SHU, ZHAN, Ph.D. Detecting Test Cheating Using a Deterministic, Gated Item Response Theory Model (2010)
Directed by Dr. Richard Luecht and Dr Robert Henson. 127 pp.

High-stakes tests are widely used as measurement tools to make inferences about test takers' proficiency, achievement, competence or knowledge. The stakes may be directly related to test performance, such as obtaining a high-school diploma, being granted a professional license or certificate, etc. Indirect stakes may include state accountability where test results are partially included in course grades and also tied to resource allocations for schools and school districts. Whether direct or indirect, high stakes can create an incentive for test cheating, which, in turn, severely jeopardizes the accuracy and validity of the inferences being made. Testing agencies and other stakeholders therefore endeavor to prevent or at least minimize the opportunities for test cheating by including multiple, spiraled test forms, minimizing item exposure, proctoring, and a variety of other preventive methods. However, even the best test prevention methods cannot totally eliminate cheating. For example, even if exposure is minimized, there is still some chance for a highly motivated group of examinees to collaborate to gain prior access to the exposed test items. Cheating detection methods, therefore, are developed as a complement to monitor and identify test cheating, afterward.

There is a fairly strong research base of statistical cheating detection methods. However, many existing statistical cheating detection methods are in applied settings. This dissertation proposes a novel statistical cheating detection model, called the *Deterministic, Gated Item Response Theory Model* (DGIRTM). As its name implies,

the DGIRTM uses a statistical gating mechanism to decompose observed item performance as a gated mixture of a true- proficiency function and a response function due to cheating. The gating mechanism and specific choice of parameters in the model further allow estimation of a statistical cheating effect at the level of individual examinees or groups (e.g., individual suspected of collaborating). Extensive simulation research was carried out to demonstrate the DGIRTM's characteristics and power to detect cheating. These studies rather clearly show that this new model may significantly improve our capability to sensitively detect and proactively respond to instances of test cheating.

DETECTING TEST CHEATING USING A DETERMINISTIC, GATED ITEM RESPONSE THEORY MODEL

by

Zhan Shu

A Dissertation submitted to
the Faculty of the Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirement of the Degree
Doctor of Philosophy

Greensboro
2010

Approved by

Dr Richard Luecht

Committee Co-Chair

Dr Robert Henson

Committee Co-Chair

APPROVAL PAGE

This dissertation has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Co-Chair	Dr. Richard Luecht
Committee Co-Chair	Dr. Robert Henson
Committee Members	Dr. John Willse
	Dr. Micheline Chalhoub-Deville
	Dr. Charles Lewis
	Dr. Terry Ackerman

Sep 10, 2010
Date of Acceptance by Committee

Aug 30, 2010
Date of Final Oral Examination

To My Dear Parents And Sister Who I Mostly Love In My Life

ACKNOWLEDGEMENTS

August 9th 2007 is the first day I was in United States. At that time, what I had was my young heart and two luggage. Three years later, I am now a Doctor of Educational Measurement and ready to take off to begin my new life journey at Educational Testing Industry. There were enlightening, challenging, trying and exciting moments during the last three years. What I learnt in this thrilled academic journey is not only the knowledge, but also the way to share, to give, to learn and to grow.

My advisor, Dr Richard Luecht, is the one who deserves my first appreciation. You are generous, knowledgeable and bright. You gave me the opportunity to develop and establish my expertise. Your profundity of academic and industrial knowledge leads me in a right direction, which is one of the most important reasons why I can develop, expand and solidify my dissertation.

Dr Robert Henson, my co-advisor of my academic committee, is a promising young scholar with solid background. It is a really exciting experience to work with you-you are such a wonderful researcher. Through the work with you, I build up my research skills and improve my writing ability. Thank you, Dr Henson!

I would also express my gratitude to Dr John Willse who helps me to build up my programming skills in the R statistic language. All my works are on the basis of this critical skill. Thank you, Dr Willse!

My appreciation also goes to Dr Terry Ackerman who creates such productive and effective learning environment for us, and to Dr Michelin Chalhoub-Deville who always gives me wise suggestions. Of course, Dr Charles Lewis, thank you for your kindness to serve as the outside academic member of my dissertation. I am especially

grateful to Dr Rick Morgan from ETS for your encouragement and recommendations during the dissertation and job application.

Finally, I want to present this dissertation as a gift to my parents, young sister and other family members for all the love and support I have received in my life.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

High-stakes tests are widely used in education, professional certification and licensure, employment, and other settings. Most tests are employed to make inferences about test takers' proficiency, achievement, competence or knowledge. The stakes may be direct or indirect. For example, direct high stakes testing might link test scores to a set of standards and cut scores for obtaining a high-school diploma, being granted a professional license or certificate, etc.. Other examinations such as the *Graduate Record Examination*, the *Test of English as a Second Language*, and the *ACT Assessment*, are used for college entrance and/or placement, where scores take on a very competitive flavor. Indirect stakes may include state accountability where test results are partially included in course grades but also tied to financial resource allocations for teachers, schools, and school districts. Whether direct or indirect, the high stakes can create an obvious incentive for test cheating. In turn, cheating severely jeopardizes the accuracy and validity of the inferences being made by a wide variety of stake-holders, including other students, parents, schools, policymakers, teachers, testing organizations (Cizek, 1999)

Ideally, test cheating should be prevented from happening altogether or the cheaters otherwise definitively identified so that appropriate inferences can be made which differentiate test takers on the knowledge and skills of interest, rather than access to illicitly acquired test materials and/or a propensity to cheat. Detection of test cheating is really independent of the policies that guide actions or sanctions applied to

cheaters. For example, in cases of individual cheating beyond the shadow of reasonable doubt, a sanction of disbarment from the test for a period of years may be imposed. When widespread cheating is evident, a remedial action may be to cancel all scores and require every examinee to retest.

To some extent, test cheating can be prevented by reducing the possibility for cheaters to gain prior access to the test materials, or by reducing the likelihood that they will have access to memorized materials or other sources (Luecht, 1998; Cizek, 1999). Obvious preventive measures include administrating multiple test forms with unique items, increasing the size of item bank and frequently updating items in the bank in computer-based test (CBT) environment, or otherwise obfuscating the items by any variety of scrambling strategies. Unfortunately, the implied strategy of adding test forms to reduce exposure has serious cost implications in terms of printing and logistics, and updating or otherwise increasing the size of item banks has both significant operational cost and technical psychometric calibration implications.

Test cheating can also be examined on a *post hoc* basis—that is, during or sometime after the test administration— where the goal is to detect by observational methods or various statistical indices those individuals who are likely to have cheated. For example, if a test taker is observed by a proctor to have been looking at a neighbor's answer sheet, we would have observational evidence of cheating. If the examinee performs statistically better on that section of the examination where the "cheating" was observed than on other sections of the test, we would have

corroborating statistical evidence. It needs to be recognized that cheating detection, at best, is largely a matter of building a compelling case of evidence.

Because of the practical limitations of preventive cheating methods—for example, having sufficient unique items and test forms to allow secure testing with minimum exposure over a three-month period—reliable *post hoc* test cheating detection methods are gaining popularity in the testing industry. Most of the cheating detection methods are statistical in nature. Examples include gain-score statistics where significantly large, observed score gains over testing events, testing days, or across sections of examination might serve as a "trigger" for addition investigation. The obvious limitation here is that there needs to be sufficient data to form a baseline for comparison and multiple scoring events to provide within-person comparisons. Other cheating detection methods rely on patterns of responses—especially incorrect responses (e.g., longest matched string of identical incorrect responses, incorrect response matches for item options having low conditional probabilities of selection)—where statistical flags indicate extremely unlikely selections (e.g., Frary, 1977; Cizek, 1999; Wollack, 2006). Most of these types of cheating detection indices have the limitation of requiring plausible source data for the comparisons. In cases such as computerized adaptive testing (CAT), where different examinees are administered different items, it is almost impossible to identify a plausible source, rendering almost useless most of the copying indices.

Other detection methods make rather strong assumptions which can be logically and easily violated in real settings. For example, Segall's (2002) cheating detection model assumes that every cheater correctly answers the subject items with 100 percent certainty. The reality is that even the best cheaters, especially those who otherwise do not have the requisite knowledge and skills to perform well on their own, have possibly incomplete representations of the answers. When this assumption is wrong, it can be shown to reduce the detection power of the statistic.

Still other methods are potentially flawed by methodology design considerations. For example, various person fit indices (Nering, 1997) tend to lose some statistical power to detect test cheating when a large proportion of the examinee population is involved in collaboration or other types of cheating. Also, because these types of fit indices merely suggest aberrant response patterns, proving that some illicit and perhaps even unknown cheating behavior is the primary cause of the pattern may be difficult.

Segall (2002) stated, "What is lacking from existing psychometric methods is an accurate procedure for measuring the level and severity of test-compromise from operational test data collected in on-demand testing programs."(p.165). In large part, this limitation stems from a lack of conditioning data (e.g., no plausible way of identifying potential sources of the cheating, or having insufficient connectivity in the data to isolate particular gain scores or aberrant response patterns). Most organizations involved in high-stakes testing are fairly certain that test cheating or

compromise—usually by item over-exposure and memorization—now widely occurs especially when many of the same items and test forms must be exposed to accommodate on-demand computer-based testing (CBT). Most of the exposure risks, in turn, stem from the seating capacity and scheduling limitations of CBT (Luecht, 2005; Drasgow, Luecht, and Bennett, 2006). The simple fact is that the existing statistical cheating detection methods cannot efficiently detect all but the most obvious types of test cheating or compromise under many of the practical constraints of CBT.

The *Deterministic, Gated Item Response Theory Model* (DGIRTM) proposed in this dissertation avoids the limitations of insufficient conditioning variables by using each examinee's own pattern of responses as the basis of comparison. In that sense, the DGIRTM shares something with person fit indices (e.g., Meijer, 1996; Nering, 1997). However, the DGIRTM also introduces an empirically known item exposure variable that helps logically infer a source of the aberrant responding, and additional assumptions about the directionality (i.e., cheaters tends to have a higher level of probability to correctly answer the items on which have been cheated than their original level determined by the real proficiency) of the response probability that jointly improve the detection power of the statistic. Most appealing is that that DGIRTM is capable of directly estimating the magnitude of the effect due to cheating and allows for independent estimation of the examinees' true proficiency, without the influence of cheating. This effect can be aggregated across groups or populations for

other purposes (i.e., to approximate the cost of cheating for a group of collaborators or to assess whether some exposure risk threshold has been reached for an item bank). Furthermore, the DGIRTM is designed to address test cheating or compromise caused by item over-exposure, item memorization, item preview, or internet collaboration. Given some initial research with the model's consistency and, accuracy of detection, and the apparent practical utility of information provided by this model, it seems to be a promising tool for practitioners to detect test cheating by either individuals or groups.

Specifically, the merits of the DGIRTM are addressed in this dissertation using simulation studies. The studies obviously do not "prove" the value of the model, since none of the data are real. However, the simulation studies do help answer two fundamental research questions relative to the DGIRTM. First, how accurately and efficiently can the parameters of the DGIRTM be estimated? This question involves the feasibility of postulating a rather mathematically complex "mixture" IRT model. A Markov Chain Monte Carlo (MCMC) estimator is introduced that appears to function as expected to estimate all of the parameters of the DGIRTM. Second, how does the DGIRTM perform in terms of its sensitivity[1] and specificity[2]?

Chapter 2 reviews some of the salient literature on test cheating and provides a taxonomy of some of the existing of statistical cheating detection methods. Chapter 3 includes a detailed introduction to the DGIRTM by including its philosophy and

[1] Sensitivity: the proportion of true cheaters who are correctly identified as cheaters by the model
[2] Specificity: the proportion of true non-cheaters who are correctly classified as non-cheaters by the model

origin, a detailed explanation on its parameters, the MCMC estimation algorithm and other relevant conceptual features of the model. This chapter further lays out the details of an intensive simulation study designed to illustrate the reliability and detection power of the DGIRTM under a variety of design conditions (e.g., the composition of cheating population, the number of cheaters, the number of items being cheated and the different degree of cheating on each cheated items, cheating effectiveness). Chapter 4 summarizes the results of this simulation research. Chapter 5 includes a discussion of the significance of the findings, limitations of the model and this research, and future research directions.

CHAPTER II

LITERATURE REVIEW

In this chapter, the definition, consequence and causes of test cheating are introduced, followed by a discussion on the necessity of diverse methodologies to prevent and detect test cheating. Next, three categories of existing test cheating detection methods (copying indices, person-fit indices and cheating model) are reviewed, including a discussion of the methodological design considerations, as well as the strengths and weaknesses of each method.

Understanding Test Cheating

Tests and other assessments often have evaluative purposes that create "stakes" for the individual test-taker/candidate or other consumers of the scores. Simply stated, the examination stakes for an individual or group might provide an incentive to cheat. For example, if a college scholarship is awarded for any student who achieves a particular test score, the financial incentive should be obvious. That is not to say that all students cheat, given the incentive, merely that there usually needs to be some type of incentive or motivation to cheat—even if that incentive is related to social standing among one's peers or some other less-direct motivation.

Cizek (1999) provided one of the first comprehensive treatments of test cheating. As he pointed out, educators, policymakers, schools, parents, teachers, employers or

students make their judgments, evaluations and inferences primarily based on the results derived by tests. Depending on the stakes, any of these groups could be directly or indirectly involved in and/or affect by cheating. For example, test scores have been used as an indicator of teaching quality and educational program efficiency in the federal No Child Left Behind (NCLB) program, as well as for other accountability-related purposes. Given the financial resources and potential sanctions tied to NCLB and state accountability, individual schools or entire districts are not immune from the incentive to possibly cheat in subtle or not-so-subtle ways. Nor are these constituencies immune from the effects of cheating, especially if resources are allocated or policies enacted on the basis of data that may contain sizeable numbers of test cheaters. Separate from accountability purposes, parents might decide whether to send their children to a school or college using that institutions prior, average test scores as important indicators of teaching quality or college preparedness training. Many state accountability systems also tie teacher raises or annual bonuses to test results or "growth" on state-sponsored assessments. There are extremely powerful incentives for teachers to demonstrate progress and those are the stakes. The point is that incentives or motivations to cheat, facilitate cheating, or simply ignore it, are not always limited to individual students or test takers.

Most testing companies or agencies responsible for large-scale testing programs take a great care and dedicate extensive resources to ensuring the validity and accuracy of scores, decisions, and other proficiency-based interpretations made on the

basis of test performance. When performance reflects nefarious behaviors beyond the required innate intelligence and learned knowledge and skills measured by a particular assessment, the validity of that assessment is seriously threatened. Inferences about proficiency in the presence of even partial cheating are no longer accurate. In fact, the test results and evaluation of proficiency for the individuals or groups involved in cheating are actually misleading.

The negative impact of test cheating tends to grow in proportion to two factors: (1) the extent of cheating (how many are cheating?) and (2) the severity (how much advantage is gained?). The impact of cheating may also be extensive, given the chain of events that follow. For example, suppose a small group of individuals cheat on a college entrance test and obtain test scores substantially above their true proficiency levels. A university admits those students. What is the impact? First, those students may have taken up admission slots and financial resources that should have gone to other, more deserving students. Second, if the test is indeed predictive of post-secondary school success, these same students are likely to require more remedial resources and may still fail the more strenuous courses or take longer to eventually graduate with mediocre grades, at best (unless, of course, they become career cheaters).

Although the impact of test cheating is well-recognized, it remains a serious challenge uncover all types of test cheating, much less to develop proactive and reactive techniques for combating cheating. As Cizek (1999) explained, test cheating

can involve individuals as well as groups. The most obvious types of cheating involve using some "source"—other than innate intelligence and learned knowledge or skills measured by the test—to respond to test items. Cheating sources can include stolen answer sheets acquired from others, posted test questions and answers on a website, a telephone or text conversation with an outside source during the examination, or merely looking at somebody else's responses. Cheating can also involve the acquisition and dissemination of test materials; for example, putting memorized test items from an active test form on a website, taking photos on online test questions with a mini-camera, etc.. Cheating can also occur when an outside source—other than the examinee(s)—intervenes to change responses. Individuals in the testing industry have many anecdotal stories about test administrators with some direct or indirect stakes in the examination results changing answers for certain students. It is interesting to note that, similar to popular crime scene investigation terminology, the phrase "test data forensics" is now being used to describe many of the investigative and analytical procedures used by testing agencies to explore the likelihood of cheating.

In the 21th century, test cheating has become a high-tech endeavor. Recording equipment, such as micro-recorders, cell phones, button-sized still cameras, or digital capture software are all useful for facilitating test cheating. For example, a group of test takers might use tiny still cameras to photograph entire test booklets, and then sell those test items to others who post the items and likely answers on a pay-per-view

website. Examinees who do not have access to the website (i.e., who do not pay) are unfairly disadvantaged. Modern technology has therefore extended the boundaries of cheating from inside a particular test administration site (e.g., a proctored testing center) to anyone with access to an internet connection and a credit card number. This challenge of using the internet and is the world-wide boundaries greatly complicates the identification of sources of cheating as well as locating the recipients (cheaters).

A cheater does not really have any advantage or potential gain if (s)he never sees the test questions memorized or that the person otherwise had access to. In that regard, if we have sufficiently large item banks and sufficient numbers of test forms to keep the data moving in an apparent random or near random pattern, cheaters will not have any particular advantage because the test questions acquired from a particular source have only a very tiny likelihood of being seen. Unfortunately, in reality, two factors in the testing industry tend to limit the size of the item banks and the number of test forms, increasing the exposure of test items and forms to potential cheaters (Luecht, 1996, 2005). The first factor is test administration seat capacity. Since the advent of large-scale, group testing in World War II, many educational and professional certification and licensure testing agencies have held large-scale testing events in auditoriums, classrooms, and any other available spaces. Proctoring is provided at the testing sites and, by testing on only one or two dates each year, extreme exposure is limited. In these situations, it is possible to use a smaller number of test forms and random spiraling of the forms to individuals to limit the exposure of potential cheaters

to examination materials[3]. However, given the pressures of modernization, many testing agencies are now considering and, in some cases, forces by economics and political pressures to adopt computer-based testing (CBT). CBT requires a "computer seat" for each examination and, given the limitations of most secure CBT providers, the large-scale testing events of the past are no longer possible. Instead, the tests may be offered 24-7, 365 days per year, or during discrete testing windows, ranging from a week to several months. The longer the window or testing period is, the greater will be the exposure of available test materials (Luecht, 1996, 2005).

