Item Level Analysis of Wrong-To-Right Erasures

Leonardo S. Sotaridona, Arianto Wibowo & Irene Hendrawan
Measurement, Incorporated

Abstract

A statistical test is proposed that can be used to detect an item compromise due to cheating using information from wrong-to-right erasures. For most class sizes observed in practice, an (exact) statistical test based on the generalized binomial distribution has been shown applicable for most commonly used levels of significance. It is necessary, however, to set the significance level at lower values, e.g., below .001, for larger class sizes in order to maintain the error rates at the nominal level.

1. Introduction

A statistical analysis of the number of wrong-to-right (WTR) erasures in statewide assessments data is becoming customary practice as part of test security protocol adopted by state education agencies to identify possible occurrences of test fraud. The unit of analysis is often on a group of examinees, e.g., class or school. The general consensus is that once a group is flagged for suspicious test taking behaviour, follow up analyses have to be undertaken to rule out alternative explanations for the flag. Furthermore, collection of collateral information in the form of additional quantitative or qualitative analyses to substantiate or refute the allegation is necessary in order to minimize false accusations. The focus of the present paper is also on the analysis of WTR erasures but the emphasis is on the *item level* instead of examinee or group of examinees. The information obtained from item level analysis of erasures, in combination with other independent analyses, e.g., item level similarity analysis (Wibowo et al., 2013a), student level similarity analysis (van der Linden & Sotaridona, 2006; Wollack, 1997) can be used as part of substantive analyses following the group-level analysis of test irregularity. Additionally, an item that is flagged from different units, either by using the method presented in this paper or other methods, e.g., item-fit and parameter drift, may indicate that an item has been seriously compromised for future operational use. In this paper, two item level statistical tests of WTR erasures are proposed and their error rates were investigated via Monte Carlo studies with one million replications using four data sets from statewide assessment program.

2. Item Level Statistical Test of WTR Erasures

2.1 Assumptions

Let the wrong-to-right (WTR) erasures of a student *s* for an item *i* (E_{is}) be a Bernoulli random variable taking values of 0 (no WTR erasure) and 1 (WTR erasure). The probability of making a WTR erasure on item *i,* $\Pr(E_{is} = 1) = p_{is}$ for a constant p_{is}. We further assumed that E_{is} and $E_{is'}$ are independent for all s and s' in unit u where $s \neq s'$. The probability of making a WTR erasure can be estimated from the data using parametric or non-parametric approach.

2.2 Estimation of p_{is}

This paper followed a similar approach of conditioning to estimate the probability of WTR erasure on item *i* as discussed in van der Linden & Jeon (2012, see equation 3). Hence, p_{is} is estimated using a subset of non-missing final responses given erasures on incorrect options as initial responses. Let S_u denotes the number of students in this subset. The proportion of students in category *d* with WTR erasures out of S_u students who made erasures on incorrect options during their initial responses can be used as (non-parametric) estimate of p_i^d.

Alternatively, p_i^d can be estimated parametrically using (multiple) logistic regression with logit transformation (or logit link function):

$$\log\left(\frac{p_{is}}{1-p_{is}}\right) = \beta_0 + \sum_{r=1}^{R} \beta_r d_r \qquad (1)$$

where β_0 is the regression intercept and β_r is the regression slope associated with the *r*th category variable (d_r). Maximum likelihood estimate of regression coefficients in equation (1) can be obtained from most standard statistical software, e.g., SAS (2002). One can also test which category variables are statistically significant using the likelihood ratio test. The present paper differs from that of van der Linden & Jeon (2012) in two ways. First, they used IRT approach to estimate the probability of WTR erasure on item *i.* Secondly, the focus of the present paper is on flagging an item instead of student.

2.3 Statistical Tests

2.3.1. Generalized (Compound) Binomial Model

The total number of WTR in unit *u* for an item *i*

$$W_{iu} = \sum_{s=1}^{S_u} E_{is} \qquad (2)$$

has a generalized binomial distribution (Lord, 1980) with parameter $P_i = \{p_{is} \mid s = 1, \ldots, S_u\}$. For a level of significance α considered important by the analyst, item *i* in unit *u* is flagged if

$$\Pr(W_{iu} \geq w_{iu}) < \alpha \,. \qquad (3)$$

The probability at the left hand side of (3) can be computed using the following recursive formula [see also Lord & Wingersky (1984) or van der Linden & Sotaridona (2006)]

$$\Pr(W_{iu} = w_{iu}) = \begin{cases} \prod_{s=1}^{S_u}(1 - p_{is}), w_{iu} = 0 \\ \dfrac{1}{w_{iu}} \sum\limits_{k=1}^{w_{iu}} (-1)^{k-1} T(k) \Pr(W_{iu} = w_{iu} - k), w_{iu} > 0 \end{cases} \tag{4}$$

where $p_{is} \neq 1$ for all *s* and

$$T(k) = \sum_{s=1}^{S_u} \left(\frac{p_{is}}{1 - p_{is}} \right)^k . \tag{5}$$

In this paper, the probability in equation (4) was computed using the *poibin* package in R (Hong, 2012).

2.3.2. Poisson Model

As noted by previous researchers, the WTR erasures is rare event (Qualls, 2001; Bishop, Balut, & Seo, 2010; Wibowo, Sotaridona, & Hendrawan, 2013), therefore the probability of WTR erasures can be very small. In this case, the probability in (4) can be estimated using Poisson model, that is

$$\Pr(W_{iu} \geq w_{iu}) = \sum_{k=1}^{w_{iu}} \frac{e^{-\mu_{iu}} (\mu_{iu})^k}{k!}, \tag{6}$$

where $\mu_{iu} = \sum\limits_{s=1}^{S_u} p_{is}$. Alternatively, μ_{iu} can be estimated using loglinear model (see for example, Agristi, 1996). The Type I error rates of a statistical test based on (4) and that based on (6) are presented in Section 4.

3. Methods

3.1 Data

Four data sets from a statewide assessment program consisting of two grades (5 & 8) and two content areas (Math and Language Arts) were used in this study. Although the original data sets include open-ended items, only the multiple-choice (MC) portion of the tests was used in the study. The number of MC items ranges from 30 to 45 and the number of option is 4 for all data sets.

