0011	0.0049	0.0037	0.0032	0.0031	0.0048	0.0033	0.0024	0.0017
	Medium	Parametric	0.0003	0.0003	0.0001	0.0001	0.0000	0.0002	0.0000	0.0000	0.0000	0.0003	0.0000	0.0000	0.0000
		Non-parametric	0.0008	0.0007	0.0003	0.0002	0.0001	0.0007	0.0005	0.0004	0.0004	0.0007	0.0004	0.0002	0.0002
	Large	Parametric	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		Non-parametric	0.0002	0.0001	0.0001	0.0000	0.0000	0.0002	0.0001	0.0001	0.0000	0.0001	0.0001	0.0001	0.0000
0.0055	Small	Parametric	0.0133	0.0123	0.0081	0.0050	0.0032	0.0111	0.0060	0.0036	0.0026	0.0116	0.0069	0.0038	0.0027
		Non-parametric	0.0195	0.0175	0.0127	0.0091	0.0060	0.0177	0.0144	0.0130	0.0127	0.0175	0.0132	0.0104	0.0084
	Medium	Parametric	0.0037	0.0031	0.0018	0.0011	0.0004	0.0028	0.0011	0.0007	0.0002	0.0030	0.0013	0.0007	0.0003
		Non-parametric	0.0058	0.0049	0.0032	0.0018	0.0010	0.0050	0.0039	0.0030	0.0032	0.0050	0.0035	0.0022	0.0017
	Large	Parametric	0.0012	0.0011	0.0005	0.0003	0.0001	0.0009	0.0004	0.0002	0.0001	0.0010	0.0004	0.0002	0.0001
		Non-parametric	0.0024	0.0019	0.0011	0.0005	0.0002	0.0020	0.0013	0.0011	0.0011	0.0019	0.0011	0.0008	0.0005
0.0105	Small	Parametric	0.0218	0.0200	0.0138	0.0097	0.0061	0.0184	0.0108	0.0070	0.0050	0.0192	0.0121	0.0076	0.0052
		Non-parametric	0.0286	0.0265	0.0195	0.0140	0.0100	0.0266	0.0217	0.0194	0.0195	0.0264	0.0205	0.0158	0.0134
	Medium	Parametric	0.0072	0.0064	0.0039	0.0024	0.0011	0.0056	0.0028	0.0016	0.0008	0.0060	0.0032	0.0018	0.0009
		Non-parametric	0.0102	0.0088	0.0059	0.0035	0.0022	0.0089	0.0069	0.0057	0.0058	0.0088	0.0062	0.0043	0.0035
	Large	Parametric	0.0029	0.0024	0.0013	0.0007	0.0003	0.0021	0.0009	0.0004	0.0002	0.0022	0.0011	0.0005	0.0003
		Non-parametric	0.0048	0.0038	0.0024	0.0013	0.0006	0.0039	0.0029	0.0023	0.0023	0.0038	0.0027	0.0017	0.0012
0.0155	Small	Parametric	0.0292	0.0269	0.0193	0.0138	0.0091	0.0252	0.0154	0.0105	0.0075	0.0261	0.0171	0.0115	0.0079