The second factor is the cost of producing items and associated test materials. Professionally developed items reportedly cost between $300 and $1,500 each to design, pilot, calibrate, and eventually use on operational test forms (Luecht, personal communication). If, for example, we determined that we needed ten times as many items to support CBT administrations within four discrete windows each year, with uniform exposures per window, the cost of test production would correspondingly increase by a factor of ten. There are few apparent economical benefits of scaling up production. And, furthermore, generating test items is not the same as stamping out parts on a manufacturing assembly line. Test developers at most testing agencies would be hard pressed to suddenly be capable of generating ten times as many high quality items, not to mention the experimental pilots/try outs and sampling design issues related to item calibration for use in scoring (Luecht, 1996; Stocking, Ward &

[3]Agencies involved in national (U.S.) and international testing still must contend with testing across time zones and even across the international date line.

Potenza, 1996; Segall, 2002; Luecht, 2005). Due to these types of practical cost limitations, many test agencies are therefore forced to expose smaller-than-desired item pools over greater periods of time, making it possible for "examinee collaboration networks" to systematically acquire and disseminate the test questions to potential cheaters (Luecht,1996).

Testing organizations endeavor to prevent test cheating through the test development and administration process by imposing strict security regulations, training of their employees, and monitoring of people and materials using various quality control mechanisms and cross-checks. Yet, test cheating can never be totally prevented, with certainty, despite the best security and quality control. One effective way of preventing cheating is to control the environment through standardization, including limiting what the examinees can bring into the testing environment, as well as using proctors and electronic surveillance measures. Trained human proctors have been effectively used for decades to observe and monitor test takers while they are taking a test, and can provide direct testimony and evidence of cheating such as overt copying or accessing an unauthorized source. However, even the most astute proctors are limited by what they can observe. For example, most proctors would be unable to see a tiny receiver embedded in an examinee's ear canal or notice carefully executed text messaging on a small cell phone. Proctoring also depends on human judgment and is subject to all of the limitations of subjectivity. If a proctor chooses to ignore certain types of minor cheating activities, the potential cheating goes unreported. If

the proctor is overly strict in his/her interpretation, false accusations could arise, raising into question the legitimacy of the proctoring. Or, if the proctor or his/her supervisor chooses to confront suspects, other innocent students might be upset and nervous and unable to focus on their tests (Cizek, 1999).

Statistical cheating detection methods provide a more reactive way to catch cheaters after the fact. Most statistical cheating detection methods are objective, as well as being cost/ time-efficient. Three categories of statistical test cheating detection methods (copying indices, cheating detection model and person fit indices) are reviewed in the following sections, with discussions of their strengths and weaknesses.

Review on Existing Statistical Cheating Detection Methods

Angoff (1974) and Frary (1977) were two of the first researchers to investigate and publish works on the detection of test cheating. However, extensive research on cheating—especially research conducted by the major testing companies and other agencies involved in operational testing—has often not been published for the simple reason of security. That is, if statistical detection methods are published and widely disseminated, copiers may discover ways to reduce the effectiveness of those techniques. Cizek (1999) can be credited with opening up this topic to more wide-spread dialogue and research interest with the publication of his book, *Cheating on Tests: How to Do It, Detect It, and Prevent It*.

Most of the existing statistical cheating detection methods can be divided into three categories: (1) answer copying indices; (2) cheating detection models that directly estimate test cheating levels or effects; and (3) person fit indices that identify unusual response patterns, without necessarily specifying the causes. Copying indices (specifically, Wollack's ω, Holland's K index, and Sotaridona's S indices) are reviewed first, followed by Segall's cheating detection model. Finally, some typical person fit indices are discussed (e.g., lz index and ECI indices).

Statistical Answer Copying Indices

Test answer copying is one of the most obvious types of test cheating, where one examinee copies the answers from one or more other examinees. The source of the copying may or may not be a willing participant in the cheating. Most answer copying indices are designed to identify test cheaters based on some policy-based thresholds related to the statistical significance level or likelihood of the similarities between the response patterns of two examinees under suspicion (cheater and source). Given a very unlikely-by-chance level of similarity between two examinees' response patterns, the plausible inference is that one of two examinees copied the answers of the other examinee. There is a plethora of answer copying indices from which to choose (e.g., Bellezza & Bellezza, 1989; Frary, Tideman, &Watts, 1977; Hanson, Harris, & Brennan, 1987; Holland, 1996; Watson, Iwamoto, , Nungester, & Luecht, 1998; Sotaridona & Meijer, 2002, 2003;Wollack, 1997, 2003;Wollack & Cohen,1998;

Wollack, Cohen, & Serlin, 2001), and most of those indexes have been demonstrated to have differential effectiveness, given the nature of the data used and assumptions made about the nature of the copying (Wollack, 2006).

Cizek (1999) classified answer copying indices as two types: (1) methods that use a theoretical distribution to compute the likelihood of observing the similarity by chance, called Type I copying indices; and (2) methods that use one or more reference samples of examinees taking the same test to estimate the likelihoods of observing particular similarities or response patterns, called Type II copying indices[4]. It should be noted that most of the detection methods shown treat incorrect answers on multiple-choice questions as equally likely (i.e., all are scored wrong and the scored dichotomous data is sufficient for the detection analysis). There are exceptions (see, for example, Angoff, 1974; Watson, Iwamoto, Nungester, & Luecht, 1998), that condition on the actual incorrect choices selected. However, nothing germane to the present research study is gained by reviewing those more elaborate methods of analysis. They are mentioned only in the spirit of completeness.

As for Type I Answer Copying Indices, generally speaking, any computed similarity index or indicator— such as counts of similarities or the length of strings of identical incorrect responses—can be viewed as having a particular sampling distribution. The probability of observing the computed statistic of interest is estimated by using a theoretical distribution such as the unit-normal cumulative

[4] These labels should not be confused with Type I and II statistic errors of inference.

z-distribution, a *t*-distribution, or some other plausible sampling distribution. Comparing the observed (computed) statistic to its theoretical sampling distribution constitutes a significance test of the null hypothesis that the observed similarity between the potential copier and the source is strictly due to chance. Rejection of that hypothesis suggests an alternative, plausible hypothesis that the similarity between the two examinees' response pattern could only have occurred due to one examinee copying from the other. The nominal significance level is typically set at a very conservative value (i.e., 1.0E-7) to minimize false-positives. Adjustments of experimental decisions errors can also be used if the number of hypothesis tests is large.

Ultimately, those examinees with significant results are identified as potential copiers (who copy answers from others) or sources (whose answers are copied by the copiers). Wollack's ω (1997) and Frary's g_2 index (1977) are the two examples of Type I answer copying indices reviewed below. (See Wollack [1997] for a more comprehensive review of answer copying indices.)

Wollack (1997) characterizes answer copying as involving two types of examinees: (1) the "source", represented by *s*, whose answers are copied by others; and (2) "copier", represented by the index *c*, who copies the source's answers. Wollack's notation h_{cs} is used to represent the number of matching item responses between the copier and the sources.

Wollack's cheating index uses item response theory (IRT), where both the test items and examinees are calibrated to a common metric, using a suitable IRT model for dichotomous items. The conditional expectation of h_{cs} is the sum of conditional probabilities, using the IRT model probabilities, $p_i \equiv \text{Prob}(u_i=1|\theta)$ and $q_i \equiv 1-p_i$:

$$E(h_{cs} \mid \varepsilon, \theta_c, U_s) = \sum_i^k p_i^{u_i} * q_i^{1-u_i} \quad (1)$$

where: ε is a vector representing the IRT item parameter estimates; θ_c is the copier's latent ability or proficiency score; U_s is the source's response pattern on $i=1,\ldots,k$ items; u_i is the source's response an individual item, i; p_i is the probability of a correct answer to the i^{th} item given the copier's ability; q_i is the probability of incorrect answer to the i^{th} item—i.e., the complement of p_i. Using the binomial distribution, Wollack provides the conditional variance of the estimates about the expected value, h_{cs}, as

$$V_{h_{cs}} = \sum_i^k p_i * q_i \quad (2)$$

The answer copying index, Wollack's ω, is therefore a simple ratio of observed deviations about the expected value to the standard error of estimate:

$$\omega = \frac{h_{cs} - E(h_{cs} \mid \varepsilon, \theta_c, U_s)}{\sqrt{V_{h_{cs}}}} \quad . \quad (3)$$

The null hypothesis for the significance test assumes that ω follows a unit-normal distribution (i.e., a Gaussian distribution). If the value of ω of an examinee is greater than or equal to the critical unit-normal value at the nominal

significance level, the copier can be identified as a potential cheater, otherwise, the examinee is not flagged. For example, given a nominal significance level of 0.001, the corresponding inverse normal critical value would be 3.09 (using absolute value). Any copiers have computed values $\omega \geq |3.09|$ would be flagged as potential copiers, requiring additional follow-up.

Frary's (1977) g_2 is conceptually similar to Wollack's IRT-based index, but only requires classical test theory statistic such as number-correct scores and item-level *p*-values. Frary's method also classifies one examinee as the source (represented by *s*) and the other one is the copier (represented by *c*). Here, h_{cs} is the observed number of shared, identical responses between the source and copier. The expected value of h_{cs} is defined as:

$$E(h_{cs}\,|U_s) = \sum_i^k P_c(U_{is}) \quad (4)$$

where: *k* is the number of items; U_{is} is the source's response to the i^{th} item; $P_c(U_{is})$ is the copiers' response probability to have the same response as the source on the i^{th} item. Frary (1977 provides extensions of the index that allow other features of the response pattern to be used (e.g., incorrect strings) calculate $P_c(U_{is})$. The variance of the $E(h_{cs}\,|U_s)$ is defined as:

$$V_{h_{cs}} = \sum_i^k [P_c(U_{is})] * [(1 - P_c(U_{is})] \quad . \quad (5)$$

The test statistic is computed as

$$g_2 = \frac{h_{cs} - E\left(h_{cs}\,|U_s\right)}{\sqrt{V_{h_{cs}}}} \quad (6)$$

Like Wollack's ω, Frary's g_2 is assumed to be normally distributed with mean of zero and standard deviation one. The null hypothesis of this statistic is: H_0: $g_2=\mu=0$. If g_2 exceeds absolute value of the critical z-value at the nominal significance threshold, the subject would be flagged for additional follow-up.

The ω and g_2 indices are highly similar, but employ a different statistical mechanism to define the expected coincidence of responses. As noted above, the ω index is IRT-based while the g_2 index is classical test theory-based. The advantage of these two indices is that they both make use of the full [scored] response vectors including both the correct and incorrect answers. However, both indices are somewhat limited in accuracy, for more extreme scores and lose substantial statistical power as the copier and source have greater numbers of shared responses. With the ω index, the copiers' estimated proficiency will be positively or negatively inflated by including the copied responses in the full response string. A substantial increase in the number of test cheaters (copiers in the population) could increase the magnitude of the scale difference between the estimated examinees' scores (with cheating) and examinees' real proficiency scores, as measured by θ. Consequently, the expected coincidence (similarity) of responses would be contaminated and drift away from the true value, resulting in a non-centrality shift in the sampling distribution of ω. The critical implication is that computed ω statistic would no longer be normally distributed with mean of zero and standard deviation one. If we could estimate the non-centrality shift, a corrected sampling distribution could be used; however, a

plausible statistical non-centrality adjustment method has not been offered in the literature.

Frary's g_2 is based on the observed data to calculate the expected number of coincident responses. This index itself will be greatly impacted by the population and item characteristics (classical item difficulties and discrimination indices). A fundamental limitation is that, similar to Wollack's index, the g_2 statistic does not distinguish an examinee's true ability from that ability plus some incremental gain due to cheating, and is therefore susceptible to the same non-centrality shift as ω, when the number of copiers increases substantially.

In summary, Type I answer copying indices such as Wollack's ω and Frary's g_2 have the advantage of making use of all of the responses (both incorrect and correct responses). As discussed below, broadly considering all the responses may be an advantage when compared to the Type II indices. However, the Type I copying indices tend to lose power and can lead to biased results when the magnitude of the cheating effectiveness or the number of cheaters in the population increases because these methods do not distinguish any examinee's true proficiency and associated responses from the portion of the response affected by cheating. Furthermore, the assumption that the counts of coincident responses due to cheating or other forms of collaboration are normally distributed may be violated, rendering the use of a Gaussian probability distribution inappropriate.

As opposed to Type I Copying Indices, Holland's *K* index (1996) and Sotaridona's *S* index (2002)， are representative examples of the Type II copying indices. Both indexes assume that the similarity between the response patterns of copiers and sources can be modeled by using either the *Binominal* or *Poisson* distributions.

Holland' K could be implemented by either using an empirical sampling distribution or a theoretical distribution. For both implementations, the examinees are divided into some number of groups, denoted *R*, based on the number of wrong answers. Each group has the same number of wrong answer which is labeled as W_r. For example, all of the examinees having only one incorrect answer form Group 1, all of the examinees having two incorrect answers form Group 2, etc.. As with the Type I indices, there is a source student, labeled by *s*, whose answers may have been copied. The label r_j is used to represent that the j^{th} examinee from the r^{th} group.

For the empirical implementation, K_j is defined as the observed proportion of examinees in one subgroup (R_c) who having the same number of incorrect answer as the j^{th} examinee and the examinees in the subgroup whose matching number of incorrect answer with the source is at least as large as the j^{th} examinee (Sotaridona, 2005). Specifically, K_j is defined as

$$K_j = \frac{\sum_{i=1}^{n} I_{ij}}{n} \quad , \qquad (7)$$

where *n* is the total number of examinees in the sub-group (R_c) that have the same

number of incorrect answers with the j^{th} examinee, and i represents the i^{th} examinee in the sub-group. The variable I_{ij} is a binary indicator that is set to zero when $M_{is}<M_{js}$, or I_{ij}= 1when $M_i>=M_j$. where M_{is} is the matching number of incorrect answer of i^{th} examinee in that sub-group with the source. M_{js} is the matching number of incorrect answer of j^{th} examinee with the source.

A small value of K_j means that the j^{th} examinee may have taken the opportunity to cheat in the test by copying strings of responses. That is, the logical directionality of the test links possible copying with decreasing values of K_j. However, the count of matching incorrect items is sensitive to the ability level of both the copier and source. For example, if the copier and source are at the same ability level, they would be expected to have a relatively high number of identical matching number of incorrect answers. As a result, more innocent test takers at similar ability levels might be misclassified as copiers using this index. In addition, this method is sample size dependent because the number of examinees in a particular subgroup is involved in this index, especially for the empirical implementation.

For the theoretical implementation, the probability of the suspect having at least as large matching number of incorrect answers with the source as other examinees in the same subgroup is modeled using the *binominal distribution*. The number of matching wrong answer between the examinee r_j and the source (s) is labeled as M_{rj}. Thus the K defined by Holland (1996) is:

$$\mathrm{K_j} = \sum_{w=M_{rj}}^{W_s} \binom{w_s}{w} p^W * (1-p)^{W_s - W} \quad (8)$$

where p is the expected probability of wrong answer matching, W_s is the number of wrong answers of the source. Note that p can be estimated in different ways. One of the more direct methods is to define p as $\frac{\overline{M_r}}{W_s}$ where $\overline{M}_r$ is the mean number of matched wrong answers between all the examinees from the r^{th} group and the source. Holland also suggested a linear regression method of approximating p. This regression approach uses the proportion of wrong answers (Q_r) of each examinee in each subgroup as the predictor variable. Holland (1996) provided an example to show how the p was predicted using linear regression:

$$p = \begin{cases} \mathrm{a} + \mathrm{bQ_r}, & 0 < Q_\mathrm{r} < 0.3 \\ \mathrm{a} + 0.3\mathrm{b} + 0.4\mathrm{b} * (\mathrm{Q_r} - 0.3), & 0.3 < Q_\mathrm{r} < 1 \end{cases}. \quad (9)$$

As a practical note, a and b must be pre-specified or otherwise estimated applying the conditional two-part regression models, based on Q_r. For example, Holland (1996) set a=0.085 and b is differently set based on specific testing settings. How these coefficients a and b were estimated for different testing contexts is not clearly presented in Holland's study (Sotaridona, 2003).

Modified K indices ($\overline{K}_1$ and $\overline{K}_2$) were proposed by using all of the data from all the sub-groups to predict the p values using linear or polynomial regression models. The estimation methods of p in these two new indices is not the focus of this dissertation, thus readers could search more information about this by referring to Holland's work.

Sotaridona (2002) similarly proposed two new indices called S_1 and S_2 to describe the probability of the suspect having at least as large matching number of incorrect answers as other examinee in the same subgroup, based on an assumed *Poisson* probability distribution. Examinees are also divided into *R* subgroups based on the number of incorrect answers, where examinees in each subgroup have the same number of wrong answers. The formula for calculating S_1 is defined as:

$$S_1 = \sum_{W=M_{rj}}^{W_s} \frac{e^{-\mu_r} * {\mu_r}^W}{W!} \quad (10)$$

Where $\mu_r = e^{\beta_0 + \beta_1 * W_r}$, and β_0 and β_1 is estimated based on the number of wrong answers and the mean number of matching wrong answers for each group (W_r). W_s is the number of wrong answer of the source (*s)*. M_{rj} is the matching number of wrong answers between the j^{th} examinee in the r^{th} subgroup and the source.

Notably, both Sotaridona's S_1 index and Holland's *K*, only use the incorrect answers, disregarding with correct answers. In contrast, the S_2 index, uses both correct answers and incorrect answers. It seems that S_2 should be a promising copying index. However, it introduces a lot of other unverified transformation and factors which would jeopardize its reliability. For example, this index is hard to distinguish examinees who are competent enough to correctly answer to a certain item from those who correctly respond to the item by guessing or cheating. Thus a complicated mathematic transformation formula, which incorporates the guessing level of each item, is introduced to identify cheating effect embedded in the correct responses, by

removing the impact of guessing on the response patterns. However, such transformation formula is not theoretically or empirically verified so far, which is somewhat speculative. For more detailed information about this index, reader could refer to Stataridona's work in 2003.