3.2. Factors, Level of Significance, and Estimation Software

The Type I error rates of the statistical tests based on equation (4) and equation (6) were investigated under four sizes of valid cases, namely, 5, 10, 30, and 100. These sizes of valid cases were chosen to show how the error rates fared across a wide range of unit or class sizes typically observed in practice. Note that the term “valid cases” in this study refers to a subset of cases as described in Section 2.2. Our emphasis on valid cases instead of the actual unit size (or class size) was motivated by the fact that for examinees in a given class or unit, e.g., 10, the actual number of valid cases is almost always less than 10 and the performance of the statistical test presented in this paper does not depend on the unit size but on the actual number of valid cases. A distribution of valid cases by class or unit size and item difficulty is presented in the Results section. Typical class sizes in the data set as reported in Wibowo et al. (2013) ranges from 10 to 59 examinees. We used significance levels in the range .01 to .05 with increments of 0.01, .0001 to .0005 with increment .0001, and .00135 to represent a level of significance based on 3 standard deviations from a one-sided statistical test that assumed standard normal distribution. Because flagging an item has less severe consequence than flagging an examinee or group of examinees, using higher values of level of significance is reasonable. SAS software version 9.1.3, SAS Institute Inc.(2002), was used to estimate the parameters of the logistic regression in equation (1). The initial parameter estimates were obtained with three category variables included in the model, namely, gender (male, female), ethnicity (black, white, hispanic, others), and proficiency level (below proficient, proficient, advanced). After considering the significance of the contribution of each category variable in the regression model and the simulation time if we are to consider different set of regressors for different items, we decided to drop both gender and ethnicity when estimating the probability of WTR erasures for all the items and only included the proficiency level as the sole regressor in the simulation studies. Maximum likelihood estimate of regression coefficients in equation (1) can be obtained from most standard statistical software, e.g., SAS (2002). In practice, one can also test which category variables are statistically significant using the likelihood ratio test. The focus of analysis is on checking the error rates of the test statistics based on equations (4) & (6) for difference sizes of valid cases and for different levels of significance.

3.3. Data Processing Steps

Select valid cases from the data set per conditioning approach discussed in Section 2.2 then estimate the regression parameters of equation (1) for all items in each grade/content. Given a subset of examinees for a certain grade/content (e.g., math grade 5):

(i) Randomly select 5 cases to represent valid cases of size 5.

(ii) Conduct a statistical test on each item using (4) and (6).

(iii) Perform (i)-(ii) for one million times and compute the empirical Type I error rates.

(iv) Repeat steps (i)-(iii) for size 10, 30 and 100.

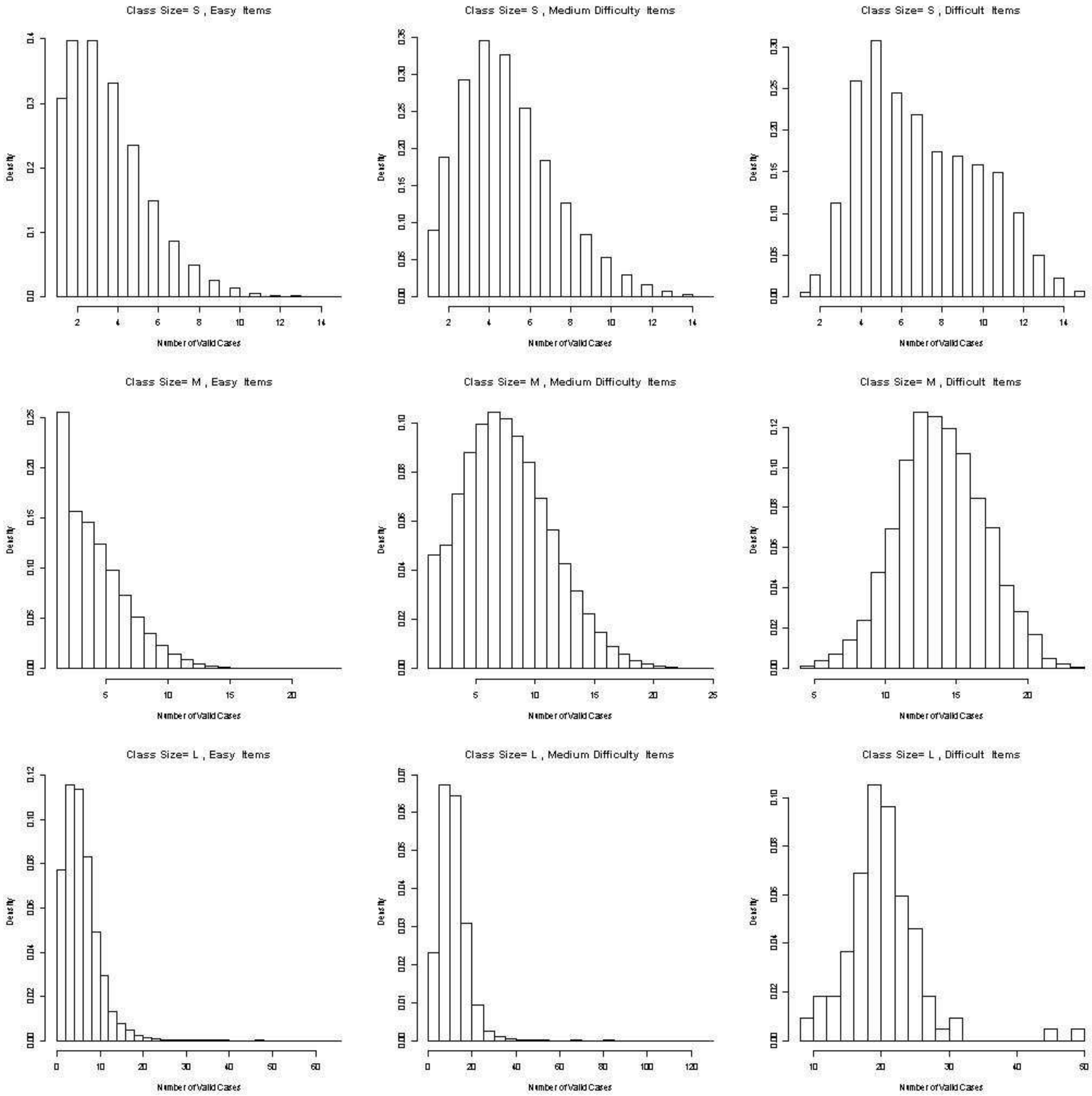
(v) Repeat (i)-(iv) for the remaining grade by content combinations.

4. Results

4.1. Number of Valid Cases by Class Size and Item Difficulty

The distributions of the number of students or valid cases from the actual data sets who made erasures on incorrect options during their initial responses and made non-missing final responses, as a function of class size and item difficulty are shown in Figure 1. In this figure, class sizes 15 or less is classified as small (S), 16 to 25 as medium size (M), and above 25 as large (L). The items with difficulty greater than .7 were classified as easy items, difficulty .3 to .7 as medium difficulty item, and difficulty below .3 as difficulty items. Clearly, the number of valid cases (or erasures) increases as the item difficulty and class size. For easy items, most valid cases are below 10 and mostly below 30 for difficult items.