				Easy				Difficult				Random			
Alpha	**Class Size**	**Method**	**Null**	**10%**	**20%**	**30%**	**40%**	**10%**	**20%**	**30%**	**40%**	**10%**	**20%**	**30%**	**40%**
		Non-parametric	0.0366	0.0338	0.0254	0.0186	0.0135	0.0339	0.0279	0.0256	0.0252	0.0338	0.0264	0.0211	0.0180
	Medium	Parametric	0.0109	0.0098	0.0062	0.0039	0.0021	0.0088	0.0045	0.0028	0.0016	0.0093	0.0052	0.0031	0.0018
		Non-parametric	0.0144	0.0125	0.0086	0.0054	0.0035	0.0128	0.0102	0.0086	0.0087	0.0126	0.0093	0.0064	0.0053
	Large	Parametric	0.0049	0.0042	0.0025	0.0013	0.0006	0.0036	0.0016	0.0008	0.0005	0.0039	0.0020	0.0010	0.0006
		Non-parametric	0.0071	0.0057	0.0038	0.0021	0.0012	0.0057	0.0046	0.0037	0.0038	0.0056	0.0041	0.0027	0.0020
0.0205	Small	Parametric	0.0361	0.0335	0.0245	0.0183	0.0122	0.0312	0.0198	0.0141	0.0102	0.0323	0.0219	0.0154	0.0107
		Non-parametric	0.0438	0.0407	0.0308	0.0233	0.0172	0.0410	0.0339	0.0314	0.0308	0.0409	0.0321	0.0262	0.0223
	Medium	Parametric	0.0148	0.0131	0.0087	0.0057	0.0032	0.0120	0.0065	0.0040	0.0024	0.0126	0.0074	0.0046	0.0028
		Non-parametric	0.0188	0.0165	0.0115	0.0074	0.0049	0.0167	0.0132	0.0113	0.0116	0.0166	0.0121	0.0088	0.0072
	Large	Parametric	0.0072	0.0062	0.0036	0.0021	0.0010	0.0056	0.0026	0.0014	0.0008	0.0059	0.0031	0.0016	0.0010
		Non-parametric	0.0097	0.0078	0.0054	0.0032	0.0019	0.0079	0.0064	0.0051	0.0053	0.0078	0.0058	0.0037	0.0031
0.0255	Small	Parametric	0.0428	0.0398	0.0296	0.0225	0.0154	0.0375	0.0247	0.0177	0.0130	0.0388	0.0267	0.0190	0.0136
		Non-parametric	0.0505	0.0472	0.0362	0.0280	0.0207	0.0474	0.0393	0.0366	0.0360	0.0472	0.0372	0.0312	0.0264
	Medium	Parametric	0.0189	0.0167	0.0112	0.0076	0.0045	0.0153	0.0087	0.0055	0.0035	0.0161	0.0097	0.0061	0.0038
		Non-parametric	0.0227	0.0201	0.0145	0.0096	0.0064	0.0204	0.0166	0.0145	0.0145	0.0202	0.0153	0.0112	0.0093
	Large	Parametric	0.0096	0.0081	0.0050	0.0031	0.0016	0.0073	0.0037	0.0020	0.0012	0.0077	0.0044	0.0023	0.0014
		Non-parametric	0.0122	0.0100	0.0071	0.0043	0.0026	0.0101	0.0084	0.0066	0.0072	0.0100	0.0077	0.0050	0.0041
0.0305	Small	Parametric	0.0492	0.0460	0.0347	0.0269	0.0186	0.0431	0.0292	0.0217	0.0162	0.0446	0.0314	0.0228	0.0167
		Non-parametric	0.0571	0.0535	0.0413	0.0324	0.0241	0.0537	0.0448	0.0421	0.0410	0.0535	0.0427	0.0359	0.0304
	Medium	Parametric	0.0229	0.0206	0.0142	0.0096	0.0059	0.0189	0.0109	0.0071	0.0048	0.0198	0.0123	0.0079	0.0050
		Non-parametric	0.0268	0.0235	0.0177	0.0118	0.0081	0.0238	0.0197	0.0174	0.0176	0.0235	0.0185	0.0137	0.0115
	Large	Parametric	0.0122	0.0104	0.0069	0.0042	0.0024	0.0094	0.0050	0.0029	0.0017	0.0100	0.0056	0.0033	0.0021

				Easy				Difficult				Random			
Alpha	**Class Size**	**Method**	**Null**	**10%**	**20%**	**30%**	**40%**	**10%**	**20%**	**30%**	**40%**	**10%**	**20%**	**30%**	**40%**
		Non-parametric	0.0150	0.0122	0.0088	0.0053	0.0034	0.0123	0.0104	0.0084	0.0090	0.0123	0.0097	0.0063	0.0054
0.0355	Small	Parametric	0.0557	0.0518	0.0397	0.0311	0.0220	0.0489	0.0333	0.0254	0.0190	0.0505	0.0362	0.0268	0.0198
		Non-parametric	0.0630	0.0595	0.0464	0.0366	0.0275	0.0598	0.0502	0.0474	0.0460	0.0597	0.0479	0.0406	0.0345
	Medium	Parametric	0.0272	0.0243	0.0171	0.0119	0.0074	0.0224	0.0134	0.0089	0.0060	0.0235	0.0150	0.0097	0.0065
		Non-parametric	0.0312	0.0275	0.0205	0.0143	0.0100	0.0277	0.0229	0.0203	0.0206	0.0275	0.0216	0.0165	0.0138
	Large	Parametric	0.0150	0.0127	0.0085	0.0056	0.0032	0.0115	0.0062	0.0038	0.0024	0.0120	0.0073	0.0042	0.0027
		Non-parametric	0.0179	0.0149	0.0109	0.0066	0.0045	0.0150	0.0124	0.0100	0.0109	0.0150	0.0114	0.0077	0.0069
0.0405	Small	Parametric	0.0618	0.0575	0.0446	0.0353	0.0254	0.0546	0.0379	0.0291	0.0219	0.0560	0.0409	0.0308	0.0228
		Non-parametric	0.0689	0.0650	0.0511	0.0410	0.0310	0.0652	0.0554	0.0524	0.0511	0.0652	0.0527	0.0451	0.0387
	Medium	Parametric	0.0315	0.0281	0.0201	0.0142	0.0092	0.0260	0.0160	0.0109	0.0073	0.0270	0.0178	0.0118	0.0080
		Non-parametric	0.0352	0.0313	0.0235	0.0168	0.0118	0.0316	0.0264	0.0232	0.0235	0.0314	0.0246	0.0191	0.0161
	Large	Parametric	0.0179	0.0153	0.0102	0.0069	0.0039	0.0139	0.0077	0.0050	0.0032	0.0146	0.0089	0.0055	0.0035
		Non-parametric	0.0210	0.0173	0.0129	0.0079	0.0055	0.0175	0.0148	0.0119	0.0128	0.0172	0.0135	0.0093	0.0083
0.0455	Small	Parametric	0.0674	0.0632	0.0495	0.0400	0.0291	0.0599	0.0424	0.0328	0.0251	0.0613	0.0454	0.0347	0.0260
		Non-parametric	0.0748	0.0707	0.0559	0.0455	0.0346	0.0709	0.0603	0.0575	0.0559	0.0709	0.0577	0.0497	0.0428
	Medium	Parametric	0.0358	0.0319	0.0234	0.0167	0.0111	0.0297	0.0187	0.0128	0.0090	0.0308	0.0207	0.0140	0.0096
		Non-parametric	0.0395	0.0351	0.0266	0.0192	0.0137	0.0354	0.0295	0.0262	0.0265	0.0352	0.0277	0.0217	0.0186
	Large	Parametric	0.0212	0.0181	0.0124	0.0083	0.0049	0.0163	0.0095	0.0062	0.0039	0.0172	0.0108	0.0067	0.0043
		Non-parametric	0.0241	0.0201	0.0152	0.0094	0.0066	0.0202	0.0172	0.0140	0.0151	0.0200	0.0159	0.0110	0.0098
0.0500	Small	Parametric	0.0729	0.0679	0.0537	0.0434	0.0323	0.0645	0.0462	0.0361	0.0280	0.0662	0.0493	0.0383	0.0290
		Non-parametric	0.0799	0.0754	0.0604	0.0493	0.0379	0.0757	0.0647	0.0617	0.0600	0.0755	0.0621	0.0538	0.0465
	Medium	Parametric	0.0396	0.0357	0.0264	0.0190	0.0127	0.0332	0.0214	0.0146	0.0105	0.0344	0.0235	0.0160	0.0112