In summary, the basic logic underlying these Type II indices is that a suspect who has a larger number of matching incorrect answer than most of other examinees in the same sub-group should be the one who copied answers from the source. A small value of both *S* and *K* indicates the high likelihood of answer copying. These two indices could be problematic sometimes. Firstly, it is hard to calculate the expected matching probability (p) for *K* and expected matching number (μ) for S_1. Actually, some new methods (e.g.,$\overline{K}_1$,$\overline{K}_2$) are proposed to calculate the *p* and the μ; however, they are still unreliable and difficult to be estimated. Secondly, the *Binominal* or *Poisson distribution* is not theoretically validated to characterize the likelihood of the incorrect answer match between the suspect and source. Next, the two kinds of indices are easily impacted by the item difficulty, examinees' ability and sample size of each subgroup. As a notice, these type II copying indices, except S_2, only utilize the information of the matched incorrect answers, but they are preferred by practitioners in real settings because they seems to have higher level of power to detect test cheating relative copying.

A number of other Type II cheating indices could have been included in this section (e.g., g_1, $\overline{K}_1$,$\overline{K}_2$,S_2), but most of them make the same relative assumptions and

focus on essentially the same data as the indices reviewed here. The copying indices are particularly developed to identify test cheating caused by direct answer copying. Although many copying indices have been demonstrated to be useful tools to detect test copying, they are limited and may be unreliable in many applied settings because of their strong assumptions on the distribution of the similarity statistics. Wollack (2006) pointed out that no one existing copying index was uniformly better than others given different amounts of or different types of answer copying.

Cheating Detection Models

Some researchers have chosen to focus on developing statistic models for more directly characterizing the severity and level of test cheating. Segall (2002) proposed an IRT-based cheating detection model to directly provide estimates of the population characteristics related to the test cheating. This model is particularly designed to detect test cheating caused by item exposure, item-preview, or item sharing via internet. The use of such models is obvious. If cheating becomes sufficiently problematic, change the item bank or otherwise implement changes to significantly reduce the benefits of prior exposure to the active item bank and test forms.

In Segall's (2002) model, the test items are separated into two mutually exclusive categories: one class of items which are *probably* compromised, as indicated by empirical exposure counts, time in use, or other indicators; the second class of items are considered to be secure due to recent release or other factors. Segall (2002)

defined the first class of items as those considered exposed to test takers beyond a span of a week, month, or a year, depending on empirical investigations or policy decisions.

Using an IRT-based response function, the probability of a correct answer to an item under Segall's test compromise model is:

$$P_{ij}\left(U_{ij}\middle|\theta_j, a_i, b_i, c_i, K_{ij}\right) = 1 + \left(1 - K_{ij}\right) * [P\left(U_{ij}\middle|\theta_j, a_i, b_i, c_i\right) - 1] \quad (11)$$

$$K_{ij} \sim Bernolli\ (P_{ij}{}^{\omega}) \quad (12)$$

$$P_{ij}{}^{\omega} = \emptyset_i * \Phi(\alpha_i + \beta_i * \omega_j) \quad (13)$$

$$\mathrm{P}\left(U_{ij}\middle|\theta_j, a_i, b_i, c_i\right) = c_i + (1 - c_i) * \Phi[a_i * (\theta_j - b_i)] \quad (14)$$

where K_{ij} is the cheating status of the j^{th} examinee on the i^{th} item (where k_{ij}= 1 denotes prior exposure of that item to that particular examinee and k_{ij}= 0 denotes lack of exposure). The j^{th} examinee's cheating tendency $P_{ij}{}^{\omega}$ is hierarchically modeled by a normal ogive model with a person parameter ω_j and item parameters α_i and β_i. The parameter $\emptyset_{\mathrm{i}}$ for the i^{th} item is assumed to be known such that $\emptyset_{\mathrm{i}}$ = 1 for exposed items, $\emptyset_{\mathrm{i}}$ = 0 for unexposed items. Equation 13 is the usual three-parameter normal-ogive IRT model for the probability of a correct response of the j^{th} examinee to the i^{th} item, with examinee ability θ_j and item parameters a_i, b_i and c_i.

Compared with copying indices, Segall's model has two distinctive advantages, from an explanatory perspective. One advantage is that, in addition to modeling the

test takers' proficiencies and item operating characteristics under non-cheating conditions, it includes population-level characteristics related to test compromise. This provides actionable information to signal a testing agency of potentially high risks of cheating and suggest the need to introduce a new item bank or less-exposed test forms. The other advantage is the model makes use of the full [scored] response matrix, including all examinees. In contrast, most of the answer copying indices compare response similarities of only pairs of examinees: the source and copier suspect.

Segall's cheating detection model does make one rather strong assumption that may limit its flexibility and, possibly, its statistical power to correctly classify examinees and estimate the magnitude of the cheating in the population. That assumption is that the test cheater will correctly answer every item on which have been cheated with 100 percent certainty. It is certainly reasonable to believe that test cheating activities should increase test cheaters' observed performance—that is, the probability of correctly answering most of the items to which they had prior access. But, it seems highly unlikely that every examinee will memorize every exposed item with certainty and be able to retrieve those answers while taking the test. If we suppose that many of the cheaters are motivated to cheat because they lack the knowledge and skills to answer the questions by honest learning/studying or innate intelligence, then it makes little sense to assume that those same cheaters can memorize and successfully recall large blocks of exposed items, at least not without

auxiliary aids such as hidden "cheat sheets", access to cell/text messaging, or other nefarious means of getting outside help. Insofar as the model is concerned, the probability of correctly answering the cheated items is not determined by the cheating activities itself. Other factors such as each cheaters' real proficiency level and his/her capability to memorize items at different levels of difficulty, given different content and surface features of the items, are also key factors to impact a cheater's probability of correctly answering the items which have been cheated by the cheater.

As a point of interest, the Deterministic, Gated IRT Model (DGIRTM) presented in this dissertation attempts to overcome that assumptive limitation by relaxing the assumption that cheating examinees response perfectly to every item that may have been previously exposed. The DGIRTM also more generally deals with the issue of estimating the impact of cheating (gain scores) in the population than is possible using Segall's model.

Person Fit Indices

An extensive number of person-fit indices (PFI) have been developed, primarily to evaluate the date-model fit of various IRT models. Examples include likelihood ratio tests such as l_o (Levine & Rubin, 1979), l_z (Drasgow, Levine,& Williams, 1985), and $ECI1_z$, $ECI2_z$, $ECI4_z$, and $ECI6_z$ (Tatsuoka, 1984; see Meijer, 1996). Only the l_z and $ECI4_z$ are reviewed in this dissertation, because prior research rather conclusively suggests that these two indices are the two of most useful tools.

It seems worth noting that person-fit indices are intended to detect aberrant patterns of responses that differ from expected responses under a particular theoretical model. Assuming that the model holds for the majority of examinees in a particular population, those examinees having highly unusual patterns can be flagged and further evaluated for likely causes of the aberrance. For example, consider an examinee getting sick during the afternoon of a full-day examination, making it difficulty for her to concentrate. Her item responses are likely to suffer and may more progressively mimic random responses toward the end of the test. The person-fit statistics should be capable of flagging the rather dramatic changes in the patterns of response, earlier versus later in the test. This example has an obvious cause that seems consistent with the data. Suppose that another examinee has exactly the same response patterns, but his performance dropped off because he no longer was able to cheat—perhaps he lost or inadvertently forgot to bring the final page of answers. Did he cheat and does the statistically determined aberrant response pattern constitute sufficient evidence to come to that conclusion? This simple example characterizes the challenge of most person-fit indices for test cheating. The power to detect aberrant response patterns is usually independent of possible causes and alternative explanations for exactly the same data with different causes.

The l_z index is a standardized ratio function of a log-likelihood function, where the likelihood of the observed responses are compared to an expected value for the population or an appropriate reference group. The l_z statistic can be expressed as

$$l_z = \frac{\ln[L(\hat{\theta})] - E\{\ln[L(\hat{\theta})]\}}{\sqrt{\text{var}\{\ln[L(\hat{\theta})]\}}} \quad (15)$$

where the log-likelihood of the observed data is given as

$$\ln[L(\hat{\theta})] = \sum_{i=1}^{n}\{u_i * \ln[P_i(\hat{\theta})] + (1 - u_i) * \ln[1 - P_i(\hat{\theta})]\} \quad (16)$$

and with an expectation and variance defined as

$$E\{\ln[L(\hat{\theta})] = \sum_{i=1}^{n}\{P_i(\hat{\theta}) * \ln[P_i(\hat{\theta})] + \left(1 - P_i(\hat{\theta})\right) * \ln[1 - P_i(\hat{\theta})]\} \quad (17)$$

and

$$\text{var}\{\ln[L(\hat{\theta})] = \sum_{i=1}^{n}\{P_i(\hat{\theta}) * \left(1 - P_i(\hat{\theta})\right) * \{\frac{\ln[P_i(\hat{\theta})]}{1-P_i(\hat{\theta})}\}^2\} \ . \quad (18)$$

Referring to Equations 15 to 18, *i* represents the items, $\hat{\theta}$ is the ability estimate (assumed to be estimated by maximum likelihood), u_i is the response of the i^{th} item, and $P_i(\hat{\theta})$ is the probability of a correct item response for a given estimated of ability, $\hat{\theta}$, based on a particular IRT model.

If examinees respond according to the chosen IRT model (i.e., if their data fit the model), l_z has a sampling distribution that is asymptotically normal with a mean of zero and standard deviation of one. Large negative values of l_z indicate misfit or unlikely response patterns. Large positive values indicate over-fit, an estimation result that rarely surfaces with real data. Some research has shown that l_z is the most efficient at detecting non-model-fitting response patterns when the test has items of varied difficulty and small lower asymptote parameters (Reise & Due, 1991). Using simulation studies, Drasgow and Levine (1986) found that l_z performed satisfactorily

with detection rates approximately 65 percent correct identification of simulated cheaters, where cheating was induced.

In a slightly more complicated way, $ECI4_z$ (Tatsuoka, 1984) was designed to isolate the difference between an observed response vector and an IRT-based or response function, incorporating both the difficulty of the item and the conditional [binomial] variance of estimation into the ratio[5]. It contains information of observed response and model predicated probability. $ECI4_z$ is defined as:

$$ECI4_z = \frac{\sum_{i=1}^{n}\{[P_i(\hat{\theta}) - u_i] * [P_i(\hat{\theta}) - \bar{P}(\hat{\theta})]\}}{\sqrt{\sum_{i=1}^{n}\{P_i(\hat{\theta}) * \left(1 - P_i(\hat{\theta})\right) * [P_i(\hat{\theta}) - \bar{P}(\hat{\theta})]^2\}}} \quad (19)$$

where $\bar{P}(\hat{\theta})$ is defined as the condition mean of $P_i(\hat{\theta})$ for a particular set of items, $P_i(\hat{\theta})$ is the model-predicted probability of a correct answer to the i^{th} item, and u_i is the observed response of the i^{th} item. Lewis (1999) asymptotically proven that $ECI4_z$ is distributed with a standardized normal distribution with a mean of zero and standard deviation of one. Large positive values may result from correctly answering more difficult items. Conversely, large negative values may result from an examinee correctly answering more easy items and fewer difficult items than would be expected on the basis of the IRT model in use.

Despite their popularity and apparent value for general data-model misfit questions related to individual examinees, l_z and $ECI4_z$ s may be fundamentally flawed for use in detecting cheaters for an obvious reason. Both of the two indices

[5] This ECI4z statistic is conceptually very similar to other, more traditional data-model fit statistics (see for example., Wright & Stone, 1979 and Yen, 1983)

make use of the estimated ability or proficiency to calculate the likelihood values. However, the estimates are potentially contaminated by some unknown mixture of guessing, or response due to cheating, and maybe even some attempts to respond based on legitimate knowledge and skills. In that case, therefore, the null hypothesis of the two indices may not be held any more. Furthermore, when the number of cheaters in a test grows, the two indices may lose their power to identify unusual response patterns due to cheating (versus other causes). As noted above, factors other than test cheating may be equally or more plausible to explain aberrant responding, where identical aberrant responses may have vastly different explanations.

Summary of the Cheating Detection Methods

Three categories of statistical cheating detection methods were introduced and discussed in this section, with examples provided under each category. These methods have all been applied, at least in research settings—less so in applied testing settings—and generally shown to be useful tools. For example, Wollack's (1997, 2006) research has suggested that the ω index consistently maintains expected level of statistical power and false positive error rates, even for small sample size applications. Segall's model was also shown to work well with simulated cheating data. For the person-fit statistics, as noted, Drasgow and Levine (1986) used simulation research to empirically demonstrate that the l_z statistic could perform satisfactorily to help detect cheating that results in aberrant response patterns.

Because most of these detection methods require only access to the scored response data and sufficient computational resources, they are obviously cost/time-efficient; at least when compared to reviewing irregularity reports submitted by test proctors or carrying out extensive, physical investigations of opportunity and motive to cheat, as well as background checks on cheating suspects. At worst, most cheating analysis results provide only one type of evidence that may or may not be plausibly linked to suspected cheating activities. Few testing organizations would responsibly accuse somebody of cheating strictly based on a statistical index. However, even when the cheating indices produce the primary flagging mechanisms for suspected cheating—based almost solely on the observed data—we might arguably want to hold those indices to a very conservative standard to avoid falsely accusing individual examinees.

As suggested in this chapter, most of the existing statistical cheating detection methods have various limitations—a point also made by Wollack (2006). There are obvious limitations such as conclusively identifying potential sources/causes and determining directionality, having somewhat speculative definitions of the parameters in some of the detection models, and a general neglect of those models to distinguish cheaters' real proficiency from their proficiency due to cheating. More specifically, almost all of the copying indices are designed specifically to detect answer copying by comparing a suspected copier to a source, using similarity-based statistics. Although these methods have been effectively used with large-scale paper-and-pencil

tests, their effectiveness seems severely curtailed for many types of computerized tests (adaptive tests and the other types of computerized testing at many small centers on multiple days). For example, it is difficult to identify possible sources when most examinees take tests at different times and in a wider variety of locations, much less extract sufficient data, especially if fewer examinees ever see exactly the same items in the same order (Iwamoto, Nungester & Luecht, 1998).

Cheating models such as Segall's (2002) IRT-based cheating model might be limited by data demands—especially for smaller-sample applications and the plausibility of the rather strong assumption about the capability of cheaters to reproduce exposed materials with certainty. However, the inclusion of way to estimate the magnitude of score gain due to cheating is a particularly appealing aspect of Segall's model that is extended in this dissertation. Finally, most person fit indices tend to confound real proficiency and proficiency due to cheating (i.e., cheating skill) and also suffer from sufficient explanation/proof of cheating as the cause of the aberrance. The lack of generalized utility of these indices and models was articulated by Dwyer and Hecht (1996):

> Our position, which is supported by both the courts and statisticians, is that one should never accept probabilistic evidence as sufficient evidence of cheating merely because a pattern of answers is deemed to be statistically improbable. In every case, reasonable competing explanations should be evaluated, limitations of the mechanical detection strategies must be taken into account, and the inherent variability in the reliability and validity of test design and administration of all but the most rigorous of standardized tests must be considered. (p.133)

In reality, building a probative argument for cheating is a matter of providing as many indicators and related, convincing evidence as possible that favors the hypothesis that cheating is the only plausible cause. This requires also demonstrating that alternative explanations or cause of the results are neither reasonable nor probable, in a relative sense. Most statistical models require assumptions. Those assumptions also need to be justified any time the model is used, not just taken on faith (or assumed to hold under some nebulous asymptotic statistical or mathematical theory. Roberts (1987) pointed out the consequence inappropriately applying statistical detection methods that lacking validity, saying " To rely on statistical evidence alone when no observational evidence is available is not only considered poor practice…but requires the prosecution to prove charges of cheating based only a probability" (p.79).

Acknowledging the apparent caveats about valid and ethical uses of test cheating detection methods, and a rather rich literature covering a plethora of statistical methods, this dissertation also recognizes the value and need to move the research forward in a credible and significant way. Toward that end, a new cheating detection model, called the *Deterministic, Gated IRT Model* is introduced and developed in Chapter 3. This new model, while certainly not immune to abuse or probative misuse, at least attempts to correct some of the technical limitations of other methods.

CHAPTER III

THE DETERMINISTIC, GATED ITEM RESPONSE THEORY MODEL

In this chapter, a new cheating detection model, called *Deterministic, Gated Item Response Theory Model* (DGIRTM), is proposed to detect and respond proactively to test cheating by groups, or individuals in paper-based tests (PBT) or computer-based tests (CBT). The new model is presented from different aspects, by describing its modeling philosophy, origin, mathematic equations, estimation framework, and its conceptual features. Then a simulation research is designed to demonstrate the DGIRTM's characteristics, followed by a brief introduction a real dataset where the DGIRTM is applied to show its usefulness.

The Origin

A New Way to Look at Test Cheating

Test cheating, as a negative factor to impact test validity, is a long-existing and challenging problem for the testing industry since the first test was administrated. The knowledge obtained from cheating activities influences an examinee' responses with respect to correctly or incorrectly answering items on which examinees cheated, and therefore their answers are not fully dependent on their true ability. The benefit

achieved by cheating is a threat to the validity and inference made based on tests. A growing benefit (i.e., positive or negative score gain) will result in an increased threat to the validity or accuracy of the inference made based on tests. Therefore, test cheating detection methods should be able to detect cheaters who obtain noticeable positive or negative benefit from test cheating, which severely jeopardizes the accuracy of the test inference. However, in real application it seems that it is not necessary to detect cheaters who actually have no or little benefit from cheating, because such cheaters who have no or little score gain (or benefit) accordingly have no or little impact on the validity or accuracy of the inference made based on tests and thus they could be treated as non-cheaters in most real applications. The DGIRTM was developed for the purpose of detecting test cheating by modeling the benefit that cheaters obtain from their cheating activities.

In this new model, cheaters have two skills, the skills are their true proficiency determined by his/her native cognitive ability and the second skill is their cheating skills determined by the combination of their true proficiency in addition to the cheating information that they have obtained. Thus, the difference between the true proficiency and their cheating ability is the benefit from their cheating. As a note, such difference between the true proficiency and cheating ability is also called *score gain* in this dissertation. It is a normal sense to believe that the score gain due to test cheating is positive (i.e., cheating ability is greater than true ability). The score gain can be used to reflect the effectiveness of the cheaters' cheating on tests. The validity

of the inference made based on tests can be jeopardized by effective cheaters with significantly large score gain, but is only slightly impacted by un-effective cheaters (those who cheat but have no or little score gain). The new model is designed to detect effective cheaters with a noticeable score gain.