Figure 1. Number of Valid Cases by Class Size and Item Difficulty



4.2. Type I Error Rates

4.2.1. By Item

Figures 2-5 shows the Type I error rates of compound binomial (circle) and poisson (triangle) models for each individual item within certain size of valid cases. The points in the graph are coordinates where the x-axis is the level of significance and the y-axis is the empirical Type I error rates. The solid line (identity line) indicates perfect agreement between the theoretical and empirical error rates (ideal scenario). Error rates below the identity line would indicate that the statistical test is conservative while those above the identity line would indicate that the statistical test is liberal. Ideally, we aimed for a statistical test with error rates that are within or are slightly conservative. Some key observations from the plots of Type I error rates:

a) For 10 and 5 valid cases, both methods were able to control the error rates within the nominal levels for all four data sets. However, the compound binomial performs better as the error rates are, for most items, closer to the identity line.

b) For 30 valid cases, the compound binomial tended to be liberal at the higher value of level of significance (.01 or higher) while the poisson model performs best. For significance levels .00135 and below, the error rates of the compound binomial are acceptable for 30 valid cases (see Appendix 7.1)

c) The error rates of the compound binomial are unreasonably large as the number of valid cases increases, e.g., for 100 valid cases, the compound binomial cannot be recommended in practice.

Figure 2. Type I Error Rates of Generalized (Compound) Binomial and Poisson, Individual Item, Math-Grade 5

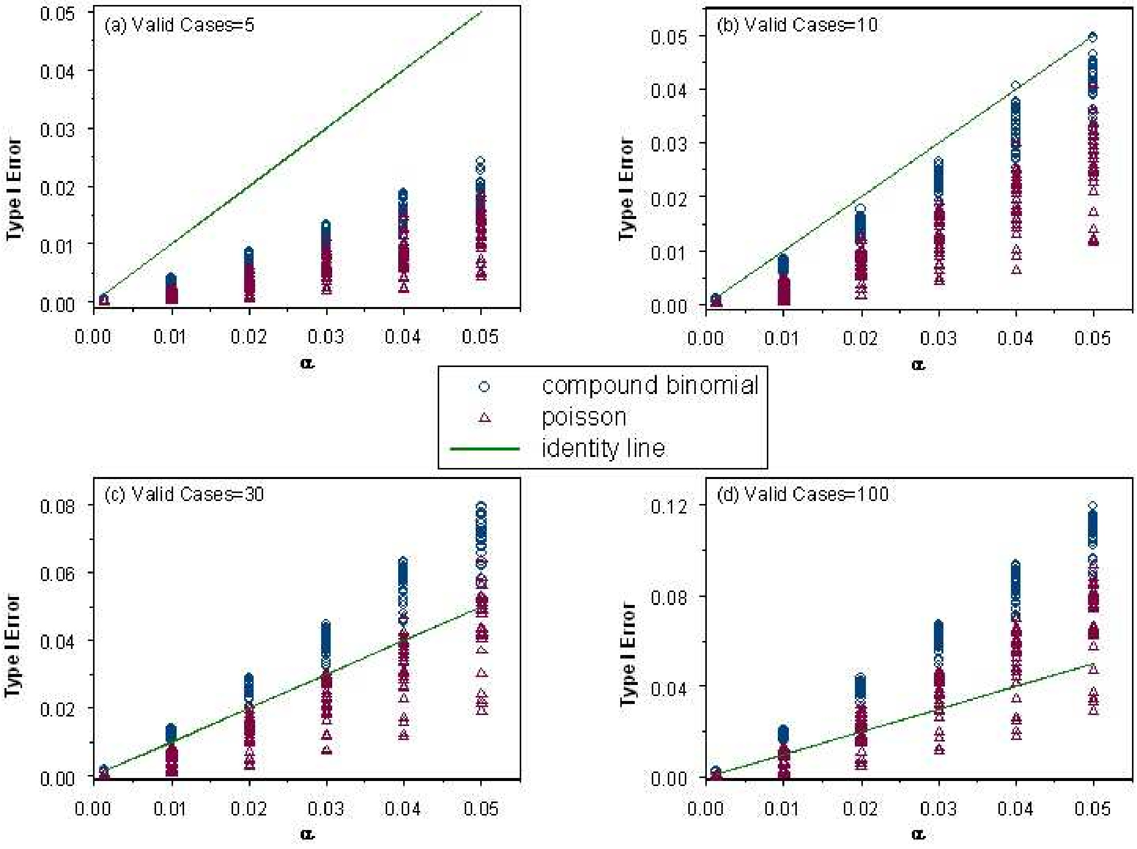


Figure 3. Type I Error Rates of Generalized (Compound) Binomial and Poisson, Individual Item, Math-Grade 8