				Easy				Difficult				Random			
Alpha	**Class Size**	**Method**	**Null**	**10%**	**20%**	**30%**	**40%**	**10%**	**20%**	**30%**	**40%**	**10%**	**20%**	**30%**	**40%**
		Non-parametric	0.0433	0.0386	0.0295	0.0213	0.0156	0.0389	0.0326	0.0290	0.0293	0.0386	0.0306	0.0240	0.0210
	Large	Parametric	0.0242	0.0206	0.0144	0.0096	0.0059	0.0188	0.0109	0.0072	0.0048	0.0198	0.0122	0.0079	0.0052
		Non-parametric	0.0270	0.0226	0.0173	0.0108	0.0079	0.0227	0.0195	0.0159	0.0172	0.0226	0.0182	0.0126	0.0112

4.4. Detection Rates

Figure 6 shows the detection rates on easy items copied. The non-parametric methods are represented by dotted lines. The number of items copied is reflected in the legend with a number, for example, PAR-3 is detection rate of parametric method on 3 items copied. Similarly, NONPAR-3 is detection rate of non-parametric method on 3 items copied. The numbers 3, 7, 10, and 13 represent 10%, 20%, 30%, and 40% copying. The parametric method consistently outperformed the non-parametric method when the percentage of items copied is 20% or more. The detection rates on difficult items are shown in Figure 7. This time, the parametric method outperformed the non-parametric method in virtually all cases. The same can be said on random copying. As expected, copying increases by % items copied. Copying on at least 40% of the items is almost always detected by the parametric method. A table showing the detection rates on all factors considered in this study can be found in Appendix A.