Conceptually, this model defines test cheaters based on the effectiveness of cheating activities, but not on whether cheaters have or have not conducted cheating activities. In other words, the model treats the test takers who are un-effective cheaters as non-cheaters, where an un-effective cheater may have cheated, but did not significantly improve his or her score in cheating. The threshold between the effective cheaters and non-cheaters is mainly determined by the level of measurement error-as determined by the DGIRTM's estimation and the confidence level to identify test cheaters (i.e., a cut point for determining who are cheaters; this cutting point will be discussed in the model estimation section). Specifically, non-effective cheaters whose score gain are within the standard error of the model estimation are more likely to be classified as non-cheaters, and otherwise effective cheaters whose score gain is above the standard error of model estimation are more likely to be identified as cheaters under the DGIRTM framework.

Essentially, the DGIRTM treats the un-effective cheaters' score gain as scoring error. From the practical standpoint, it is safer to treat un-effective cheaters as non-cheaters, because the severity of the impact of their score gain on test score validity is at the level that is the same as the standard error of the model estimation.

From the statistical standpoint, it is difficult to distinguish the un-effective score gain from the standard error of examinee scoring. From the educational and legal standpoint, we would prefer to believe that test takers are honest and innocent in their tests unless presented with strong evidence otherwise.

As discussed in Chapter 2, test cheating due to item preview, item memorization or internet collaboration now noticeably exists due the evolvement of computerized testing, which provides more administration windows with a higher frequency. In such cheating due to item preview, item memorization or internet collaboration, it is reasonable for us to argue that examinees are only possible to review the exposed items (i.e., the items have been used in previous forms) and they are impossible to have access to the unexposed items (i.e., the items have not been used). Generally speaking, the DGIRTM is developed to detect test cheating conditional on the item exposure status (exposed or not exposed) in a test by modeling examinees' score gain.

The General Cheating Model

A statistic model is proposed to describe test cheating context by distinguishing cheaters' true proficiency from their proficiency due to cheating. The model is called the *general cheating mode,* which is the origin of the DGIRTM. In this model, two latent traits are used to separately characterize a cheater's real knowledge level and cheating. One is test takers' true ability (θ_t), which is the latent trait to characterize test takers' real knowledge level, and the other ability is their cheating ability (θ_c),

which is the latent trait to describe the examinees performance due to cheating. As a note, it is argued that cheating ability is a combination of examinee true ability and a score gain. Mathematically, the score gain (labeled by Δ) is the difference of examinee cheating ability and true ability ($\Delta=\theta_c-\theta_t$). As stated before, in reality it is reasonable to believe that the score gain is greater than zero.

Test takers will only rely on their true ability to answer items when they do not cheat on tests, while they will use their cheating ability to solve items when they cheat. Based on this logic, the expected probability of a correct answer to an item according to the *general cheating model* is defined as equation 20, equation 21 or equation 22:

$$\mathrm{P}\left(\mathrm{U_{ij}}=1|\theta_\mathrm{t},\theta_\mathrm{c},\mathrm{I_i},\mathrm{T_j}\right)=(1-\mathrm{I_i})*\mathrm{P}\left(\mathrm{U_{ij}}=1\middle|\theta_\mathrm{t}\right)+\mathrm{I_i}*\left[\left(1-\mathrm{T_j}\right)*\mathrm{P}\left(\mathrm{U_{ij}}=1\middle|\theta_\mathrm{t}\right)+\mathrm{T_j}*\mathrm{P}\left(\mathrm{U_{ij}}=1\middle|\theta_\mathrm{c}\right)\right] \tag{20}$$

$$\mathrm{P}\left(\mathrm{U_{ij}}=1|\theta_\mathrm{t},\theta_\mathrm{c},\mathrm{I_i},\mathrm{T_j}\right)=\left(1-T_j\right)*P\left(U_{ij}=1\middle|\theta_t\right)+T_j*\left[(1-I_i)*P\left(U_{ij}=1\middle|\theta_t\right)+I_i*P\left(U_{ij}=1\middle|\theta_c\right)\right] \tag{21}$$

$$\mathrm{P}\left(\mathrm{U_{ij}}=1|\theta_\mathrm{t},\theta_\mathrm{c},\mathrm{I_i},\mathrm{T_j}\right)=\left(1-T_j*I_i\right)*P(U_{ij}=1|\theta_t)+T_j*I_i*P(U_{ij}=1|\theta_c) \tag{22}$$

And

$$P\left(U_{ij}=1\middle|\theta_t\right)=\Phi(\theta_t-b_i) \tag{23}$$

$$P\left(U_{ij}=1\middle|\theta_c\right)=\Phi(\theta_c-b_i) \tag{24}$$

where T_j is the j^{th} examinee's cheating tendency which is a parameter to represent the cheating degree at the examinee level, I_i is the i^{th} item's cheating difficulty which is a parameter representing how many examinees can cheat on the i^{th} item. As a special

note, T_j and I_i are continuous variables in the *general cheating model*, where T_j is re-defined as a dichotomous gating variable and I_i become a dichotomous model input to define the item exposure status in the DGIRTM. $\mathrm{P}(\mathrm{U_{ij}} = 1|\theta_t)$ and $\mathrm{P}(\mathrm{U_{ij}} = 1|\theta_c)$ are the probability of a correct answer given an examinee's true ability and cheating ability, and b_i is the i^{th} item's difficulty. Equation 20 and 21 are essentially equivalent, the difference is that equation 20 represents that test cheating is modeled from item prospective, and in equation 21 the test cheating is modeled from examinee prospective. Both equation 20 and equation 21 could be simplified into equation 22. The statistic model represented by equation 22 is the simplified *general cheating model.*

The *general cheating model* has a mixed structure by using two latent traits to characterize the characteristics of each examinee. Specially, the probability of a correct response is defined conditional on the product of examinees' cheating tendency and items' cheating difficulty ($T_j * I_i$, in Equation 22). Given this information, the probability then depends on either the examinees true ability or cheating ability (i.e., examinees will only use their cheating ability if they are, in fact, cheaters and they have seen the item). In this dissertation, the measurement part of the general cheating model is chosen to be *Rash model* based, because the *Rash model* is well-accepted and applied by practitioners in real settings.

The distinctive feature of *the general cheating model* is that it makes use of two categories of examinee's abilities together with examinee's cheating tendency and

item cheating difficulty to generally characterize test cheating contexts. Unlike the person fit indices, copying indices and Segall's cheating model, *the general cheating model* provides a new philosophy to characterize test cheating. Under *the general cheating model*, when I_i decreases (i.e., the i^{th} item becomes increasingly difficult to be cheated on), examinees will increasingly rely on their true ability to solve the item, no matter how strongly they want to cheat. In addition, when an item is easily assessed and cheated by examinees, examinees with a low degree of cheating tendency tend to answer the item using their true ability, and examines with a high degree of cheating tendency would like to respond to the item based on their cheating ability. Although *the general cheating model* defines an effective philosophy for cheating, this model is not identified. Thus, a modified cheating model, called *the Deterministic, Gated Item Response Theory Model (DGIRTM),* is derived from *the general cheating model*, which will be presented in the next section.

The Deterministic, Gated IRT model

The Parameters of the Deterministic, Gated IRT Model

The *Deterministic, Gated IRT model* (DGIRTM) is still based on a Rash model that is conditional on whether a person is a cheater or a non-cheater by using a set of two abilities (true ability and cheating ability) to characterize cheaters real knowledge and cheating severity, but T_j and I_i are modified to fix the estimation indeterminacy of the *general cheating model*. The DGIRTM is defined as equation 25:

$$\mathrm{P}(\mathrm{U_{ij}}=1)=\mathrm{P}(\mathrm{U_{ij}}=1|\theta_{\mathrm{t}})^{1-\mathrm{T_j}}*[(1-\mathrm{I_i})*\mathrm{P}(\mathrm{U_{ij}}=1|\theta_{\mathrm{t}})+\mathrm{I_i}*\mathrm{P}(\mathrm{U_{ij}}=1|\theta_{\mathrm{c}})]^{\mathrm{T_j}} \tag{25}$$

And

$$P(U_{ij}=1|\theta_t)=\Phi(\theta_t-b_i) \tag{26}$$

$$P(U_{ij}=1|\theta_c)=\Phi(\theta_c-b_i) \tag{27}$$

where θ_t *is* the true ability determined by examinee's cognitive nature, θ_c is the cheating ability determined by examinees' cheating skill (a combination of true ability and score gain), b_i is the item difficulty and ϕ is the logistic function . T_j in equation 25 is not continuous parameter any more as it defined in the general cheating model. It is an indicator, or gated, variable used to label cheating status which is defined as:

$$T_j=\begin{cases}1, & \text{when cheating}\\ 0, & \text{when no cheating}\end{cases} \tag{28}$$

where T_j=1 represents that the j^{th} examinee cheats when given the opportunity, and T_j=0 represents that the j^{th} examinee does not cheat. No middle phase between cheating and non-cheating exists in the DGIRTM. The T_j is called a gated variable because it classifies examinees into two groups (cheating and non-cheating group). When an examinee obtains a noticeable score gain from their cheating activities, the gated variable T_j will assign him or her to the cheating group, he or she will remain in the non-cheating group otherwise. The tendency of assigning examinees into the cheating group increases when their score gain increases.

The second modification to the general cheating model is that the latent variables I_i in equation 25 is now defined as model input that describes item status with respect to possible exposure (i.e., possibility that cheaters can cheat on this item). Note that I_i can be dichotomously or continuously defined based on specific applied settings. In the continuous case, I_i is the probability that an examinee cheats on the i^{th} item in general. For example, Mcleod, Lewis & Thissen (2003) made an assumption that item cheating difficulty has a non-linear relationship with item-difficulty, which is defined in equation (29):

$$p(m_i|b_i) = \frac{1}{1+\exp(-b_i)} \tag{29}$$

where b_i is the i^{th} item's difficulty; $p(m_i|b_i)$ refers to the probability of the i^{th} item being memorized given its item difficulty. Such probability of being memorized could be used as I_i to represent the items' cheating difficulty. Other reasonable assumptions with respect to item cheating difficulty could be incorporated into this model as a way to identify cheaters according to specific needs.

The variable I_i can also be dichotomously defined relative to item status. Specifically, I_i can be used to specify item exposure status, which is defined as:

$$I_i = \begin{cases} 1, \text{if exposed} \\ 0, \text{if unexposed} \end{cases} \tag{30}$$

where I_i=1 means that the i^{th} item has been exposed, and I_i=0 means that the i^{th} item has not been exposed. The dichotomous definition of I_i implies that items have been divided into two categories: exposed items and unexposed items. Empirically,

examinees could only have beneficial information of those exposed items, but no chance to assess item information of those unexposed items. As stated previously, it is argued that score difference between the exposed items and unexposed items should be due to test cheating by the item over-exposure or item preview. In this dissertation, the dichotomous definition of I_i is adopted because the exposure information about items is easily accessed by practitioners in real settings.

As a demonstration, when T_j=1 and I_i=1, $P(U_{ij}=1|\theta_t, \theta_c, b_i, T_j)=P(U_{ij}=1|\theta_c, b_i)$,which implies that when examinees cheat and the items have been exposed, the probability that the j^{th} examinee correctly answers the i^{th} item is defined by that examinee's cheating ability. When T_j=1 and I_i=0, $P(U_{ij}=1|\theta_t, \theta_c, b_i, T_j)=P(U_{ij}=1|\theta_t, b_i)$, which means that the probability that the j^{th} examinee correctly answers the i^{th} item is defined by his true proficiency, because the j^{th} examinee could not cheat on an unexposed item although he wants to cheat. In addition, when T_j=0 and I_i=1, $P(U_{ij}=1|\theta_t, \theta_c, b_i, T_j)=P(U_{ij}=1|\theta_t, b_i)$, which means that the probability of a correct answer is defined by the j^{th} examinee's true ability, because the j^{th} examinee does not want to cheat although he has chance to cheat on the i^{th} exposed item. Thus, successful cheating on a test is jointly determined by both examinees' cheating status and the status of the items on the test. Under the DGIRTM definition, test takers only successfully cheat on a test when they want to cheat and items have been exposed.

Generally, the DGIRTM has the model input I to specify item exposure status and four model parameters which will be estimated based on the response data (e.g.,

item difficulty b, examinees' true ability θ_t, examinees' cheating ability θ_c and cheating status T). Cheating ability θ_c is the latent trait to characterize cheaters' cheating level, which is essentially a combination of true ability and score gain, and the true ability θ_t, is the latent trait to characterize examinees' real knowledge level or competence. The two latent trait abilities (cheating and true ability abilities) and one single item parameter (item difficulty) imply that cheaters' cheating activities only change cheaters' ability and have no impact on item characteristics.

In the DGIRTM, the true ability parameter is scored mainly by the unexposed items and the cheating ability is scored mainly based on the exposed items. Given a certain examinee, if his/her two latent traits scored by the two different sets of items exhibit a noticeable difference (i.e., a score gain above scoring error), the gated variable T will trigger and assigns him/her into the cheating group. A higher score gain results in a higher likelihood to be treated as a cheater. In the model estimation section, the algorithm of how the model could assign the examinees with noticeable score as cheaters is presented.

Model Estimation of the Deterministic, Gated IRT model

Markov Chain Monte Carlo (MCMC; Patz & Junker, 1999a, 1999b; Mislevy, Almond, Yan, D., & Steinberg,1999; Templin & Henson 2006; Templin, Henson, Templin, & Rousso, 2008; Henson, 2009. etc…), as a representative Bayesian estimation algorithm, is now gaining popularity in educational measurement field. As

Patz and Junker (1999a & 1999b) stated, MCMC is a sampling method to simulate posterior distributions for multivariate variables based on the prior information available-that is a Bayesian approach, so the features of posterior variables could be characterized based on the simulated posterior distributions. In Bayesian estimation, a variety of prior distributions of the model parameters could be easily incorporated into estimation. Hierarchical Bayesian inference via MCMC sampling (i.e., a MCMC estimation procedure embedded with a hierarchical parameters) is widely used (e.g., in diagnostic classification models) due to its ability to deal with multi-dimensions as well as computer's increasing computational power. As Fu pointed out (2005) that

> Unlike the Bayesian model estimation, where the parameter space is searched to find the modal point of the posterior distribution, Bayesian inference with MCMC draws a sufficient number of samples from the posterior distribution, and then makes inferences based on the distribution of these samples, such as the mean and variance of the distribution. (p.96).

Considering the two dimensions of the model, a Hierarchical MCMC algorithm is adopted to estimate the DGIRTM. Specifically, the prior distribution of each parameter in the DGIRTM is defined as following:

$$\theta_t \sim N(0,1) \tag{32}$$

$$\theta_c \sim N(0,1) \tag{33}$$

$$b \sim N(0,1) \tag{34}$$

$$T = 1, when\ \theta_t < \theta_c \tag{35}$$

The true proficiency *(θ_t)*, cheating ability (θ_c) and item difficulty (*b)* have independent standardized normal distribution as their priors, and the cheating status (*T*) is governed by the relationship between the prior of true proficiency (θ_t) and the prior of cheating ability (θ_c). When the prior of an examinee's true proficiency is greater than the prior of his/her cheating ability, *T* will be assigned 0 as its prior, which implies he or she is proposed to be a non-cheater; otherwise, *T* will be assigned 1 as its prior, which means that she or he is proposed to be a cheater. The theoretical base of the MCMC is not the focus of this dissertation, but the detailed MCMC estimation algorithm for this DGIRTM in R statistic language is provided in the Appendix I of this dissertation. As a note, in the set of R code in the Appendix, the mean of cheating ability is hierarchically modeled, allowing for its change according to the samples, as opposed to a fixed mean zero in equation 33. The benefit of allowing a changing mean of the cheating ability is to solve, or at least reduce, the scale shift problem when cheating size is large (the scale shift problem will be discussed in the Chapter 4).

As implied in Equation 35, a cheaters' cheating ability is always greater than their true ability as is defined in the prior distribution for *T.* Thus, examinees have a higher probability of correctly answering items with beneficial knowledge from cheating than without cheating. From a practical standpoint, those cheaters who obtain positive score gain from their cheating activities are those who bias our inference of ability from test scores. The degree of such bias grows given an

increasing score gain. It should be noted that even though cheaters might have negative score gains due to the wrong information they obtained from their cheating activities, such cheaters with negative score gain will not be of interest because they have already punished themselves by their own cheating activities.

The DGIRTM only provides an item difficulty to represent the item characteristics, which are constrained to be equal between the cheater and non-cheater for a given item. The DGIRTM estimated by the Hierarchical MCMC also provides three parameters to characterize each examinee. One is the examinees' true proficiency (θ_t), which is a parameter representing examinee's real knowledge level. The second examinee parameter is the examinees' cheating ability (θ_c) which is a parameter to characterize the level of ability when that examinee cheats. Finally, the last parameter related to examinees is the gated variable *T*, which is essentially an indicator if the ability characterized by the exposed items (i.e., cheating ability) is significantly greater than the true ability mainly characterized by the unexposed items.

Given an examine, the probability (represented by $\tilde{T}$) that his/her cheating ability is significantly greater than his/her true ability is the estimated mean of the posterior distribution of *T* (the posterior distribution of *T* is obtained via MCMC sampling procedure). Statistically, throughout the MCMC estimation algorithm, $\tilde{T}_j$ describes the proportion of proposals where T_j=1 among all the proposed T_j, which measures the posterior probability that the j^{th} examinee's cheating ability is greater than its true ability. Therefore, $\tilde{T}_j$ could be explained as a p-value of a significance

test on the difference between the two latent traits scored by exposed and unexposed items for the j^{th} examinee. The directionality of the magnitude of $\tilde{T}$ positively links to difference between the cheating ability and true ability (i.e., a greater difference leads to a greater level of $\tilde{T}$). Therefore, $\tilde{T}$ could be directly explained as the probability that examinees' cheating ability is significantly greater than their true ability. Most importantly, $\tilde{T}_j$ does not define the probability that examinee is a cheater, but instead defines the probability that the cheating ability is higher than the true ability. However, conditional on the item exposure status in the DGIRTM where cheating is the reasonable explanation of a significant difference between the cheating ability and true ability, $\tilde{T}$ could indirectly serve as an indicator that examinees cheat or not.