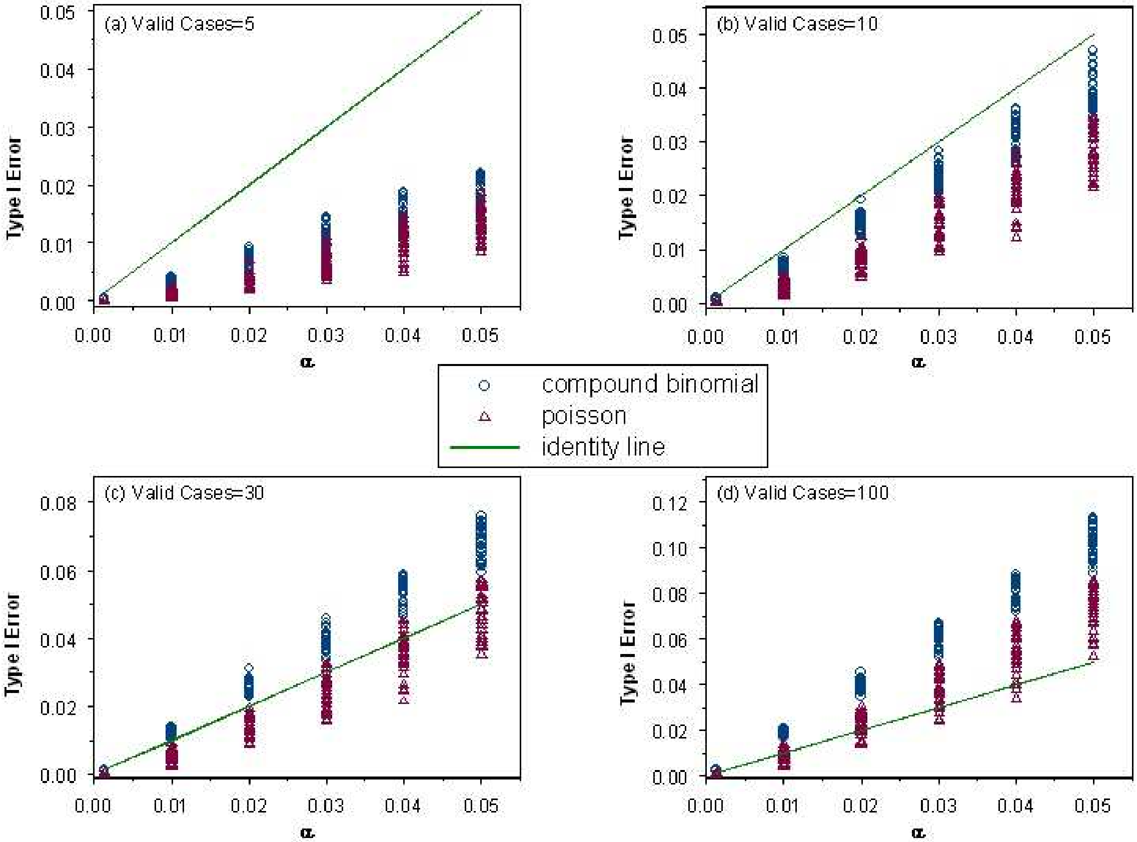


Figure 4. Type I Error Rates of Generalized (Compound) Binomial and Poisson, Individual Item, Reading-Grade 5

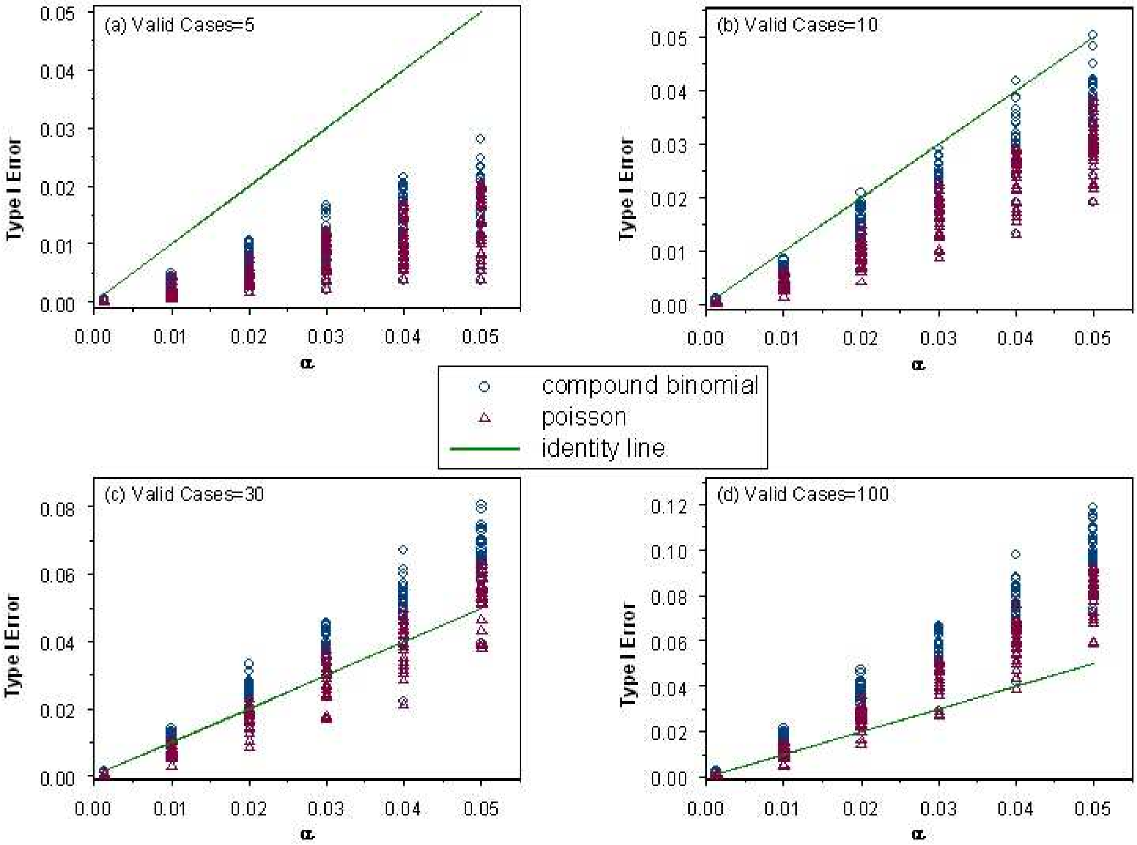
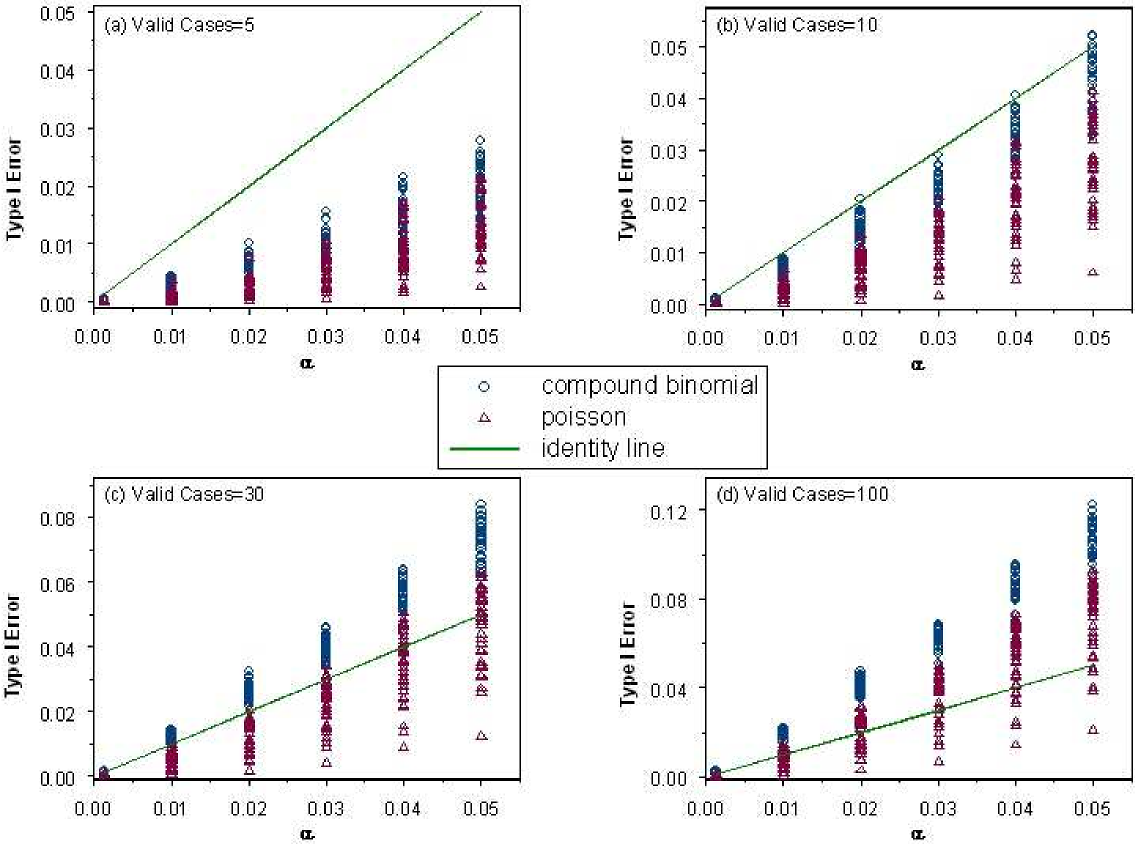