Figure 6. Detection Rate – Copying on Easy Items

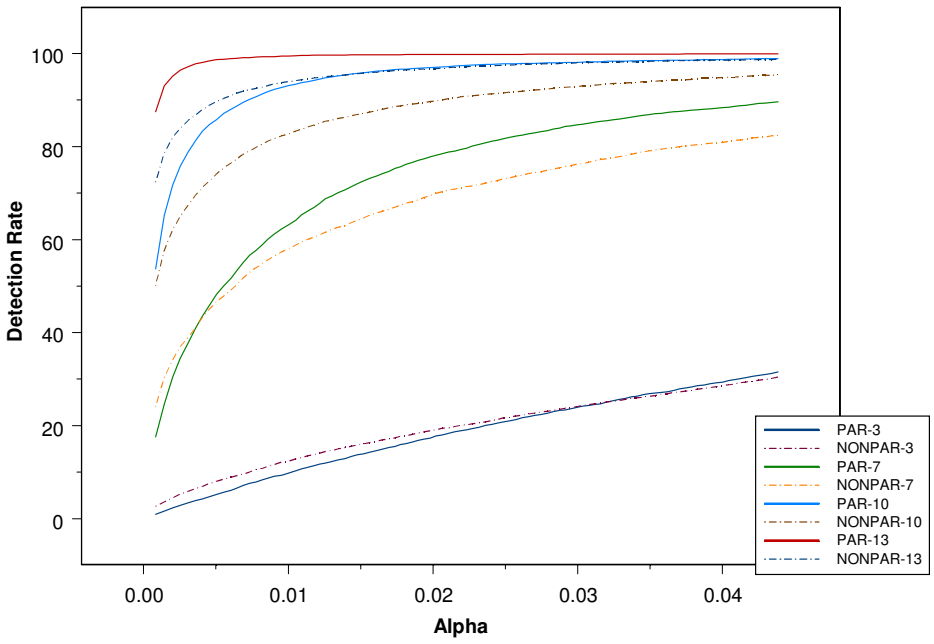


Figure 7. Detection Rate – Copying on Difficult Items

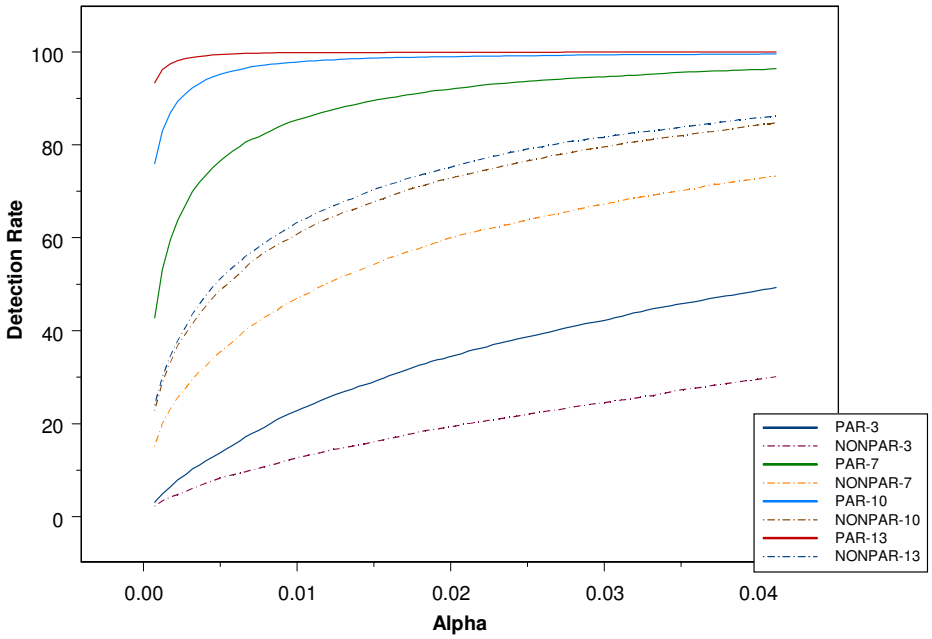
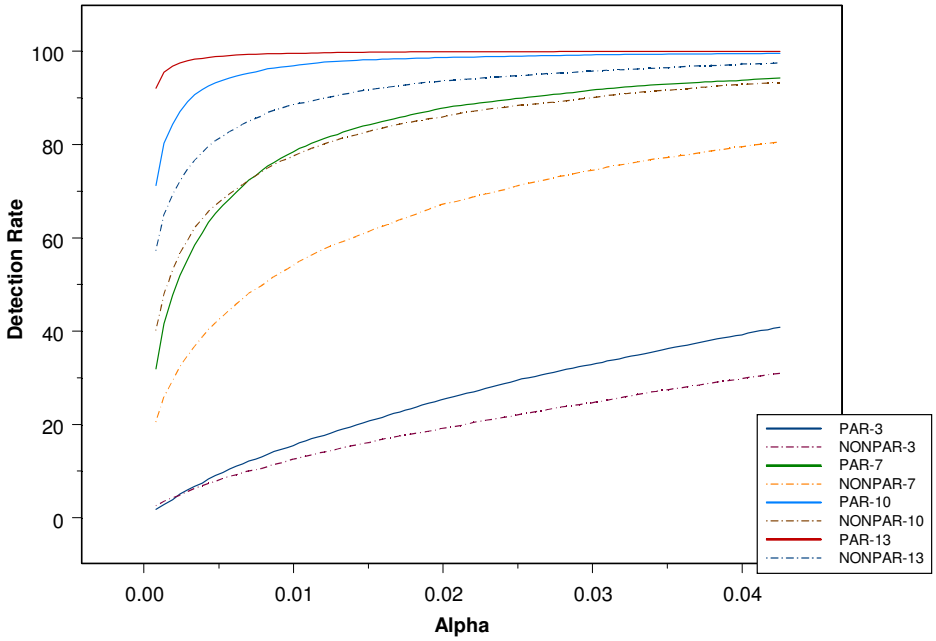


Figure 8. Detection Rate – Random Copying



4.5. Detection Consistency between the Parametric and Non-Parametric Method

In this section we compared the consistency of the two methods for detecting group-level cheating. The decision consistency was computed on three significance levels - .001, .01, & .05. The results are shown in Table 2. Classes were simulated for cheating then the percentages of those classes detected or not detected are reflected as entries in columns 4-7 in Table 2. The first 3-columns in Table 2 are the level of significance (Alpha), type of items copied, and % of items copied, respectively. The 4th & 5th columns are percentages detected by the parametric method while the 6th and 7th columns are percentages not detected by the parametric method. The last column in Table 2 is the % perfect agreement, e.g., % detected by both methods plus % not detected by both methods. The % perfect agreement is the sum of entries in column 4 and column 7.

More than 90% of classes detected by non-parametric method were also detected by the parametric method when % copying is at least 20% and the types of items copied are difficult or random. When the % copying is at least 30%, almost all classes detected by non-parametric method were also detected by the parametric method. Although the detection rates of the parametric method is high (above 90%), with at least 30% copying, the % perfect agreement is lower (below 90%) due to lower detection rates of the non-parametric method. The parametric method performed worst when there is only 10% copying. Using both methods however resulted to a significant increase in the detection rates for the 10% copying but not for 20% copying or more.