Examinees could be classified as cheaters or non-cheaters by setting a cut point P_c ($P_c \in (0,1)$) for $\tilde{T}$, as shown in equation 36,

$$T_j = \begin{cases} 1, \tilde{\mathrm{T}}_{\mathrm{j}} \geq \mathrm{p_c} \\ 0, \tilde{\mathrm{T}}_{\mathrm{j}} < \mathrm{p_c} \end{cases} \quad (36)$$

If the $\tilde{T}_j$ is greater than P_c, then examinees should be classified as cheaters (T_j=1); if the $\tilde{T}_j$ is less than the P_c, examinees should be classified as non-cheaters (T_j=0). Specifically, a greater $\tilde{T}$ results in a higher chance to be classified as a test cheater under the DGIRTM. The selection of cutting point P_c is a critical factor to impact the False Positive and False Negative rate of the classification. When P_c increases, it will correspondingly decrease the False Positive rate of classification,

since less non-cheaters are mis-classified as cheaters; when P_c decreases, it will accordingly increases the False Positive rate because more non-cheaters will be mis-classified as cheaters.

In MCMC, the posterior distribution of cheating status *T,* as determined by Bayesian Theorem, is jointly determined by its prior and the cheating information carried by the real response data. In terms of the prior of *T*, the prior probability for each examinee being a cheater is 0.5 (i.e., the probability of T=1), which is determined by the relationship between the prior of the true proficiency and the prior of cheating ability. Specifically, the prior of *T* is equal to the prior probability that true ability is greater than the cheating ability. Because the prior for true ability and cheating ability is defined such that they are independent normal distributions with mean zero and variance equal to unity, this prior probability of T=1 is 0.5.

The amount of cheating information is determined by cheaters' score gain. As stated above, the score difference of interest is essentially the difference between examinee's ability scored by the unexposed items and that scored by the exposed items. As implied by the DGIRTM, ability by the exposed items is assumed to be always greater than the ability scored by the unexposed items for cheaters and otherwise there would be no difference. Such score gain that goes above the scoring error level would result in a high value of $\tilde{T}_j$. When the cheating ability is greater than the true ability, there is enough information for the DGIRT model via MCMC to "accept" the assumption that the cheating ability is greater than the true ability, but

when the score gain is small (i.e., below the standard error of the model estimation), not enough information is available for the DGIRTM via MCMC to "accept" that the cheating ability is greater than the true ability.

With respect to non-cheaters, no score difference theoretically exists between the exposed and unexposed items, because non-cheaters use their true ability to answer both (i.e., no cheating information is available in real data and thus the priors will determine the posterior distribution). As a result, the marginal posterior distribution of a non-cheater's true ability should be identical to the marginal posterior distribution of the cheating ability. In addition, there should be no association between the cheating ability and the true ability given an examinee, as determined by the fact that the cheating and true ability are independently proposed by two standardized normal distribution. Thus, the probability of the cheating ability being higher than the true ability for the posterior distribution of a non-cheater will equal 0.5, theoretically. That is, a $\tilde{T}_j = 0.50$ is an indication that an examinee is not a cheater (as opposed to $\tilde{T}_j = 0.00$).

With respect to test cheaters, a score difference theoretically exists because cheaters use their real competence to answer unexposed items and use their cheating ability to respond to exposed items. Such score difference or score gain by cheaters increases along with their increasing cheating effectiveness. Cheating effectiveness is used to describe how effective cheaters' cheating activities are, which represents the cheating information embedded in the real response data. Effective cheaters will have

noticeable score gain and thus the posterior mean for the cheating ability is expected to be higher than the posterior mean of the true ability. As a result, the probability of the cheating ability being greater than the true ability will be higher than 0.5 (which is the assumed probability of the prior distribution). Notice that an un-effective cheater is expected to perform in a similar way as a non-cheater and, thus, cannot be distinguished from a non-cheater (i.e., the model treats an un-effective cheater as a non-cheater). Therefore, a $\tilde{T}_j = 0.50$ is also an indication that an examinee is a un-effective cheater who acts similarly as non-cheater, and in contrast, a $\tilde{T}_j$ close to one indicates that the j^{th} examinee is a highly likely test cheaters. The closer to one, the higher certainty that the cheating ability is greater than the true ability with a noticeable difference, which implies a high probability of being a cheater.

Because non-cheaters or un-effective cheaters' ability scored by the unexposed items would be exactly equal to their ability scored by the exposed items, $\tilde{T}_j$ of non-cheaters/un-effective cheaters should be equal 0.5. As a result, if 0.5 is set as $\tilde{T}_j$'s cut point, around one half of non-cheaters would be incorrectly classified as cheaters. In other words, we only have a 50 percent confidence level to classify the j^{th} examinee as a test cheater. However, a basic requirement of test cheating detection model is to sensitively identify cheaters with a small degree of error, because of our testing evaluation purpose to allocate right talents into right learning or working positions and to treat innocent test takers with fairness. Testing agencies might also expose themselves to potential law suit charges when an innocent test taker is classified as

test cheaters. Thus, as an introduction to the DGIRTM in this dissertation, the cut point for $\tilde{T}_j$ is set as 0.9, as opposed to 0.5 which might be used in a traditional two class model, which implies the j^{th} examinee is classified as a test cheater or non-cheater with a 90 percent of confidence level. As a note, practitioners could also choose 0.95 as the cut point for $\tilde{T}_j$, which implies a 95 percent confidence level.

Hierarchical MCMC, as the model estimation algorithm, offers flexibility to incorporate a variety of priors and conduct a large number of dimensions of estimation while allowing for the conceptual expectations on the DGIRTM. The features of the DGIRTM are conceptually discussed and compared with the reviewed cheating detection methods in next section.

Features of the Deterministic, Gated IRT model

Generally, the DGIRTM is a Rash based mixture model. It incorporates two categories of latent abilities to separately characterize examinees' real knowledge level and cheaters' cheating degree. The model input related to item status classifies test items as exposed and unexposed items, and therefore an examinee is expected to perform better on the exposed items than on the unexposed item, if he or she is cheating. The cheating status parameter is a direct index to represent the level of the score difference between exposed and unexposed items, and thus indirectly serves as an indicator that the flagged examinees by the DGIRTM might cheat on the exposed items.

Under the DGIRTM framework, the item difficulty is fixed on a single scale for both cheaters and non-cheaters (i.e., the item difficulty is equivalent for cheating and non-cheating cases). It is assumed that the cheating scale determined by the exposed and unexposed items are the same, and the difference is that through examinees who cheat on test appear to have a higher ability along with the fixed scale. Conceptually, we believe that the ability of examinees improves and the item characteristics do not change in cheating context.

The model input is dichotomously defined to represent exposed and unexposed items, which helps to define inner person comparisons conditional on the two groups of items (exposed or unexposed items). As already stated, a cheaters' cheating ability is defined by their response to exposed items and their true ability is estimated by their response to unexposed items. A noticeable difference between the two categories of abilities (true ability and cheating ability) is detected as an evidence of test cheating. One direct cause of a statistical difference between the true and cheating ability is that examinees have the opportunity to cheat on exposed items and no opportunity to cheat on unexposed items.

Essentially, the cheating probability is estimated by the model to represent the severity of the score gain. Statistically, it could be explained as a p-value of a significance test. A significant value of the cheating probability represents a notable score gain which is above the random error of examinee scoring. In real settings, it is the group of effective cheaters (e.g., who cheat and obtain a significant score gain)

who severely damage the test validity. The DGIRTM is designed to identify such kind of effective cheaters. In addition, notice that because this "test" for cheating is within a person, it is not necessary to have a large proportion of cheaters in the sample to allow for the identification of a cheating group.

As compared to cheating detection indices, this model also measures the level of cheaters' real competence and the severity of their cheating activities. Unlike some copying indices and person fit indices where cheaters' true ability is confounded with their cheating information, the DGIRTM distinguishes cheaters' real competence from their cheating skill. As a result, estimation and identification of cheaters is not dramatically affected by the proportion of cheaters relative to non-cheaters, like other indices (e.g., *lz* index). In addition, this model could incorporate outside information related to item status. Specifically, the parameter I_i could be defined as a continuous variable with a value interval [0,1] rather than a dichotomous input. When the model input (*I)* is used to represent the item status relative to exposure status, the DGIRTM is particularly useful to detect test cheating caused by item exposure, as discussed in this dissertation. If the model input is defined as items status relative to different item points of its administration, the DGIRTM is able to monitor examinees' growth over time.

The DGIRTM is designed to detect effective cheaters by distinguishing cheaters' real knowledge level from their cheating skills. Its features conceptually make the DGIRTM a promising tool to identify test cheating due to item over-exposure.

The Simulation Design

In this section, a simulation study is designed to illustrate the characteristics of the *Deterministic, Gated IRT model* and its power to detect test cheating. Three categories of conditions are considered in this simulation design. The conditions about item characteristics are first introduced, which is followed by a description of the simulation of examinees' ability. Next the conditions about cheating characteristics are presented. At the end of this chapter, a data generation model (a model used to generate response data) is introduced and an example process is described to demonstrate how the response data is generated. As a practical note, the number of examinees is set as 2000 in every condition, and the chain length of each MCMC chain is 5000 where the first 3000 steps serve as the burning period. Although in actual applications of this model the chain must be much longer, because the true model is known this chain length and burn-in is sufficient to achieve convergence and reliable estimates.

Simulated Item Characteristics

Three variables describing item characteristics are considered, which are item number, item difficulty location and standard deviation. Test length is a key factor determining test reliability. Although the DGIRTM is designed to detect test cheating at the examinee level, the information provided by items plays a critical role on the accuracy of cheating detection. In this study, two levels of test length are considered:

(1) a short test with 40 items; and (2) a long test with 80 items.

Other than test length, item difficulty, or location, is a key factor impacting the test information location, while the standard deviation of item difficulty is a key factor determining the magnitude of test information. Test information location refers to the point along the ability distribution where a test has the greatest amount of information. Amount of test information refers to the magnitude of information at a certain difficulty point. Two conditions of item difficulty location are simulated to control this information location. One condition is that the mean item difficulty for the exposed items is equal to the average of the cheaters' cheating ability, and the mean item difficulty of unexposed items is equal to the average of the true ability of all the test takers. A second condition is that the mean of item difficulty is set as zero. In terms of the standard deviation of item difficulty, in one condition the standard deviation is set as 0.5, which represents a high degree of information at the difficulty location, and in the other condition the item standard deviation is set as one, representing a normal degree of amount of information.

Such design is practically meaningful for the computerized adaptive tests (CATs) and normal CBT. In CATs, items are typically selected based strictly on examines estimated ability level. Those items in a CAT test normally have a mean at the mean of the examinees' estimated ability with a relatively small standard deviation. Thus, the performance of the DGIRTM under the settings which are similar to CAT applications is computed, especially given the fact that the DGIRTM is designed to

detect test cheating caused by item over-exposure. In contrast, items are usually distributed with a mean zero and a unit standard deviation in normal CBT testing settings.

The DGIRTM model separates items in a test into two categories: exposed items which can be cheated by examinees and unexposed items which cannot be cheated by examinees. The number of exposed items and unexposed items should inevitably impact the accuracy of cheating detection. In terms of the number of exposed items and unexposed items, three designs are considered in this simulation: a balanced design where the number of exposed items is equal to that of exposed items, an over-exposed design where the number of exposed items is greater than that of unexposed items, and a under-exposed design where the number of exposed items is less than that of unexposed items. Specially, 50 percent of items were exposed in the balanced items, 70 percent of items were exposed in the over-exposed design and 30 percent of items were exposed in the under-exposed design. As a summary, all the conditions related with item characteristics considered in this simulation are listed in the Table 1.

Table 1. Item Conditions

Conditions	Value
Test length	C(40, 80)
Proportion of exposed items in a test	C(0.3, 0.5, 0.7)
Item difficulty location	$b_u=M_t$, $b_e=M_c$ or $b=M_d$
Standard deviation of item difficulty	0.5 and 1

b_u =the mean of unexposed item difficulty, b_e =the mean of exposed item difficulty, b= the mean of all-item difficulty,M_t= the mean of all examinees' true ability, M_c= the mean of the cheaters' cheating ability, M_d= the mean of the difference between the true ability and cheating ability.

Simulated Cheating Characteristics

Cheating characteristics, including Cheating size, Cheating degree and Cheating effectiveness, are also considered in this simulation design. Cheating size is used to refer to how many examinees in a test are involved in cheating activities. Cheating degree represents how many cheaters cheated on specific exposed items. Cheating effectiveness is used to describe how effective cheating activities are (i.e., how much score gain cheaters obtain from their cheating activities). These three factors together determine the amount of test cheating information embedded in response data.

Different tests could have different rates of cheating. For example, a test with items which have been exposed for a long period of time may be cheated on by a large number of test takers, but a test with recently exposed items may have lower rates of cheating. Three levels of cheating size are considered: a low level cheating (5 percent of test takers cheating in a test), medium level cheating (35 percent of test takers cheating in a test), and high level cheating (70 percent of test takers cheating in a test). The 70 percent cheating size is simulated to represent the organized concert cheating (i.e., group cheating), which has been known to occur in reality. For instance, some test takers share items they remembered by internet (called internet collaboration), or some individuals purposely memorize items and sell them to test takers afterwards.

In addition to Cheating size, each exposed item has been cheated on by a different number of cheaters. Essentially, each exposed item has a different cheating

degree. Two levels in terms of cheating degree are considered. In the low cheating degree condition 50 to 80 percent of cheaters cheat on a given exposed item. In the high cheating degree condition 80 to 100 percent of cheaters cheat on an exposed item.

Cheating effectiveness is the last factor we considered in terms of cheating characteristics. As stated above, some cheaters are less likely to correctly respond on all the cheated items. Therefore, even though cheating can improve the overall observed performance, it may be to different degrees because of many real factors (e.g., their true ability, the degree of correctness of source information) besides the fact that cheaters conduct cheating activities. Cheaters, therefore, are not necessarily equally effective. Specifically, some cheaters might be effective enough to correctly answer all cheated items that should originally be incorrectly answered according to their true ability. Some cheaters might still incorrectly answer some of the exposed items although they cheat on the exposed items. Thus, cheaters in this study are classified as high-effective cheaters, medium-effective cheaters and low-effective cheaters.

The high-effective cheaters have the largest score gain from their cheating activities, low-effective cheaters have the smallest score gain from their cheating activities, and medium-effective cheaters' score gain is between that of the high-effective cheaters and that of the low-effective cheaters. To be detailed, the delta of the high-effective cheaters is simulated by Beta (9, 4)*3 (called *Delta1*), the delta of medium-effective cheaters is simulated by Beta (5, 5)*3 (called *Deta2*) and that of

low-effective cheaters is simulated by Beta(1.5, 5)*3 (called *Delta3*). The distributions for the simulated cheating gain scores, Delta1, Delta2, and Delta3, are shown in Figure 1 to Figure 3. The empirical statistic characteristics of Delta1, Delta2, and Delta3 are provided in Table 2.

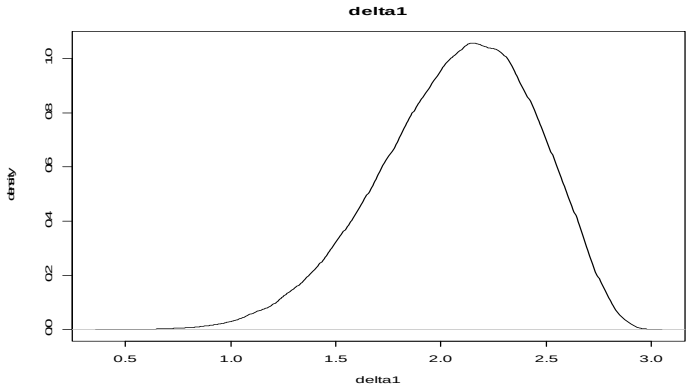


Figure 1. The Delta Distribution of High-Effective Cheaters

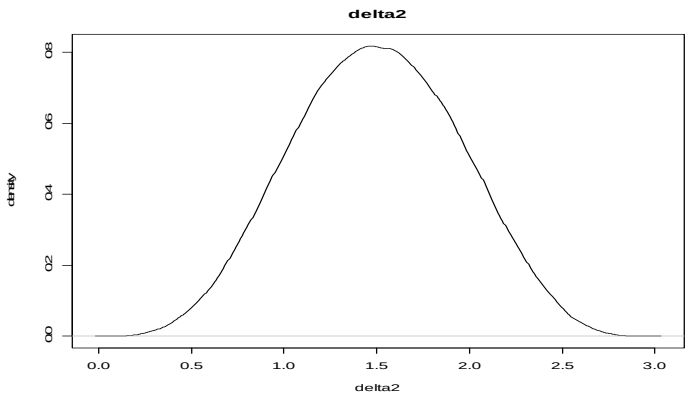


Figure 2. The Delta Distribution of Medium-Effective Cheaters

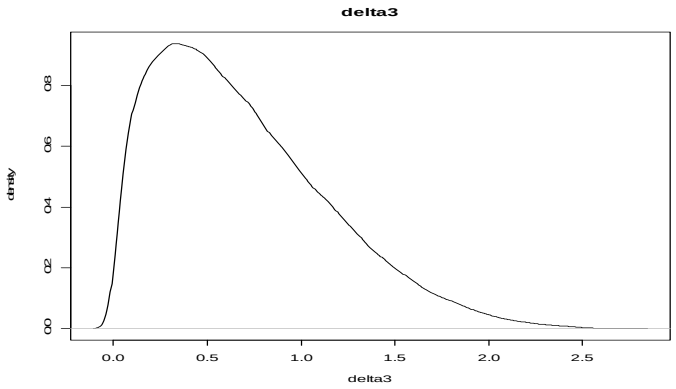


Figure 3. The Delta Distribution of Low-Effective Cheaters

Table 2. The Characteristics of Delta1, Delta2, and Delta3

Delta	Delta1=Beta(9,4)*3	Delta1=Beta(5,5)*3	Delta1=Beta(1.5,5)*3
Mean	2.078	1.500	0.692
SD	0.369	0.453	0.462
Min	0.443	0.088	0
Max	2.965	2.928	2.738

As a further illustration, Delta 1 is negatively skewed, Delta 2 is normally distributed and Delta 3 is positively skewed. In Delta 1, most of cheaters tend to be effective cheaters, but some of them are not as effective as others, in Delta 3, most of the cheaters are non-effective, but a small proportion of cheaters might be effective. Based on the cheating characteristics, the two latent traits are described in next section.