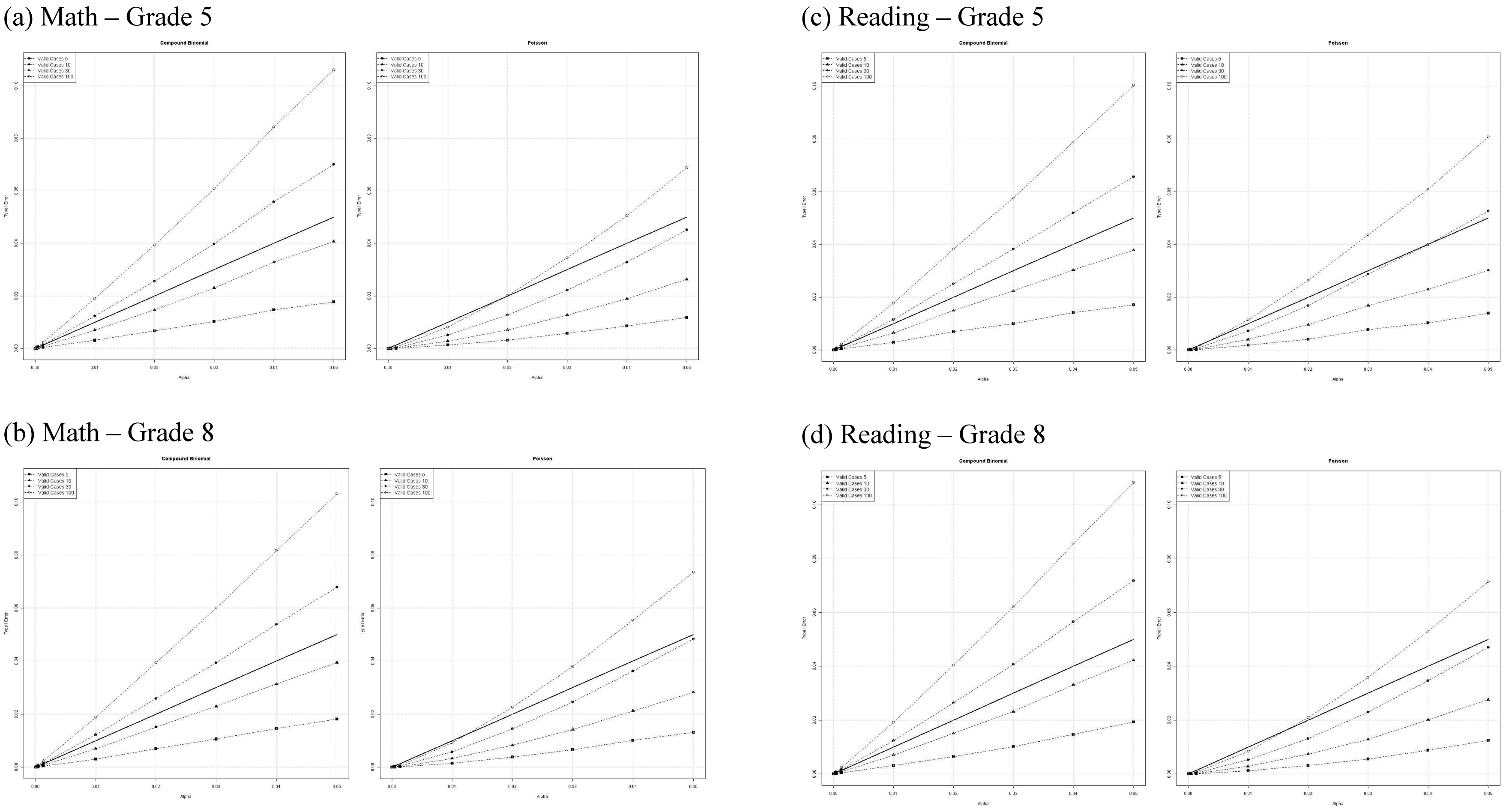
Figure 5. Type I Error Rates of Generalized (Compound) Binomial and Poisson, Individual Item, Reading-Grade 8



4.2.1. Mean Error Rates by Grade/Content

Figures 6 shows the mean Type I error rates when all items are aggregated together and for large values of level of significance (.01-.05). Consistent with earlier results, the compound binomial performs best for 10 and 5 valid cases. If one has to use the compound binomial test for 30 valid cases, one has to set the level of significance at lower values, e.g., below .001 in order to maintain the error rates close to the nominal level (see Figure 7).