Table 2. Detection Consistency

			Parametric Detection		Parametric Non-Detection		
Alpha	Type of Item Copied	% Item Exposed	Non-parametric detection	Non-parametric non-detection	Non-parametric detection	Non-parametric non-detection	% Perfect Agreement
0.0010	difficult	10%	1.15%	3.80%	2.24%	92.80%	93.95%
		20%	16.39%	36.79%	3.60%	43.22%	59.61%
		30%	27.00%	55.92%	1.82%	15.26%	42.27%
		40%	29.55%	66.58%	0.48%	3.39%	32.94%
	easy	10%	0.65%	0.94%	2.90%	95.51%	96.16%
		20%	15.79%	8.81%	14.54%	60.87%	76.65%
		30%	47.94%	17.27%	9.62%	25.17%	73.12%
		40%	76.19%	16.86%	2.25%	4.70%	80.89%
	random	10%	0.92%	1.96%	2.62%	94.51%	95.43%
		20%	18.09%	23.48%	7.87%	50.56%	68.66%
		30%	43.90%	36.35%	3.96%	15.79%	59.69%
		40%	63.44%	32.01%	1.49%	3.06%	66.51%
0.0100	difficult	10%	6.78%	14.69%	5.04%	73.48%	80.26%
		20%	42.22%	42.06%	3.10%	12.62%	54.84%
		30%	58.48%	39.03%	0.68%	1.81%	60.29%
		40%	61.26%	38.50%	0.09%	0.15%	61.42%
	easy	10%	4.36%	5.47%	8.05%	82.12%	86.48%
		20%	45.53%	17.71%	12.43%	24.33%	69.86%
		30%	79.87%	13.26%	2.84%	4.03%	83.90%
		40%	93.67%	5.80%	0.28%	0.25%	93.92%
	random	10%	5.36%	9.68%	6.82%	78.14%	83.50%
		20%	47.25%	30.54%	6.07%	16.14%	63.38%
		30%	75.78%	20.90%	1.14%	2.18%	77.96%
		40%	87.88%	11.61%	0.33%	0.19%	88.07%
0.0500	difficult	10%	23.14%	26.20%	7.00%	43.66%	66.80%
		20%	71.94%	24.44%	1.36%	2.26%	74.20%
		30%	84.49%	15.02%	0.25%	0.25%	84.74%
		40%	86.07%	13.83%	0.07%	0.03%	86.10%
	easy	10%	17.64%	13.89%	12.78%	55.69%	73.33%
		20%	77.34%	12.27%	5.09%	5.29%	82.64%
		30%	94.83%	4.08%	0.60%	0.49%	95.32%
		40%	98.69%	1.23%	0.07%	0.02%	98.71%
	random	10%	20.58%	20.23%	10.39%	48.81%	69.39%
		20%	77.83%	16.39%	2.67%	3.12%	80.94%

			Parametric Detection		Parametric Non-Detection		
Alpha	**Type of Item Copied**	**% Item Exposed**	**Non-parametric detection**	**Non-parametric non-detection**	**Non-parametric detection**	**Non-parametric non-detection**	**% Perfect Agreement**
		30%	92.98%	6.51%	0.26%	0.26%	93.24%
		40%	97.40%	2.55%	0.04%	0.01%	97.41%

5. Discussions

The use of routine statistical test to screen potential occurrence of testing irregularities serves not only as additional layer of test score integrity check, it also serves as a deterring factor and can provide useful information that can be used by state education agency to gauge the effectiveness of test irregularities prevention procedures currently in place. Two enhancements to the statistical procedure proposed by Sotaridona & Choi (2007) that can be used to detect possible occurrence of systematic cheating on statewide assessment program were presented in this paper. Empirical and simulation studies were conducted to investigate the usefulness of the new method under varying conditions such as class size and number and difficulty of items copied.

The first enhancement was to use to the polytomous item response theory (IRT) model to estimate the item response probability instead of the non-parametric approach presented in the previous paper. Recall that the previous method uses the proportion of examinees who responded to an item option as an estimate to the item response probability. Such an estimate is known to be population dependent and sensitive to changes in examinee behaviour, for example, when there is significant copying in a class, the estimate is directly affected. Conditioning on the number-correct score was suggested to minimize the effect of population dependency. The enhancement capitalizes on the fact that when the IRT model holds, parameter estimates are invariant of the examinee population and therefore circumvent the population-dependency problem by the previous method.

The second enhancement is an adjustment to the normalization procedure aimed to bring the distribution of the test statistic consistent with the assumed distribution. Results from empirical and simulation studies showed that the proposed normalization made the distribution of the test statistic practically identical to standard normal. Consequently, significant improvements in the error rates was noted, in particular, the error rates are

consistently below the nominal level for medium to large class sizes. For small class sizes, the error rates are close but slightly higher than the nominal level. This result is not surprising because for small classes, the number of pairs will also be small resulting to larger variability in the distribution of the test statistic. For small size classes, setting the level of significance lower by certain value Δ than the actual target when conducting the statistical test is recommended if the analyst want to ensure the actual error rates are below the target level of significance. The value Δ could be derived from the actual data or through simulation using the approach used in this study. Note that the adjustment is not necessary for medium and large class sizes.