The Simulation Design about the True and Cheating Ability

Cheating ability is considered as a combination of true ability and score gain in this simulation. True ability (θ_t) was simulated by the standardized normal distribution. The *Delta* (also called *score gain*) refers to the amount of ability

difference between the true ability and cheating ability. As a note, the *Delta* is always greater than zero, because logically we believe that cheaters will have a positive score gain due to beneficial information cheaters obtain from their cheating activities.

$$\theta_t \sim N(0,1) \tag{37}$$

$$\theta_c = \theta_t + \Delta \tag{38}$$

$$\Delta \sim \text{Beta}(\alpha, \beta) * A - B \tag{39}$$

where the parameters *A* and *B* in equation (39) are scaling factors to change the lower and upper limits of the Beta distribution (this distribution is also commonly referred to as a four-parameter Beta distribution). Specifically, *A* is set as 3 and *B* is set as 0 in this simulation design.

Empirically, examinees who are competent normally tend to rely on their own knowledge and competence to respond to items. However, those examinees who are not competent enough would be more likely to seek for cheating to help them respond to items, and thus they could obtain a greater test score. From the practical standpoint, test cheaters who are competent are less likely to obtain a noticeable score gain, because there is not much room for them to improve their scores. In contrast, cheaters who are less competent might be more likely to obtain a noticeable score gain than cheaters who are actually competent. In other words, test cheaters who are less competent might have a greater degree of invalidating the inference made based on tests. Therefore, in this simulation design, 60 percent of cheaters are sampled from the low ability students whose true ability is less than -0.5, 30 percent of cheaters are

randomly sampled from the medium ability students whose true ability is between -0.5 and 0.5, and 10 percent of cheaters are randomly selected from the high ability students whose true ability is greater than 0.5. The composition of test cheating population is represented in Table 3.

Table 3. Cheaters' Distribution

Conditions	Cheating size=5%	Cheating size=20%	Cheating size=70%
True ability< -0.5	60%×5%×2000	60%×35%×2000	60%×70%×2000
True ability=U(-0.5, 0.5)	30%×5%×2000	30%×35%×2000	30%×70%×2000
True ability> 0.5	10%×5%×2000	10%×35%×2000	10%×70%×2000

As an illustration, when the cheating size is 5 percent (the first column in Table 4), and thus total number of cheaters in this case is 2000*5%=100. 60 (60%*100) out of the 100 cheaters are low ability students whose true abilities are below -0.5, 30 (30%*100) out of the 100 cheaters are the students whose true abilities are between -0.5 and 0.5, and 10 (10%*100) out of the 100 cheaters are capable students whose abilities are greater than 0.5. Together with the cheating effectiveness, the different categories of test cheaters are listed in Table 4.

Table 4. Joint Conditions of Ability Distribution

Conditions	Delta1	Delta2	Delta3
True ability< -0.5	Yes	Yes	Yes
True ability[-0.5,0.5]	Yes	Yes	Yes
True ability> 0.5	Yes	Yes	Yes
Category	High-effective cheaters	Medium-effective cheaters	Low-effective cheaters

Note: "Yes"= the joint conditions of column and row is considered in this research.

As shown in Table 4, the two conditions (i.e., the composition of examinees' real proficiency and cheating effectiveness) are fully crossed to create three categories of test cheaters: high-effective, medium-effective and low-effective cheaters. All the conditions considered in this simulation design are listed in Table 5 as an overall summary.

Table 5. Joint Conditions

Conditions	Number of conditions
Test length	2
Proportion of exposed items	3
Item Difficulty location	2
Standard deviation of item difficulty	2
The cheating degree on exposed items	2
Cheating size	3
Cheating category	3
Guessing level	1

A total number of joint conditions considered in this research is 432 (2×3× 2 × 2 ×2×3×3). A small number of replications is a common practice when using MCMC as the estimation algorithm, so each joint condition is replicated for 10 times in the simulation.

Data Generation Process

Considering the complexity of data generation, a more general model called *data generation model* is used to generate the response data, which is defined as:

$$P(U_{ij}=1)=(1-s)*P(U_{ij}=1|\theta_t)+s*\ P(U_{ij}=1|\theta_c) \quad (40)$$

where $S=T*I$. T is the dichotomous cheating parameters for each examinee, which is a

vector $N\times 1$; I is the model input relative item exposure status, which is a vector $1\times J$. Thus S is a matrix with $N\times J$, which is a joint parameter used to define the probability that a specific examinee cheats on a given item. N is the number of examinees and J is the number of items.

As an important note, I is not as a dichotomous parameters any more as it is in the *Deterministic, Gated IRT model*. In the *data generation model*, I_i is defined as:

$$I_i = \begin{cases} 0, & \text{unexposed items} \\ \epsilon\ \mathrm{U}[l_l, l_u], & \text{exposed items} \end{cases} \quad (41)$$

Where, I_i will be set as 0 if the i^{th} item is unexposed items, otherwise I_i will be uniformly sampled from U [l_l, l_u] (l_l is the lower limit and l_u is upper limit of the uniform distribution). As defined above, U [l_l, l_u] has two levels: U(0.5, 0.8) and U(0.8, 1). I_i=0 for those unexposed items, which implies that examinees have no chance to cheat on i^{th} item. I_i= U [l_l, l_u] for those exposed items, which implies that each exposed items could be randomly cheated by different cheaters with different cheating degree.

Cheaters are randomly sampled from the specific true ability domain (defined in Table 4) by controlling the total number of cheating. As an illustration, the case that cheaters are high-effective cheaters is used as an example to present the sampling process.

1) Identifying target domain (D). In terms of the true ability, the target domain consists of three sub-domains, where $D=C(D_1,D_2,D_3)$. $D_1=C(\theta_t < -0.5)$,

D_2=C(-0.5<θ_t< 0.5), and D_3= C(θ_t> 0.5). The cheating ability equals to the sum of true ability and delta1, which is $\theta_c = \theta_t + Delta1$.

2) Determining sample size from each sub-domain. For example, if the 20percent of examinees cheat in the test, the total number of cheaters is 20%*2000=400. Then cheaters in the sub-domain D_1 is 400*60%=240, the cheaters from sub-domain D_2 is 400*30%=120, and the cheaters from sub-domain D_3 is 400*10%=40.
3) Sampling cheaters from sub-domains with the correct cheating size. The *T* of each cheater is assigned a value 1, or it will be assigned a value 0.

In order to make sure each sub-domain has enough samples, a population with 100,000 samples is simulated. Then cheaters and non-cheaters are sampled from this big population by following all the simulation specification. All the conditions related to items, two categories of abilities, cheating characteristics and cheating population are jointly integrated together in the simulation process.

Comparison Baseline

The *lz* index and a simple t-test is used as the baseline for comparison to demonstrate the model's improvement in capability to detect test cheating. The *lz* index, like the DGIRTM, is used to detect test cheating at the examinee level. As stated before, previous research (Reise & Due, 1991, Drasgow & Levine, 1986) shows that the *lz* index is among the best indices to detect misfit response pattern caused by test cheating.

As opposed to use the existing *lz* index as comparison-baseline, a simple t-test is designed which originally comes from the model itself. The model separates the items into two sets (exposed and unexposed items). The DGIRTM could be essentially described as a model to identify the examinees with significant score difference between the two sets of items. Similarly, the t-test is designed based on the two sets of items without a complex model equation and estimation algorithm. In the t-test, each examinee is separately scored using an IRT model by his/her exposed items and unexposed items, thus each examinee has a pair of ability or theta scores (i.e., the latent score for each examinee), one theta score is obtained on the basis of the exposed items and the other theta score is on the basis of the unexposed items. Then a t-test on the score difference between the pair of theta scores is conducted for each examinee. The t-test is essentially a non-central t for two estimates of ability, where a pooled standard error of estimate is used in the denominator. That is,

$$t = \frac{\widehat{\theta}_1 - \widehat{\theta}_2}{SE_{pooled}} \tag{42}$$

where the numerator of the t-test is the observed score difference for estimates of θ respectively computed using the exposed items and unexposed items. As noted above, the pooled standard error of estimate is used as the denominator of the t-test. The null hypothesis is that the difference is only observed by random error in estimation and should be normally distributed with a mean zero and a unit standard deviation. The null hypothesis is that the difference by chance, should be normally

distributed with a mean zero and a unit standard deviation. A significant value of the t-test (i.e, the observed statistic value of the t-test for an examinee is greater than the preset critical value, 0.05) might imply an alterative: an unusual score difference between exposed and unexposed items exists which should not be due to chance. By comparing to the t-test, the model's advanced ability in separating cheaters from innocent test takers can be fully exhibited.

Two indices (i.e., sensitivity and specificity) are used to evaluate the accuracy and reliability of DGIRTM as well as the *lz* index and the t-test. *Sensitivity* is the proportion of true cheaters who are correctly detected as cheaters by this model, and *Specificity* refers to the proportion of real innocent test takers who are correctly classified as non-cheaters by this model. The sensitivity is an index to measure the model's power to detect test cheaters, while the specificity is an index to represent the degree of misclassifying non-cheaters as test cheaters. Given a greater sensitivity, the power to detect test cheaters grows. In contrast, a greater specificity implies a smaller degree of error. In one word, a good test cheating detection model should acquire an ability to maintain a high level of both sensitivity and specificity. In this way, the cheating detection method can efficiently identify test cheaters and yet derive a small degree of error.

CHAPTER IV

RESULT AND ANALYSIS

In this chapter, the results of the simulation study are summarized to illustrate the accuracy and reliability of the model under a variety of conditions (e.g. cheating effectiveness, test length, information location, etc….). The DGIRTM's specificity is first presented followed by a detailed discussion on the DGIRTM's sensitivity in every joint condition. Next, the accuracy of the model estimation is described by using root mean square error and correlation between the estimation of the DGIRTM and the true value. Finally, as a comparison baseline, the *lz* index and a simple t-test (which will be fully presented in the following section) are incorporated in this dissertation to demonstrate the model's improvement in capability to detect test cheating.

Model's Specificity

Given that every examinee should be treated with fairness, to correctly identify the innocent test takers is as important as to sensitively detect test cheaters. The DGIRTM seems to be powerful for correctly identifying innocent test takers based on its consistently high level of specificity in every joint condition.

Table 6. The Sensitivity and Specificity of the Informative Test with 50% Exposed Items

Conditions			High Effective		Medium Effective		Low effective	
Degree	Length	Cheat Size	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
0.8-1	Short	70%	45.60%	**98.90%**	34.20%	**99.10%**	18.10%	**99.30%**
		35%	76.60%	**96.20%**	56.40%	**97.40%**	25.90%	**98.00%**
		5%	84.80%	**95.00%**	66.40%	**96.30%**	28.70%	**96.80%**
	Long	70%	82.00%	**97.70%**	62.00%	**98.60%**	28.30%	**98.90%**
		35%	92.50%	**95.80%**	79.40%	**96.70%**	38.50%	**97.60%**
		5%	94.50%	**94.70%**	83.80%	**95.40%**	45.00%	**96.50%**
0.5-0.8	Short	70%	32.97%	**98.77%**	24.93%	**98.98%**	12.82%	**99.20%**
		35%	62.54%	**96.99%**	42.16%	**97.65%**	17.97%	**97.98%**
		5%	72.10%	**94.97%**	53.00%	**96.78%**	20.70%	**96.84%**
	Long	70%	65.97%	**97.78%**	45.16%	**98.67%**	19.21%	**99.03%**
		35%	85.77%	**96.02%**	66.06%	**96.89%**	27.37%	**97.95%**
		5%	89.30%	**94.49%**	72.00%	**95.54%**	33.90%	**96.50%**

Degree= cheating degree; Length=Test length; short= test with 40 items; long= test with 80 items; High Effective=high effective cheaters; Medium Effective=medium effective cheaters; Low Effective=low effective cheaters. 70%=70% examinees cheat in the simulated test; 35%=35% examinees cheat in the simulated test; 5%=5% examinees cheat in the simulated test; 0.8-1=each exposed items are cheated by 80%-100% test cheaters; 0.5-0.8=each exposed items are cheated by 50%-80% test cheaters.

Table 7. The Sensitivity and Specificity of the Normal Test with 50% Exposed Items

Conditions			High Effective		Medium Effective		Low effective	
Degree	Length	Cheat Size	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
0.8-1	Short	70%	37.26%	**99.95%**	27.19%	**99.67%**	15.22%	**99.10%**
		35%	69.56%	**99.05%**	50.46%	**98.49%**	23.41%	**97.73%**
		5%	80.60%	**97.99%**	61.50%	**97.16%**	27.10%	**96.36%**
	Long	70%	78.38%	**99.62%**	56.02%	**99.20%**	24.82%	**99.00%**
		35%	91.87%	**98.53%**	75.63%	**97.83%**	35.77%	**97.50%**
		5%	94.90%	**97.06%**	81.80%	**96.58%**	43.40%	**96.15%**
0.5-0.8	Short	70%	24.14%	**99.95%**	18.34%	**99.53%**	10.32%	**98.83%**
		35%	49.49%	**98.85%**	33.10%	**98.72%**	16.16%	**97.58%**
		5%	60.10%	**97.72%**	42.30%	**97.51%**	19.70%	**96.55%**
	Long	70%	55.36%	**99.58%**	35.94%	**99.33%**	17.07%	**98.73%**
		35%	79.07%	**98.25%**	57.37%	**97.98%**	25.39%	**97.68%**
		5%	84.60%	**97.15%**	68.00%	**96.36%**	31.40%	**96.21%**

Degree= cheating degree; Length=Test length; short= test with 40 items; long= test with 80 items; High Effective=high effective cheaters; Medium Effective=medium effective cheaters; Low Effective=low effective cheaters. 70%=70% examinees cheat in the simulated test; 35%=35% examinees cheat in the simulated test; 5%=5% examinees cheat in the simulated test; 0.8-1=each exposed items are cheated by 80%-100% test cheaters; 0.5-0.8=each exposed items are cheated by 50%-80% test cheaters.

As an illustration, the detailed specificity as well as sensitivity of the DGIRTM across test length, cheating degree, cheating effectiveness, and cheating size is presented in Table 6 and Table 7.

The specificity in Table 6 and Table 7 is stably distributed around 96 percent, which implies that almost all the innocent test takers (or non-cheaters) are correctly identified. Specifically, the minimum specificity in these two tables is 94.49 percent (located in the column 2 and last row in Table 6) and the maximum specificity is almost 100percent. Each of the conditions considered in this simulation study has no or only slight impacts on the model's specificity.

As a further demonstration, the specificity within different levels of cheating effectiveness (high, medium and low effectiveness) is box-plotted in Figure 4 to exhibit the DGIRTM's stable specificity in all the conditions. As shown in Figure 4, the specificity in all different cases is greater than 90 percent with fairly small variance. Especially, when 70 percent of test takers are test cheaters, the model's specificity is almost 1, which means the model makes no mistake in detecting non-cheaters. The factors, including test length, test information, proportion of exposed items, cheating size and cheating effectiveness, only slightly impact the model's specificity. One of the valuable characteristics of the DGIRTM is its capability to maintain a high level of specificity across different conditions, which ensures that the innocent test takers are correctly identified and treated with fairness.

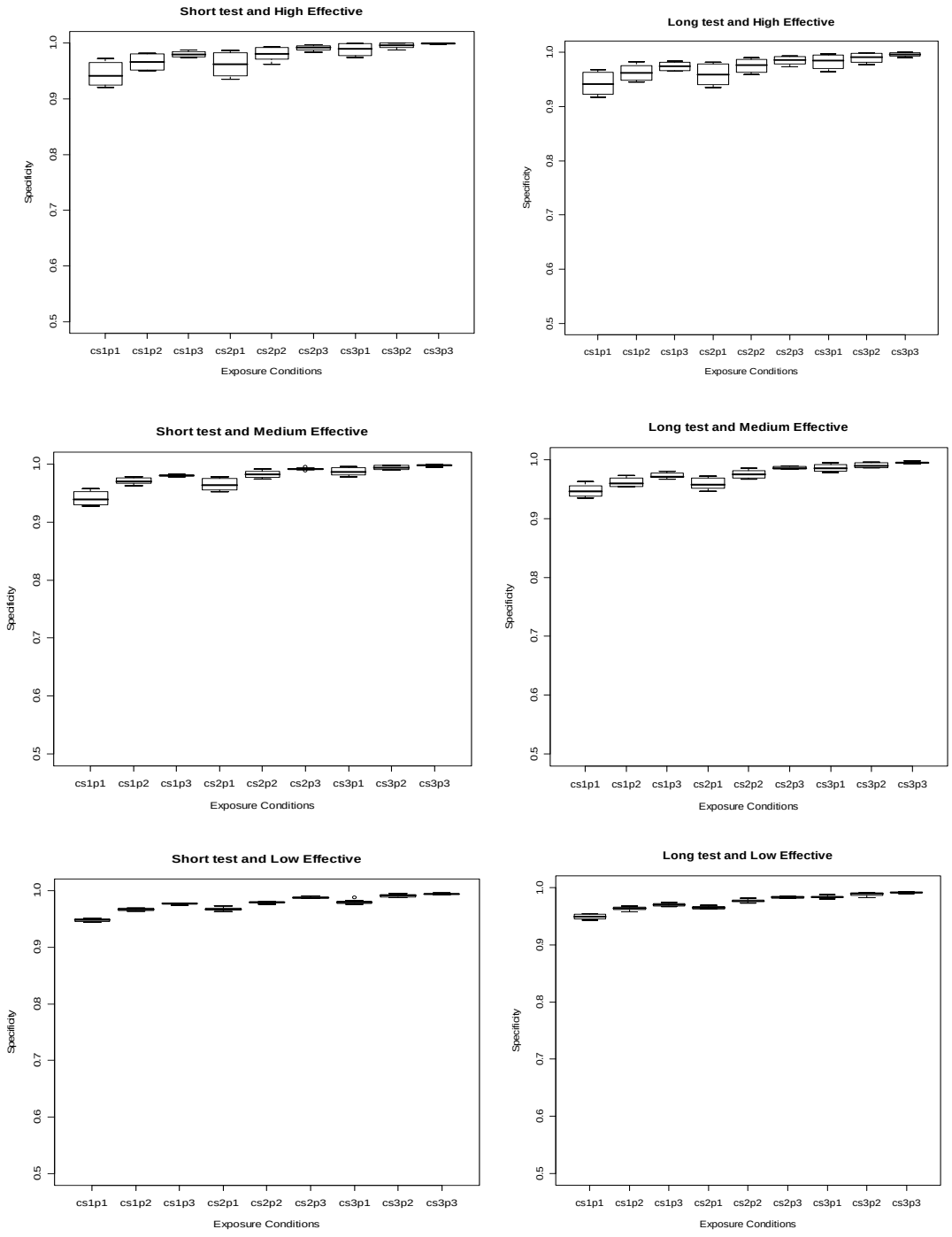


Figure 4. The Specificity across Different Conditions. “cs1”=cheating size with 5% cheaters; “cs2”=cheating size with 35% cheaters; “cs3”=cheating size with 70% cheaters. “p1”= 30% items are exposed; “p2”=50% items are exposed; “p3”=70% items are exposed; “High Effective”= high effective cheaters; “Medium Effective”= medium effective cheaters; “Low Effective”= low effective cheaters.