Figure 6. Mean Type I Error Rates of Generalized (Compound) Binomial and Poisson



(a) Math – Grade 5

(b) Math – Grade 8

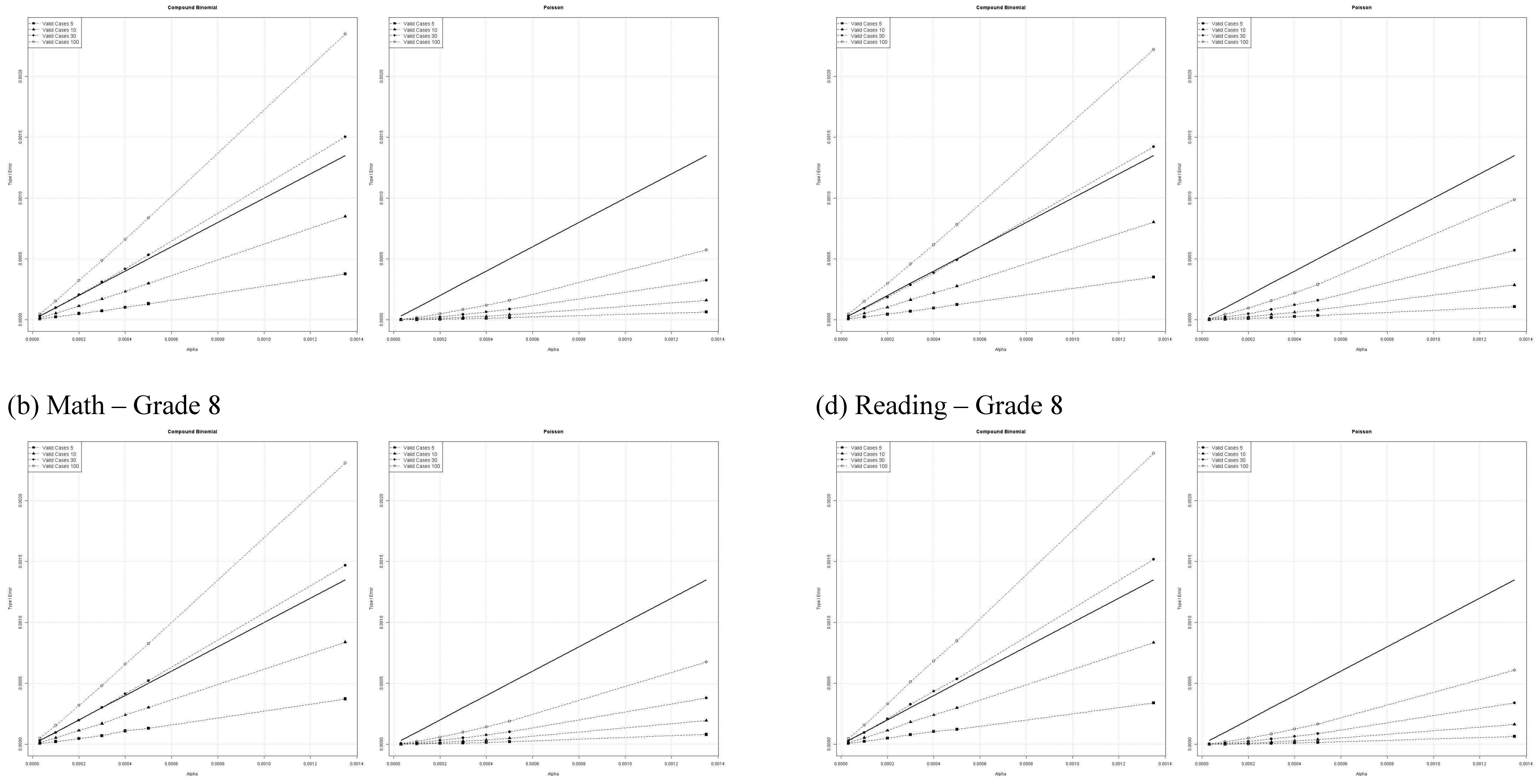
(c) Reading – Grade 5

(d) Reading – Grade 8

Figure 7. Mean Type I Error Rates of Generalized (Compound) Binomial and Poisson for Small Level of Significance

(a) Math – Grade 5

(c) Reading – Grade 5



(b) Math – Grade 8

(d) Reading – Grade 8

5. Discussions

An evaluation of the characteristics of an item is a very important aspect of test development to ensure that each item that comprised a test contributes optimally to sound measurement. Some important item indicators that are used, traditionally, by test specialist for judging the characteristics of an item includes item fit, bias, difficulty, discrimination, and information. Because of the recent prevalence of cheating on statewide assessment tests, it is important to have an item indicator that pertains to cheating, e.g., an indicator that test practitioners can use to evaluate whether or not an item has been compromised due to test irregularities. Although an item fit, to some degree, could be used to measure item compromise, it was not designed for this purpose and hence, its power to detect item misfit due to cheating is known to be very low and the error rate is high. The statistical tests presented in this paper are designed specifically to detect item compromise using information from wrong-to-right erasures. For a statistical test of item compromise using information from item response similarity, refer to the paper by Wibowo et al. (2013). It has been shown that for most class sizes seen in practice, an (exact) statistical test based on the generalized (compound) binomial distribution is applicable, particularly at .01 to .05 range of significance level. For larger class sizes, however, it is necessary to conduct the test at lower significance level, e.g., below .001, in order to maintain the conservative nature of the test. A similar approach of analysis proposed in the paper by Wibowo et al. (2013) can also be applied here where the prevalence of item compromise across classes or units are collected then use this collective information as supporting evidence whether or not to exclude an item for operational use. Another approach of analysis is to look at the number of compromised items in a class and use this information as supporting evidence that there might be class-wide incidence of cheating.

6. References

Agresti, A. (1996). An introduction to categorical data analysis. NY: Wiley.

Bishop, S., Bulut, O., & Seo, D. (2010). Modeling Erasure Behavior. *NCME.* New Orleans.

Hong, Y. (2012). On computing the distribution function for the Poisson binomial distribution. *Computational Statistics & Data Analysis*. Available online, http://dx.doi.org/10.1016/j.csda.2012.10.006

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement*, 8, 452–461.

SAS Institute Inc. (2002), SAS 9.1.3 Help and Documentation, Cary, NC: SAS Institute Inc.

Qualls, A. (2001). Can Knowledge of Erasure Behavior Be Used as an Indicator of Possible Cheating? *Educational Measurement* , 9-16.

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.