Results further showed that the parametric method exhibited promising detection rates and are consistently more powerful than the non-parametric method. Almost all cases detected by the non-parametric method were also detected by the parametric method for 20% copying or more. Hence, if the focus of the analysis is detection of extreme cases of copying, the former method is not needed when the latter method can be used. However, if detection at the lower % copying is important, e.g., 10% copying, combining both methods would help improve the overall detection rates. As expected, the detection rates increase with class size, item difficulty and number of items copied.

Note that only one data set was used in the study due to very large computation time required to run the simulation. Use of more additional data sets will improve generalizability of the results. One thing to consider when using the parametric approach to improve detection is the longer processing time it required especially if one will use the adjustment Δ to align the error rates with the nominal level. For future research, one could explore the use of dichotomous IRT model instead of polytomous IRT model that was proposed in this paper. When using the dichotomous, it would be interesting to know wether collapsing the options

into dichotomous, e.g., 1/0 as correct/incorrect, would results to significant loss of information.

6. References

Angoff, W.H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Associations*, 69, 44-49.

Bay, M. L. G. (1994). Detection of copying on multiple-choice examinations (unpublished doctoral dissertation, Southern Illinois University, 1987). *Dissertation Abstracts International*, *56(3-A)*, 899.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 46, 443-459.

Cizek, G. J. (1999). *Cheating on tests: how to do it, detect it, and prevent it.* NJ: Lawrence Erlbaum.

Frary, R.B., Tideman, T.N., & Watts, T.M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics, 6*, 152-165.

Holland, P.W. (1996). Assessing unusual agreement between the incorrect answers of two examinees using the K-index: statistical theory and empirical support (Research Report RR-94-4). Princeton, NJ: Educational Testing Service.

Lewis, C. & Thayer, D.T. (1998). The power of the K-index (or PMIR) to detect copying (Research Report RR-98-49). Princeton, NJ: Educational Testing Service.

Sotaridona, L.S., & van der Linden, W.J., & Meijer, R.R. (2006). Detecting answer copying using the kappa statistic. *Applied Psychological Measurement, 30*, 412-431.

Sotaridona, L.S., Choi, S. (2007). A non-parametric approach to detect disproportionate number of identical item responses on a test. A paper presented at the annual meeting of the National Council on Measurement and Education, April 9-13, 2007, Chicago, Illinois.

Thissen, D. (1991). MULTILOG user's guide (Ver._). Chicago: Scientific Software, Inc.

van der Linden, W. J., & Sotaridona, L.S. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics,* 31, 283-304.

van der Linden,W. J. & Hambleton, R.K. (1997). *Handbook of modern item response theory*. NY: Springer-Verlag.

Wollack, J.A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement*, 21, 307-320.

-

Appendix A. Detection Rates by Class Size, Item Difficulty, and Number of Items Copied

			Easy				Difficult				Random			
Alpha	**Class Size**	**Method**	**10%**	**20%**	**30%**	**40%**	**10%**	**20%**	**30%**	**40%**	**10%**	**20%**	**30%**	**40%**
0.0005	Small	Parametric	1.6846	21.2815	52.1348	83.8794	4.5600	42.9690	73.2210	91.2728	2.9335	33.5170	69.1548	89.7072
		Non-parametric	4.0372	26.9353	48.6785	69.3534	3.3401	18.1792	25.5591	26.7904	3.8629	24.1229	40.3427	56.0452
	Medium	Parametric	0.7067	16.9014	53.7102	88.7637	2.8269	43.1612	76.9729	93.9593	1.4134	32.6135	71.6431	93.0516
		Non-parametric	2.0024	23.8811	50.1178	72.4656	1.9140	14.1784	22.1830	22.8580	2.0612	19.7809	39.9764	56.9149
	Large	Parametric	0.3571	13.4161	55.6746	90.7889	1.4683	42.0104	78.3730	95.4746	0.8333	28.7545	73.4524	93.9928
		Non-parametric	1.5476	20.1442	51.9048	76.3221	1.2698	12.2147	19.8649	21.6520	1.4286	16.4998	40.0159	59.3750
0.0055	Small	Parametric	7.7839	51.2902	85.0131	98.2024	16.4392	74.5723	94.2202	99.1882	12.5182	66.0191	92.5937	98.5793
		Non-parametric	10.9207	48.9417	72.2335	87.5616	10.5141	37.5761	49.1141	51.2612	10.8045	44.3317	66.6570	78.8634
	Medium	Parametric	5.9187	51.8623	88.7220	99.2488	14.6054	78.7480	95.7597	99.4992	10.4535	69.1393	94.5819	99.1236
		Non-parametric	8.2744	48.4194	77.7974	91.8023	8.0389	36.6823	49.9853	52.4390	8.4217	44.6948	69.5614	83.6984
	Large	Parametric	4.0873	52.1826	91.1508	99.5194	12.1429	80.3765	97.0238	99.8398	7.6984	70.6448	95.9127	99.5995
		Non-parametric	7.0635	50.3805	80.3571	94.5112	7.1825	34.8819	50.5761	54.2101	7.0238	44.5334	72.7273	86.7788
0.0105	Small	Parametric	12.5182	62.0180	91.1414	99.2172	23.9036	81.8788	96.6018	99.6231	17.1362	75.9061	95.4981	99.1302
		Non-parametric	15.4807	56.5961	78.9718	91.4758	14.5513	45.4335	57.5370	59.2636	14.9288	52.8849	73.6277	85.2421
	Medium	Parametric	10.0707	64.8513	94.0518	99.6244	21.7609	86.0094	97.8799	99.8122	15.3710	79.4366	97.3498	99.6557
		Non-parametric	11.5135	58.8106	84.5406	94.8999	11.3074	46.6980	60.1648	62.8205	11.5135	54.2723	78.5399	88.9862
	Large	Parametric	7.4206	66.3196	95.7143	99.8799	20.3571	87.5851	99.0476	100.000	13.3333	80.6167	98.0556	99.8398
		Non-parametric	10.9921	61.9543	87.5794	96.9151	10.3571	46.4958	62.4553	65.6375	10.7540	55.6268	81.5006	92.7083
0.0155	Small	Parametric	15.6840	68.5996	93.9297	99.4781	28.8702	85.8800	97.6474	99.7391	21.3767	80.8060	97.0375	99.5941