Model's Sensitivity

Compared to the specificity which is consistently stable across different conditions, the sensitivity of this model is greatly impacted by cheating effectiveness, cheating size, cheating degree, information and test length. Generally, the test length, cheating effectiveness, cheating degree and test information increase the model's sensitivity, but the cheating size decreases the model's sensitivity.

The Impact of Cheating Effectiveness

As stated before, the model is designed to detect effective cheaters who have noticeable score gain and might treat the low effective cheaters who have no or little score gain as innocent test takers. As a result, the sensitivity of the model to detect test cheaters should increase along with cheaters' cheating effectiveness (score gain). The results in Table 6 and Table 7 show that the model is able to detect effective cheaters but loses its power at low-effective cheaters. For example, the sensitivity of the model is 94.9 percent when the cheaters effectively cheat in a normal test with 80 items and cheating degree within 0.8 to 1, but it sharply drops to 43.4 percent when the cheaters low-effectively cheat on the exposed items. In other words, 94.9 percent of the high effective cheaters are correctly identified, but only 43.4 percent of the low effective cheaters are correctly detected by the model. The sensitivity across three different

levels of cheating effectiveness for the informative test is presented in Figure 5 and for more straightforward illustration is presented in Figure 6.

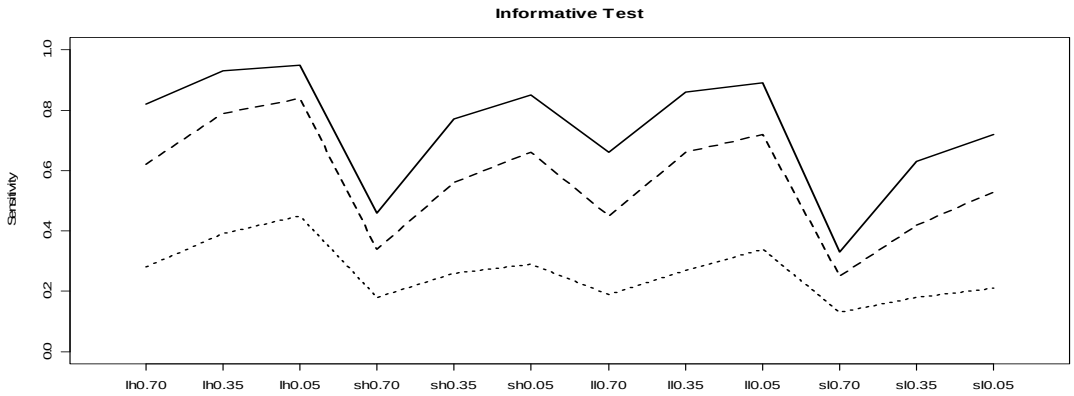


Figure 5. The sensitivity of the Informative Test. "lh0.70"= the long test, 0.8-1 cheating degree and 70% cheating size; "sh0.35"=the short test, 0.8-1 cheating degree and 35% cheating size; "ll0.70"= the long test, 0.5-0.8 cheating degree and 70% cheating size; "sl0.35"=the short test, 0.5-0.8 cheating degree and 35% cheating size.

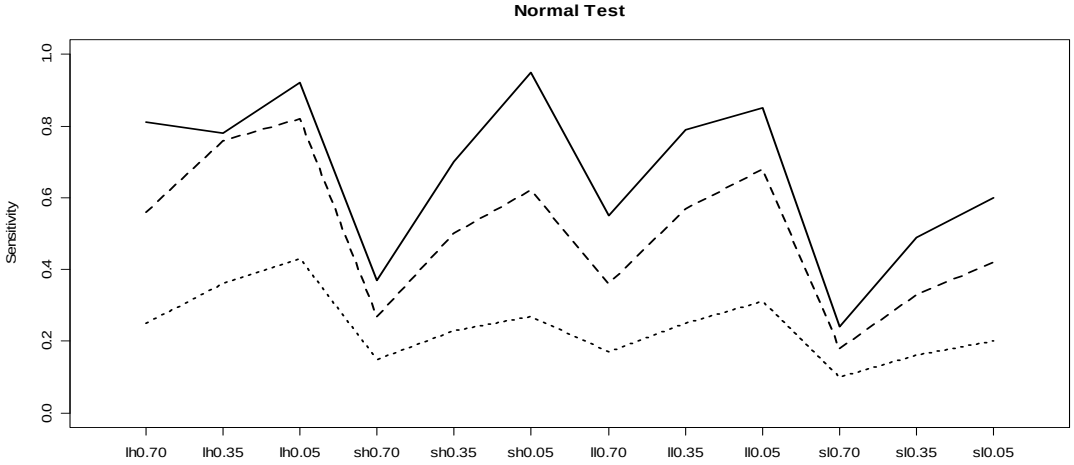


Figure 6. The sensitivity of the Normal Test. "lh0.70"= the long test, 0.8-1 cheating degree and 70% cheating size; "sh0.35"=the short test, 0.8-1 cheating degree and 35% cheating size; "ll0.70"= the long test, 0.5-0.8 cheating degree and 70% cheating size; "sl0.35"=the short test, 0.5-0.8 cheating degree and 35% cheating size.

In both Figure 5 and Figure 6, the solid line (represents the sensitivity in the high effective cheaters cases) is the highest, and the dotted line (represents the sensitivity in the low effective cheaters cases) is the lowest and the dash line (represents the sensitivity in the medium effective cheaters cases) is located at the middle. Based on the position of the three lines in Figure 5and Figure 6, high effective cheaters are mostly likely to be detected, medium effective cheaters are less likely to be identified and the low-effective cheaters are least likely to be detected.

Other than the stable specificity of the DGIRTM, another valuable characteristic of this model is that test cheaters are more likely identified by the model when their cheating activities become increasingly effective. In real settings, those effective cheaters obtains significant score gain, and such score gain will lead us to make wrong inference on the cheaters. Effective cheaters severely invalidate the inferences made based on tests. Actually, the degree of invalidation would be increasingly worse when cheating effectiveness goes up. To detect such kind of cheaters, therefore, is practically meaningful and important for stakeholders. Together with the model's stability and accuracy in specificity, the model is capable of detecting effective cheaters and makes a small degree of mistakes in cheating detection.

Impact of Test Length and Number of Exposed Items

Test length is a critical factor impacting this model's sensitivity. In this simulation study, two conditions about test length are considered: a short test with 40

items and a long test with 80 items. The average sensitivity within the short test and long test along different cheating size is presented in Table 8, and the detailed sensitivity in each condition within short/long test is plotted in Figure 4 and Figure 5.

As shown in Table 8, the average sensitivity within long test cases (represented by the second, fourth and sixth row in Table 8) is uniformly greater than the sensitivity in the corresponding conditions within short test cases(represented by the first, third and fifth row in the Table 8). Based on the results, the model is more sensitive to detect test cheaters in tests with a larger total number of items than those with a smaller total number of items.

Table 8. The Average Sensitivity by Test Length

Cheating Effective	Test length	5%			35%			70%		
		cs1p1	cs1p2	cs1p3	cs2p1	cs2p2	cs2p3	cs3p1	cs3p2	cs3p3
High Effective	S-Test	69.53%	**74.31%**	71.25%	56.28%	**63.98%**	58.34%	33.96%	**34.54%**	25.34%
	L-Test	88.64%	**90.61%**	87.69%	84.52%	**87.04%**	82.56%	65.37%	**69.87%**	61.18%
Medium Effective	S-Test	55.19%	55.34%	46.70%	41.94%	45.03%	38.28%	26.04%	25.84%	20.36%
	L-Test	73.98%	76.08%	71.05%	65.45%	69.42%	61.35%	45.95%	49.18%	42.11%
Low Effective	S-Test	28.03%	24.09%	18.98%	22.19%	20.80%	14.04%	15.06%	14.14%	9.88%
	L-Test	40.43%	38.98%	33.96%	32.07%	31.67%	26.40%	22.45%	22.17%	18.45%

cs1"=cheating size with 5% cheaters; "cs2"=cheating size with 35% cheaters; "cs3"=cheating size with 70% cheaters. "p1"= 30% items are exposed; "p2"=50% items are exposed; "p3"=70% items are exposed."S-Test"= short test with 40 items; "L-Test"=long test with 80 items.

Table 9. The Sensitivity by Test Information and Cheating Degree in High Effective Cheating

Length	Cheating degree	cs1p1	cs1p2	cs1p3	cs2p1	cs2p2	cs2p3	cs3p1	cs3p2	cs3p3
Short Test	High cheating degree	**81.30%**	**84.80%**	**81.20%**	**67.29%**	**76.63%**	**71.80%**	**41.56%**	**45.62%**	**36.80%**
		74.00%	80.60%	77.30%	60.69%	69.56%	65.13%	36.65%	37.26%	25.33%
	Low cheating degree	**66.60%**	**72.10%**	**71.80%**	**53.64%**	**62.54%**	**57.19%**	**31.37%**	**32.97%**	**25.69%**
		57.30%	60.10%	53.60%	45.46%	49.49%	40.37%	26.13%	24.14%	14.76%
Long Test	High cheating degree	**94.30%**	**94.50%**	**92.40%**	**91.46%**	**92.47%**	**90.27%**	**76.96%**	**82.01%**	**76.05%**
		92.20%	94.90%	91.80%	89.87%	91.87%	88.51%	73.29%	78.38%	70.54%
	Low cheating degree	**88.20%**	**89.30%**	**86.30%**	**82.81%**	**85.77%**	**80.89%**	**60.18%**	**65.97%**	**58.99%**
		80.90%	84.60%	81.60%	74.83%	79.07%	70.50%	51.14%	55.36%	38.19%

"cs1"=cheating size with 5% cheaters; "cs2"=cheating size with 35% cheaters; "cs3"=cheating size with 70% cheaters. "p1"= 30% items are exposed; "p2"=50% items are exposed; "p3"=70% items are exposed

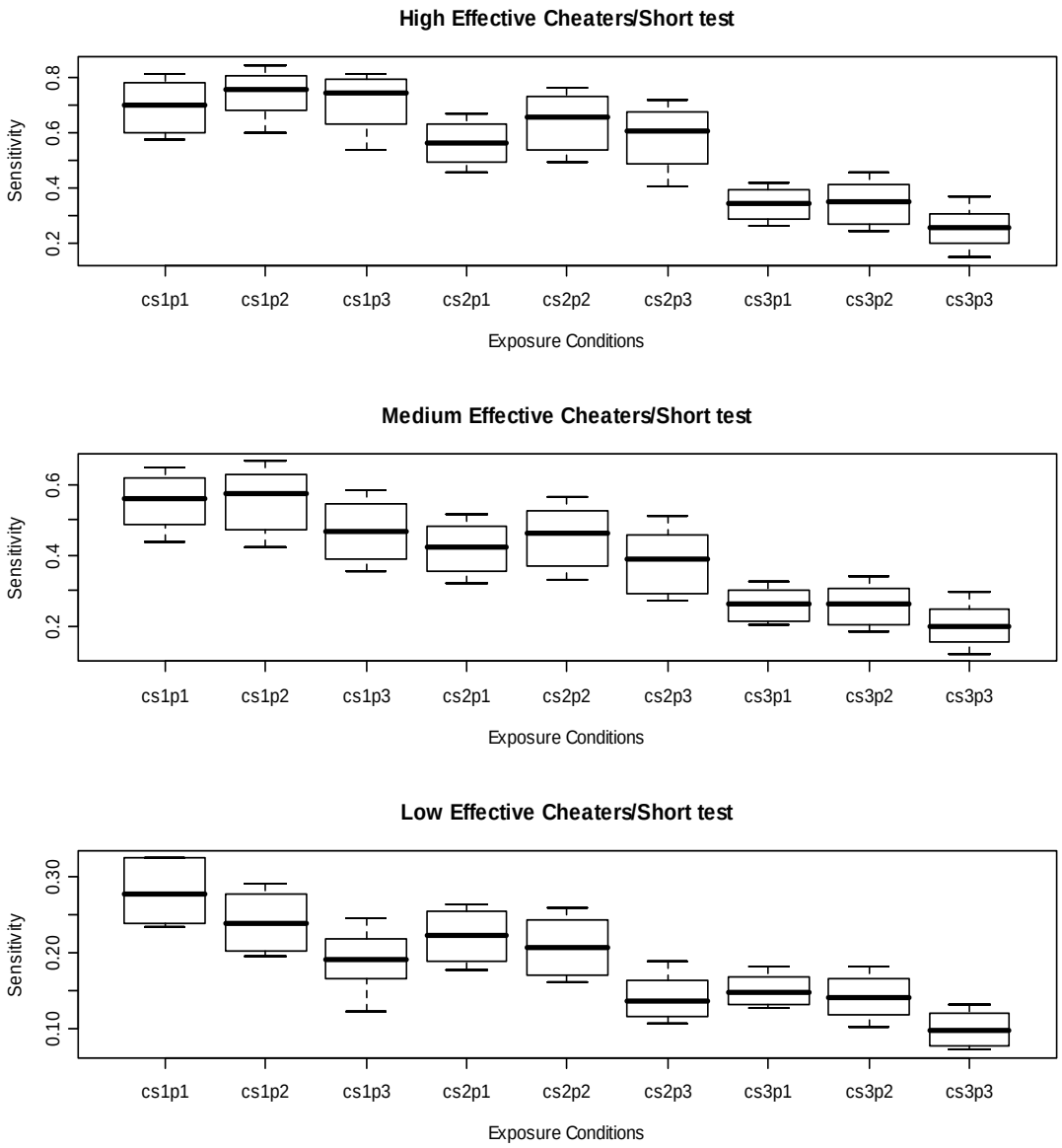


Figure 7. The Sensitivity of Short Test. “cs1”=cheating size with 5% cheaters; “cs2”=cheating size with 35% cheaters; “cs3”=cheating size with 70% cheaters. “p1”= 30% items are exposed; “p2”=50% items are exposed; “p3”=70% items are exposed.

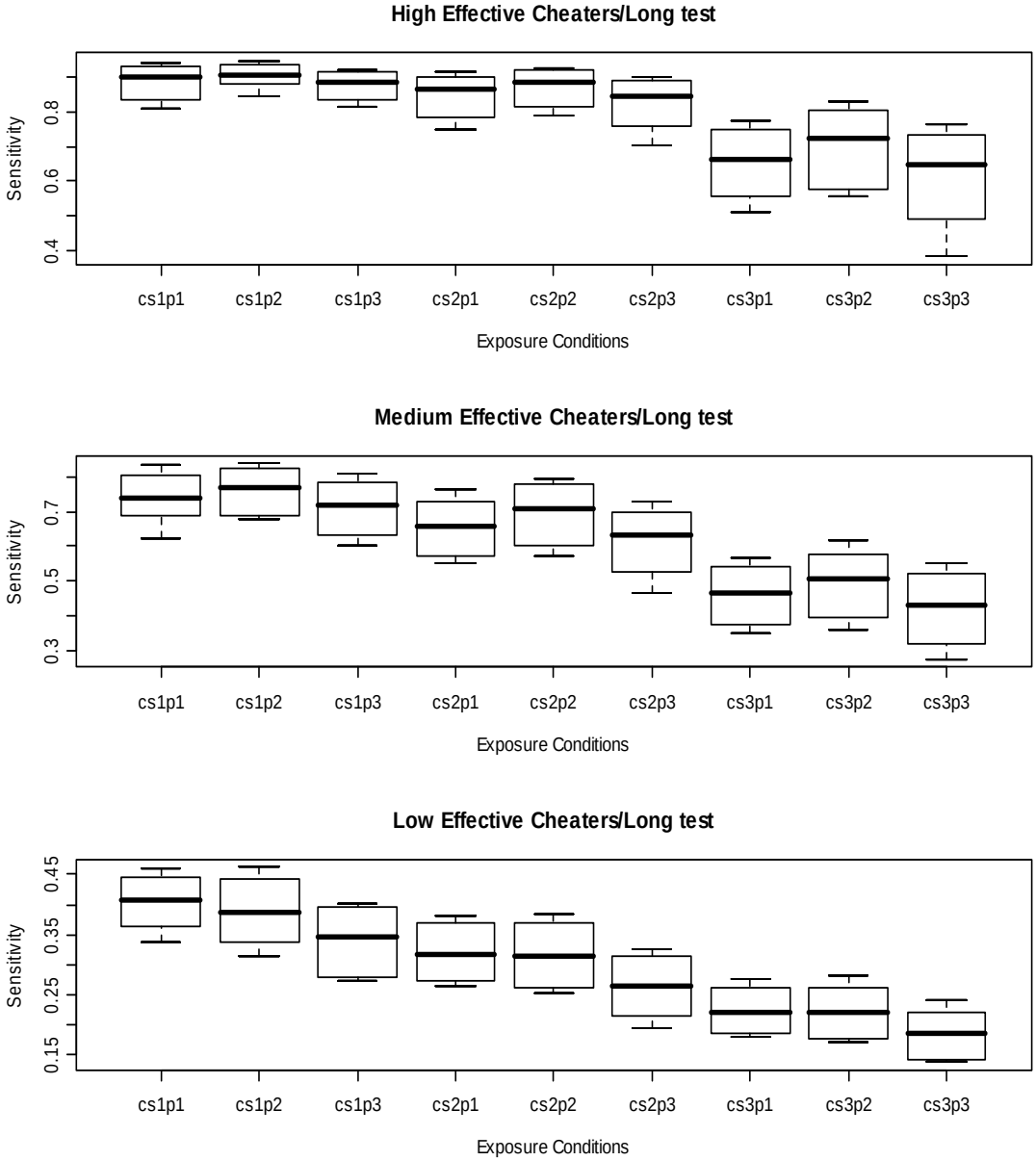


Figure 8. The Sensitivity of Long Test. “cs1”=cheating size with 5% cheaters; “cs2”=cheating size with 35% cheaters; “cs3”=cheating size with 70% cheaters. “p1”= 30% items are exposed; “p2”=50% items are exposed; “p3”=70% items are exposed.