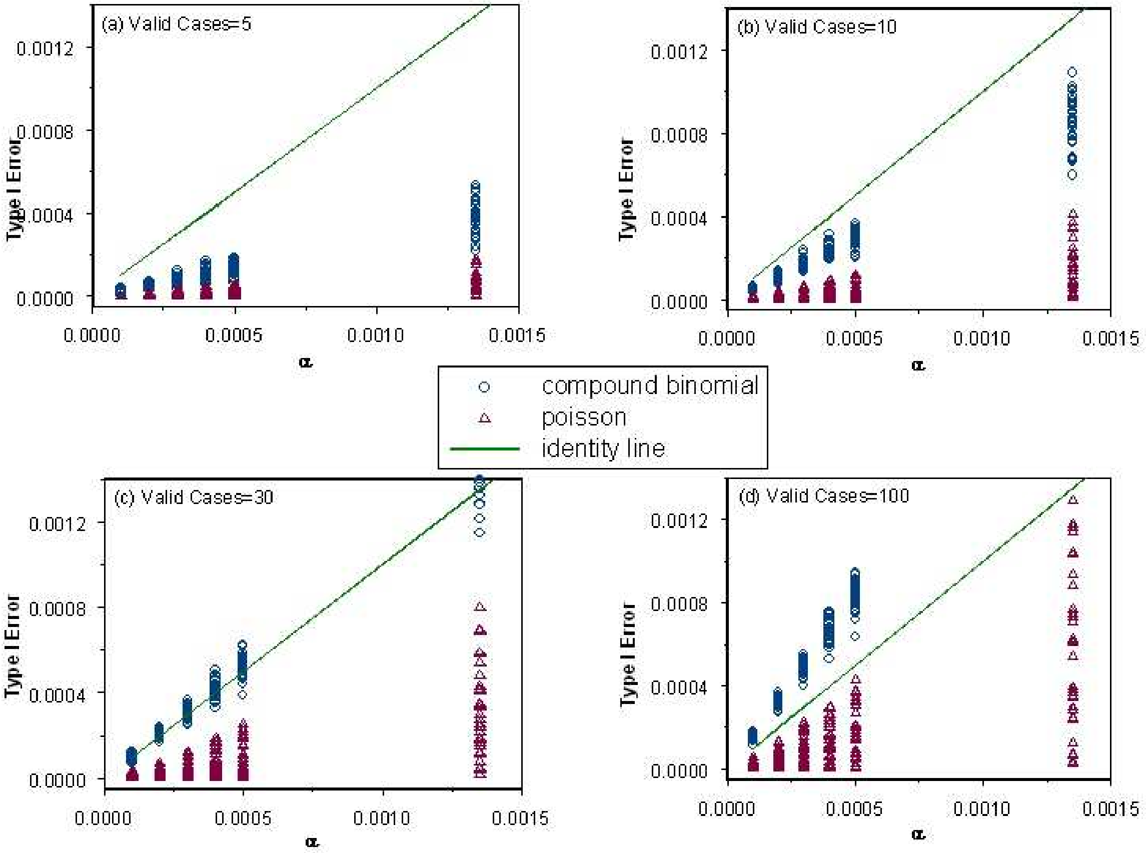
Wibowo, A., Sotaridona, L.S., Hendrawan, I. (2013a). *Statistical models for flagging unusual number of wrong-to-right*. A paper presented at Annual Meeting of the National Council on Measurement in Education, April 26-30, 2013, San Francisco, California.

Wibowo, A., Sotaridona, L.S., Hendrawan, I. (2013b). *Item level analysis of response similarity*. A paper accepted for presentation at the 2nd Annual Conference on Statistical Detection of Potential Test Fraud, October 17-19, 2013, Madison, Wisconsin.