			Easy				Difficult				Random			
Alpha	**Class Size**	**Method**	**10%**	**20%**	**30%**	**40%**	**10%**	**20%**	**30%**	**40%**	**10%**	**20%**	**30%**	**40%**
		Non-parametric	18.4723	61.3221	82.7186	93.3604	17.4267	50.7683	63.2588	65.0913	18.2980	58.1328	78.1005	88.1995
	Medium	Parametric	13.8987	73.3959	96.3192	99.7809	26.8846	89.2645	98.7927	99.8748	19.6113	84.6948	98.4393	99.7496
		Non-parametric	14.8115	63.8498	88.0153	96.4643	14.3993	52.1127	66.4019	68.5428	14.5171	60.6573	83.3677	91.8648
	Large	Parametric	10.7143	74.8098	97.4603	100.000	26.3492	91.1894	99.2063	100.000	18.2143	85.8630	99.0079	99.9600
		Non-parametric	13.6111	68.3620	91.0714	97.7965	13.1349	54.8658	69.3286	72.7747	13.4524	62.9155	85.8674	94.9519
0.0205	Small	Parametric	18.8789	73.4996	95.5272	99.5941	33.0526	88.8084	98.2283	99.7970	25.1815	83.9664	97.6184	99.7101
		Non-parametric	21.0863	65.0623	85.1873	94.7521	20.1568	54.9435	66.9765	69.1215	21.1153	61.9310	81.1211	90.2870
	Medium	Parametric	16.9317	77.8091	97.0259	99.9061	31.5077	91.4554	99.1166	99.9061	23.6455	87.6056	98.7633	99.8748
		Non-parametric	17.5795	68.9202	90.6066	97.2778	17.1378	57.5274	71.6681	73.8899	16.9317	65.5712	85.9582	93.7109
	Large	Parametric	13.8889	80.9371	98.4524	100.000	31.5476	93.1117	99.3254	100.000	21.9841	89.1069	99.2857	99.9600
		Non-parametric	15.9127	72.7673	93.1349	98.4375	15.8333	60.1522	74.2551	77.3456	15.6746	69.0028	89.2418	96.1939
0.0255	Small	Parametric	22.0738	76.8919	96.3985	99.6811	36.6250	90.5190	98.4026	99.8260	28.2893	86.4888	98.1412	99.7970
		Non-parametric	23.2356	67.9037	87.1043	95.6799	22.4223	58.3357	69.5033	72.6008	23.2936	65.2943	83.3575	91.7077
	Medium	Parametric	19.3168	81.3772	97.7915	99.9061	35.3946	92.8951	99.2933	99.9061	26.9140	89.7653	98.9694	99.9061
		Non-parametric	19.6113	72.8951	92.4028	97.9036	19.4346	62.0657	75.2868	77.7986	19.4935	70.1721	88.2838	94.7747
	Large	Parametric	16.7460	84.2211	99.0079	100.000	35.7540	94.9539	99.5238	100.000	26.4286	91.1093	99.4048	100.000
		Non-parametric	18.2937	75.9712	94.7619	98.8381	18.2540	64.2371	78.2678	80.5533	18.5714	72.9676	91.5046	97.0753
0.0305	Small	Parametric	24.6297	79.5013	96.8632	99.6811	39.3262	91.8527	98.6640	99.8260	31.7165	88.1415	98.4026	99.8260
		Non-parametric	25.1815	70.5422	88.4694	96.2598	24.3392	60.9452	72.2045	75.0362	25.7043	68.6576	85.2745	92.9545
	Medium	Parametric	22.3793	84.1628	98.2038	99.9374	38.5159	94.0219	99.4111	99.9061	29.9470	91.2363	99.2933	99.9374
		Non-parametric	22.1731	75.7746	93.3451	98.1852	21.7609	64.8826	78.4937	80.8943	21.9965	73.3020	89.4613	95.3379
	Large	Parametric	19.4841	87.0645	99.0873	100.000	39.4444	96.0352	99.6032	100.000	29.8810	92.4710	99.5238	100.000