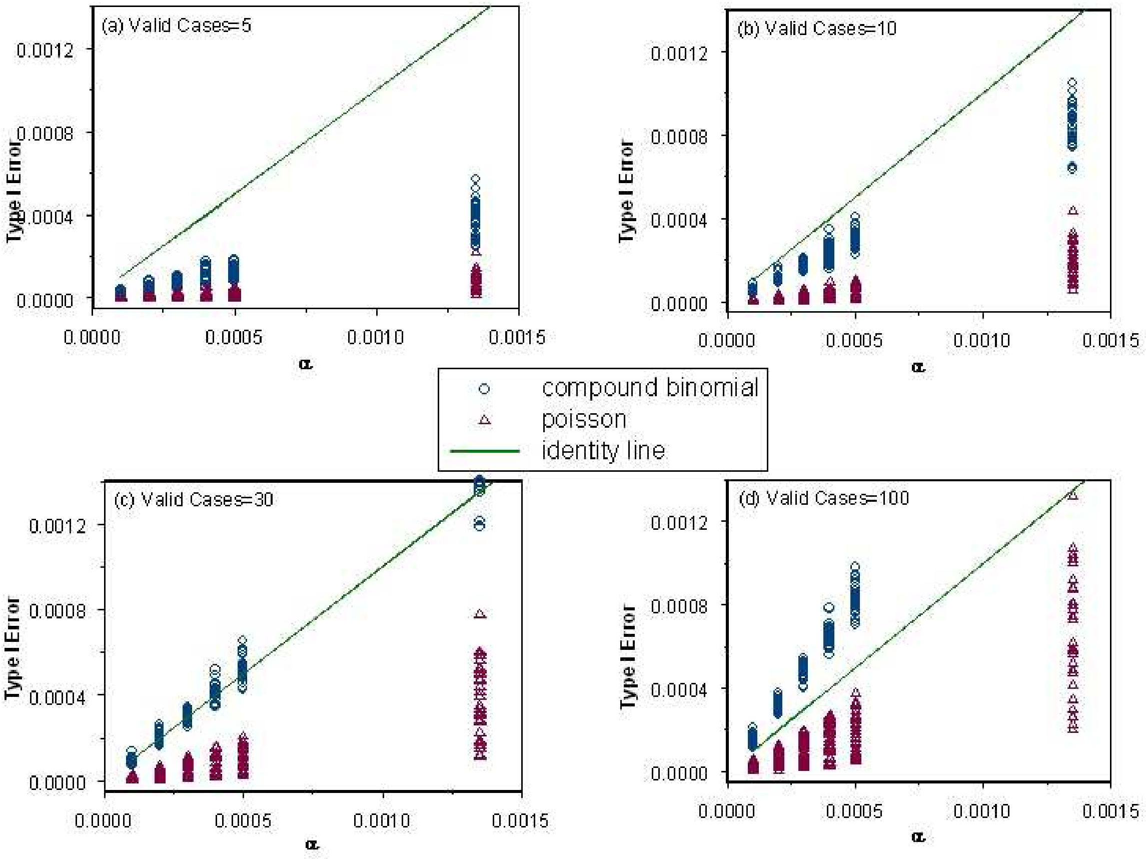
7. Appendices

7.1. Type I Error Rates for Small Level of Significance

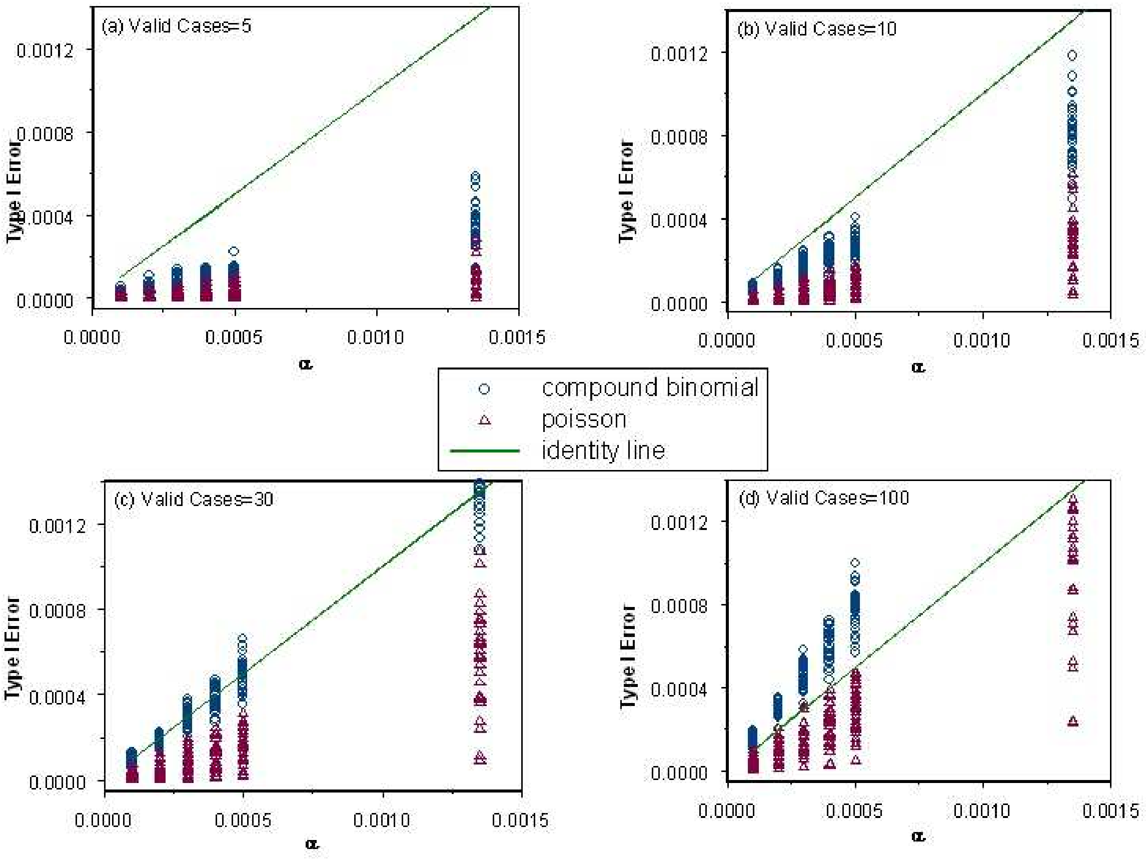
Math, Grade 5



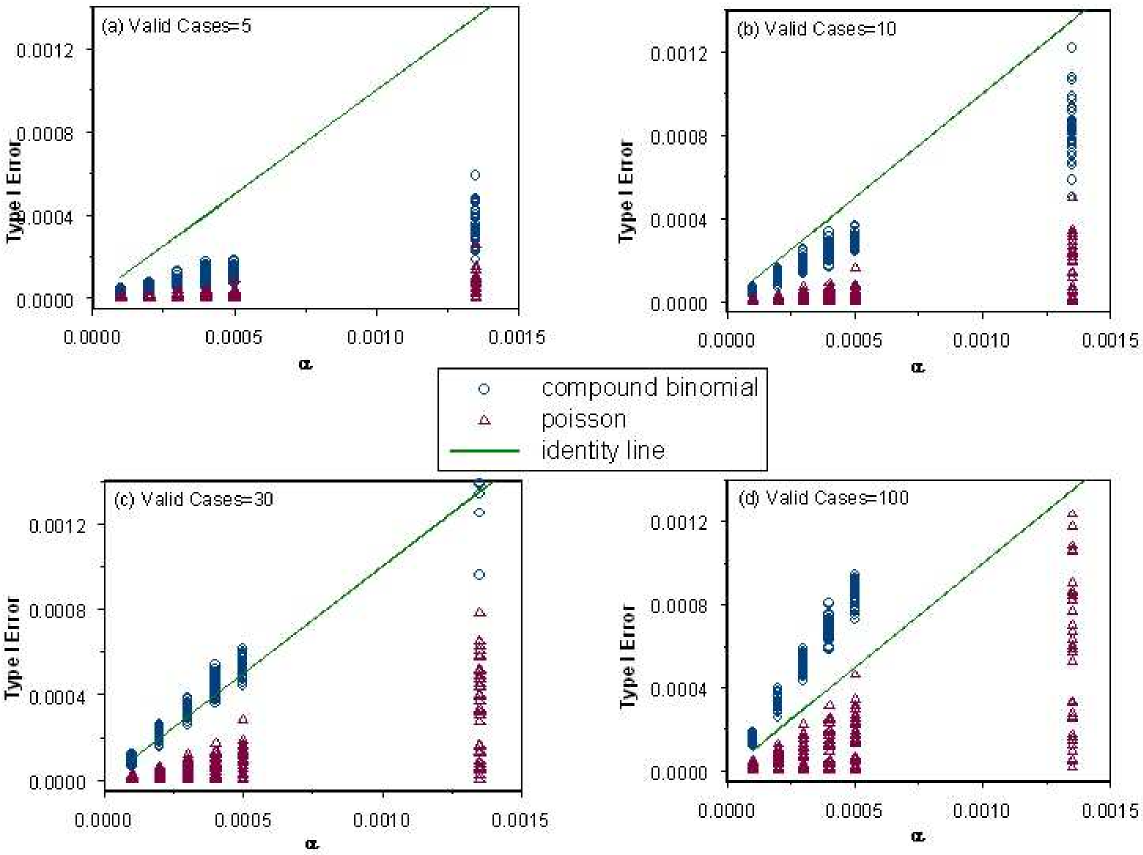
Math, Grade 8


(a) Valid Cases=5
(b) Valid Cases=10
(c) Valid Cases=30
(d) Valid Cases=100
Type I Error
0.0000
0.0004
0.0008
0.0012
α
0.0000
0.0005
0.0010
0.0015
compound binomial
poisson
identity line

Reading, Grade 5



Reading, Grade 8


(a) Valid Cases=5
(b) Valid Cases=10
(c) Valid Cases=30
(d) Valid Cases=100
Type I Error
α
compound binomial
poisson
identity line