			Easy				Difficult				Random			
Alpha	**Class Size**	**Method**	**10%**	**20%**	**30%**	**40%**	**10%**	**20%**	**30%**	**40%**	**10%**	**20%**	**30%**	**40%**
		Non-parametric	20.7937	78.7745	95.8730	99.3590	20.3968	67.9616	81.0886	83.5204	20.5556	75.8510	92.6955	97.6763
0.0355	Small	Parametric	26.8371	81.9368	97.2408	99.7391	41.8530	92.6645	98.9254	99.8260	34.0982	89.7072	98.7801	99.8260
		Non-parametric	26.7790	73.1226	89.7473	96.8977	26.2562	63.6996	75.2251	77.0948	27.8246	71.2960	86.6105	93.9403
	Medium	Parametric	24.4700	86.2598	98.5571	99.9374	41.6078	94.9296	99.4994	99.9687	32.7444	92.6135	99.4111	99.9687
		Non-parametric	24.3227	78.3099	94.4346	98.4355	23.4688	67.6369	80.5825	82.6141	24.0577	75.8685	91.1098	96.0889
	Large	Parametric	22.0238	89.0669	99.2857	100.000	42.6190	96.7161	99.7222	100.000	32.8968	93.8726	99.7222	100.000
		Non-parametric	23.0159	81.4577	96.6667	99.4391	22.9762	70.7249	83.2340	86.2069	22.4206	78.4541	94.2040	98.1170
0.0405	Small	Parametric	28.9573	83.7634	97.5312	99.7391	44.5832	93.6213	98.9834	99.8260	36.4217	91.0699	98.8963	99.8260
		Non-parametric	28.5216	75.0942	90.7639	97.3036	28.1441	66.2801	77.3453	78.7765	29.4511	73.2676	88.2951	94.6361
	Medium	Parametric	26.8551	88.1064	98.9105	99.9687	44.7291	95.6182	99.6172	99.9687	35.6596	93.6463	99.5289	99.9687
		Non-parametric	26.0895	80.6260	95.1708	98.6233	25.5889	69.7653	82.7596	84.3965	26.4134	77.9030	92.1696	96.6834
	Large	Parametric	25.0397	90.6688	99.3651	100.000	45.9524	97.2367	99.7619	100.000	35.7540	94.3933	99.7222	100.000
		Non-parametric	25.1984	84.1410	97.4206	99.5994	24.9206	73.3680	85.1808	87.5301	25.5556	81.2575	95.3156	98.5577
0.0455	Small	Parametric	30.9323	85.1841	97.7345	99.7970	46.4711	94.3752	99.0706	99.8260	38.6872	91.6208	99.0125	99.8550
		Non-parametric	30.3224	76.4280	91.7804	97.5935	29.7705	68.2227	78.9428	80.4871	30.6709	74.9203	89.8344	95.4769
	Medium	Parametric	28.9753	89.3271	99.1461	99.9687	47.1731	96.3067	99.6172	99.9687	38.3098	94.2410	99.6172	99.9687
		Non-parametric	27.7385	82.3161	95.8481	98.7484	27.7974	72.0501	84.4660	86.1163	28.7986	79.9061	93.1116	97.0901
	Large	Parametric	27.7381	91.5899	99.4841	100.000	48.5317	97.5170	99.8016	100.000	38.8492	95.3945	99.8016	100.000
		Non-parametric	27.5397	85.3024	97.6587	99.7196	27.5794	75.5707	86.8097	89.3745	27.9762	83.4602	95.8714	98.9984
0.0500	Small	Parametric	32.8493	86.4019	98.1412	99.8260	48.2719	94.8971	99.1868	99.8550	40.5751	92.2296	99.1577	99.9130
		Non-parametric	32.0070	78.1386	92.4775	97.8834	31.0485	69.8173	80.8888	81.8788	32.0941	76.6599	90.2992	96.1148
	Medium	Parametric	31.0071	90.4225	99.2344	99.9687	49.2344	96.7762	99.6761	99.9687	40.4888	94.9609	99.6172	99.9687

			Easy				Difficult				Random			
Alpha	**Class Size**	**Method**	**10%**	**20%**	**30%**	**40%**	**10%**	**20%**	**30%**	**40%**	**10%**	**20%**	**30%**	**40%**
		Non-parametric	29.4170	83.6620	96.4959	98.8736	29.7114	73.9906	85.9370	87.4609	30.6537	81.0955	93.8770	97.5282
	Large	Parametric	30.4365	93.0316	99.5238	100.000	50.9524	97.9175	99.8413	100.000	41.5476	95.9952	99.8016	100.000
		Non-parametric	29.6032	86.7841	98.0159	99.8397	29.4841	77.2127	88.4783	90.4170	29.8413	85.0220	96.4272	99.1987