Like Table 8, Figure 7 and Figure 8 also show that the DGIRTM is more sensitive in the long test cases than it is in the short test cases. The DGIRTM classifies items as two types (exposed and unexposed items). An increase in the total number of

items means an increase in the number of both exposed and unexposed items. The model becomes more sensitive to detect test cheaters when there is an increase in the number of both exposed items and unexposed items.

Other than the difference in the total number of exposed and unexposed items, the relationship between the number of the exposed items and that of the unexposed items also plays a critical role to impact the model's sensitivity. As exhibited by Table 8, the sensitivity of the conditions with 50 percent exposed items (which is the number of exposed items equal to that of unexposed items) is greater than the case with 30 percent exposed items where the number of exposed items is less than that of unexposed, and the case with 70 percent exposed items where the number of exposed items is greater than that of unexposed items. For example, when the cheating size is 5percent, the sensitivity of the case with 50percent exposed items in the short test for effective cheaters is 74.31 percent, which is greater than the sensitivity of the corresponding cases with 30percent exposed items (69.53 percent) and 70 percent exposed items (71.25 percent). Figure 6 is plotted to demonstrate the impact of test length (the total number of items) and relationship between the number of exposed items and that of unexposed items.

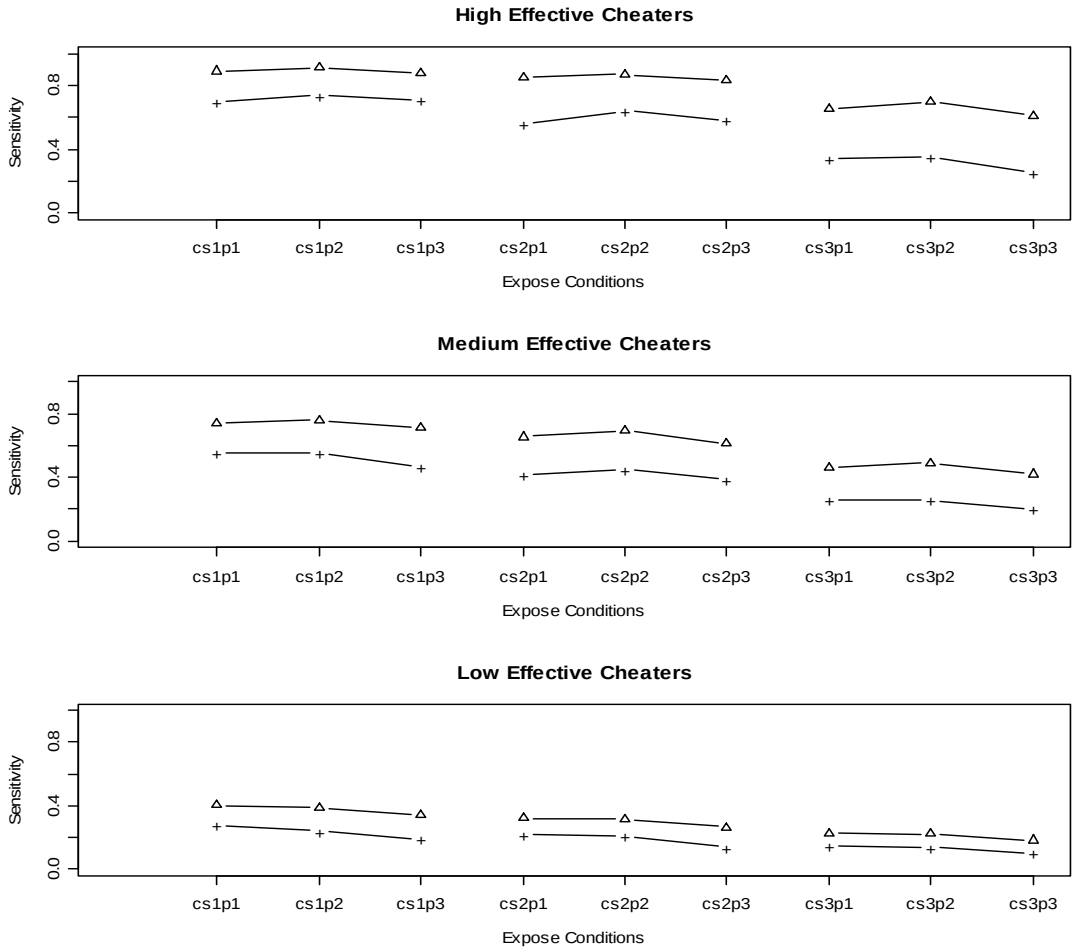


Figure 9. The Average Sensitivity by Test Length. "+"= the short test; "Δ"= the long test; "cs1"=cheating size with 5% cheaters; "cs2"=cheating size with 35% cheaters; "cs3"=cheating size with 70% cheaters. "p1"= 30% items are exposed; "p2"=50% items are exposed; "p3"=70% items are exposed.

In Figure 9, the sensitivity in the long test cases (represented by "Δ") is uniformly greater than that in the short test cases (represented by "+"). Within each cheating size (each group of three points linked together by lines), the middle points representing the cases with 50 percent exposed items are higher than the other two points which represents the 30 percent and 70 percent exposed items. This cheating model requires a decent number of both exposed and unexposed items to reliably

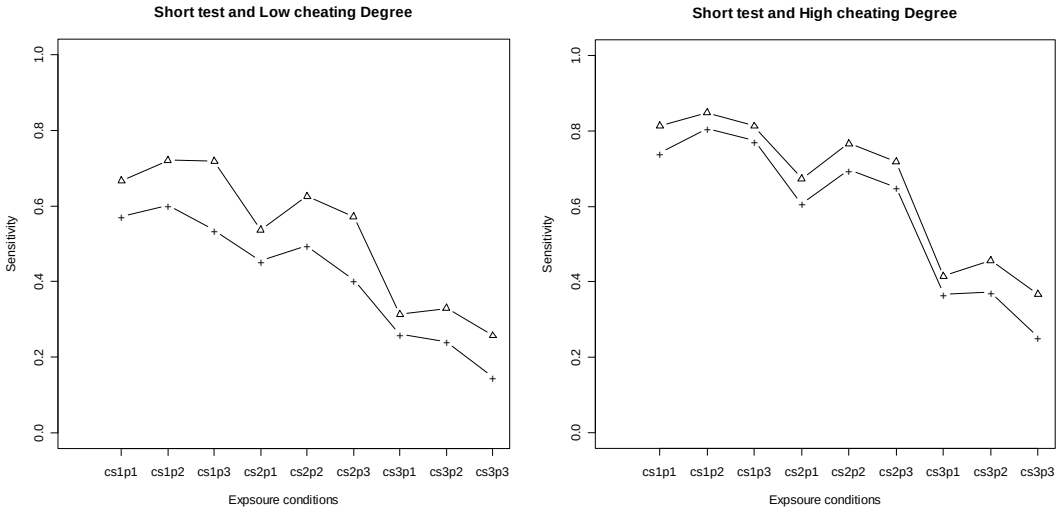
detect test cheaters. A decrease of either exposed or unexposed items could undermine this model's power to detect test cheaters, even though the test has a large total number of items. Generally, the model seems to be more sensitive when the number of exposed items is less than that of unexposed items. In contrast, the model is slightly more conservative when the number of the exposed items is greater than that of unexposed items.

As a summary, an item number increase in either exposed or unexposed items can improve the model's power to detect test cheaters. Given a certain total number of items in tests, the model would derive a greater degree of sensitivity in tests with a balanced number of exposed and unexposed items (the number of exposed items equal to that of unexposed items) than in tests with an unbalanced number of exposed and unexposed items (the number of exposed items is either greater or less than that of unexposed items).

Impact of Test Information, Cheating Degree and Cheating Size

The impact of test information, cheating degree and cheating size is discussed only in high effective cheating cases because the impact of these three factors has the same pattern in the three different level of cheating effectiveness. The sensitivity across cheating degree, information and cheating size in high effective cheating cases is presented in Table 9 and also plotted in Figure 10.

In Table 9, the sensitivity of the informative test[6] (bolded rows) is uniformly higher than that of the normal test[7] (not bolded rows) in every condition. Test information can improve the model's power to detect test cheaters. Similarly, the sensitivity of the conditions in the high cheating degree cases (exposed items being cheated by 80 percent-100 percent of the whole cheaters) is uniformly higher than that of the corresponding conditions in the low cheating degree cases (exposed items being cheated by 50 to 80 percent of the whole cheaters). An increasing cheating degree can lead to a higher degree of cheating information, and thus make the model more sensitive in terms of cheater detection. However, the model's sensitivity decreases along with the increasing cheating size. It seems that the model is not efficient to identify cheaters when a large scale of cheating activities occurs in tests, such as cheating by group.



[6] Informative test: a test with exposed item difficulty distribute with a mean equal to the mean of cheating ability and standard deviation 0.5, and unexposed item difficulty with a mean equal to the mean of true ability and standard deviation 0.5.

[7] Normal test: a test with item difficulty with a mean equal to the mean of true and cheating ability and standard deviation 1.

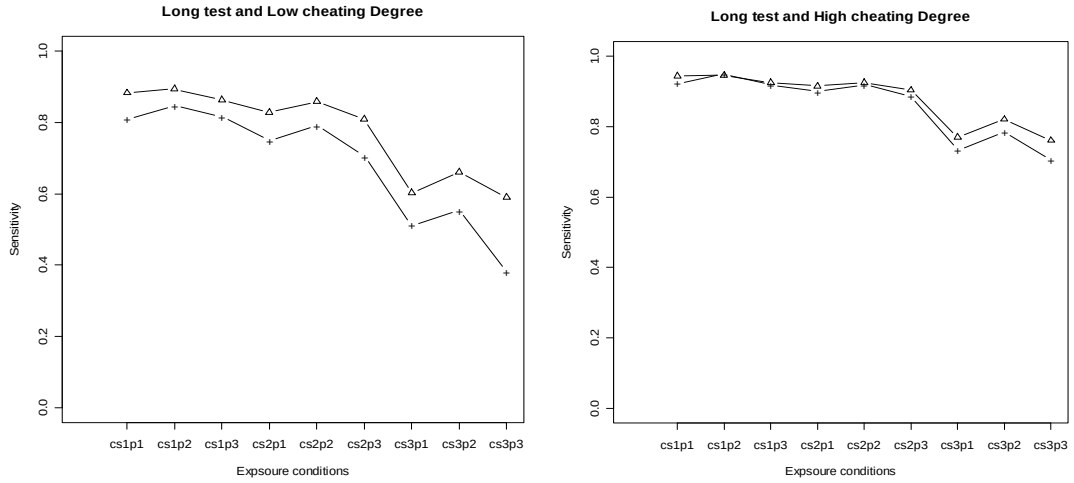


Figure 10. The Sensitivity by Test Information and Cheating Degree. "+"= the normal test; "Δ"= the informative test; "cs1"=cheating size with 5% cheaters; "cs2"=cheating size with 35% cheaters; "cs3"=cheating size with 70% cheaters. "p1"= 30% items are exposed; "p2"=50% items are exposed; "p3"=70% items are exposed.

As shown in Figure 10, the sensitivity of the high cheating degree cases (the two plots on the right hand) is greater than that of the low cheating degree cases (the two plots on the left hand). The decreasing line means that the sensitivity decreases along with the increasing cheating size. The line represented by "Δ" is always higher than the line represented by "+", which implies that the sensitivity of the informative test is uniformly higher than that of the normal test.

The informative test with targeted location and higher degree of information amount enables the model to derive a greater level of accuracy relative to cheating identification. As a note, such feature is practically important, especially for the computer adaptive tests (CAT). In CAT context, items are purposely selected according to the average of examinee ability, which results in the selected set of items have a mean equal to examinee's ability and small standard deviation. In other words,

the model might be highly useful in monitoring test cheating relative to item exposure in CAT settings.

Like the cheating effectiveness, cheating degree is also a critical factor to impact the DGIRTM's cheating sensitivity. The DGIRTM's sensitivity grows when cheating degree increases. In other words, the cheaters with a higher level of cheating degree are more likely to be identified by the DGIRTM.

As opposed to the positive impact of test information and cheating degree on the model's sensitivity, the cheating size has a negative impact on the DGIRTM's sensitivity. The impact of the cheating size should be due to the scale shift of the exposed items. Under the DGIRTM, the scale for the estimated item difficulty and examinee score is expected to be determined by the unexposed items. The scale determined by the unexposed items is the scale of the true ability for all examinees taking tests and items. The score difference is obtained based on this scale, by comparing cheaters' cheating ability and their true ability. However, when the cheating size increases, the true ability scale for all the examinees is hard to align on the scale determined by unexposed items, but moves forward to a mixed scale determined by both the exposed and unexposed items instead. The scale shift moves further towards to the mixed scale determined by both the exposed and unexposed items when the cheating size goes up. This shift apparently appears with a cheating size above 50 percent. As a result, the score difference between the exposed items and unexposed items are reduced because the true ability for all examinees moves closer

to cheaters the cheating ability, and thus the sensitivity of the DGIRTM model correspondingly decreases when the cheating size

Model Estimation Accuracy

The DGIRTM provides both cheaters and non-cheaters true ability estimation as well as cheaters' score gain, which helps stake-holders (e.g., teachers or universities) to make correct inference about the students' real knowledge level and cheating degree. The ability to distinguish cheaters' true ability and cheating skill is another unique advantage of the DGIRTM. The average Root Mean Square Difference (RMSD) of item difficulty and examinees' true ability in the high effective cases is presented in Table 10. The RMSD of item difficulty in all the conditions in high effective cheating cases is plotted in Figure 8 to demonstrate the variance of RMSD. As a note, the accuracy of the model estimation is only discussed within the high effective cheating cases, because the estimation accuracy in the high effective cheating is of the same degree as the other two levels of cheating effectiveness.

Table 10. The Average RMSD of Item Difficulty by Test Length in the High Effective Cheating

Type	Test length	5%			35%			70%		
		cs1p1	cs1p2	cs1p3	cs2p1	cs2p2	cs2p3	cs3p1	cs3p2	cs3p3
Item Difficulty	Short Test	0.16	0.20	0.22	0.18	0.17	0.17	0.38	0.40	0.41
	Long Test	**0.20**	**0.23**	**0.25**	**0.13**	**0.14**	**0.14**	**0.29**	**0.31**	**0.36**
True Ability	Short Test	0.33	0.36	0.39	0.37	0.40	0.45	0.49	0.53	0.61
	Long Test	**0.23**	**0.25**	**0.27**	**0.22**	**0.24**	**0.27**	**0.28**	**0.30**	**0.36**

"cs1"=cheating size with 5% cheaters; "cs2"=cheating size with 35% cheaters; "cs3"=cheating size with 70% cheaters. "p1"= 30% items are exposed; "p2"=50% items are exposed; "p3"=70% items are exposed.

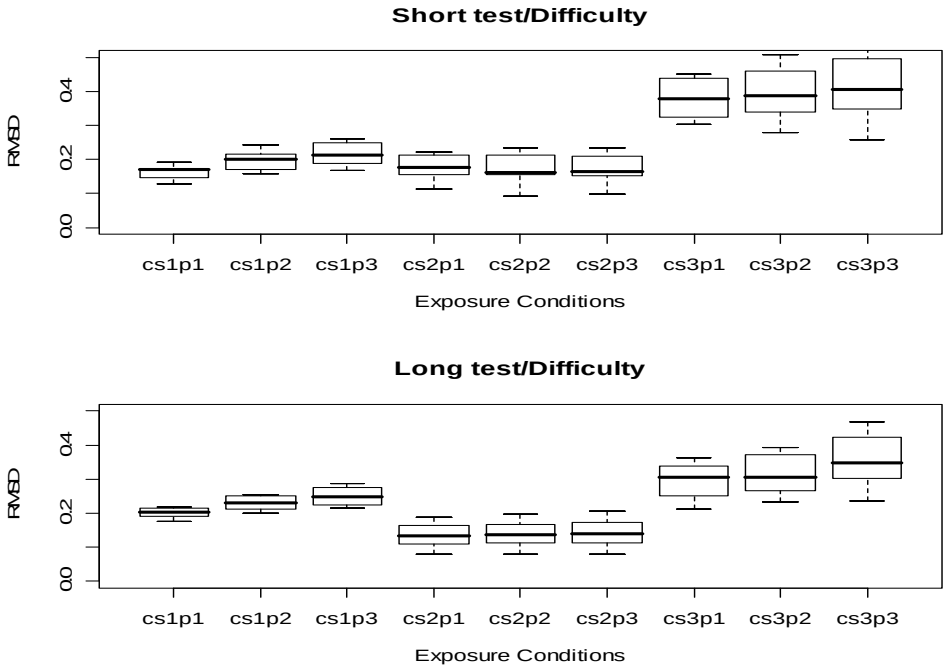


Figure 11. RMSD of Item Difficulty. "cs1"=cheating size with 5% cheaters; "cs2"=cheating size with 35% cheaters; "cs3"=cheating size with 70% cheaters. "p1"= 30% items are exposed; "p2"=50% items are exposed; "p3"=70% items are exposed.

In Table 10, the RMSD in the long test cases is greater than the RMSD in the corresponding cases in the short test. The RMSD in the 70 percent cheating size is the greatest (i.e., the maximum RMSD of item difficulty is 0.36) and that in the 35 percent cheating size is the smallest (i.e., the minimum RMSD is 0.13). Figure 11 also shows that the variance of the RMSD in the 70 percent cheating size case is noticeably greater than that in the other two cheating size cases, but the variance of the RMSD in the 5 percent and 35 percent cheating size cases is relatively small and stable. The RMSD of examinee score in all conditions is plotted in Figure 12.

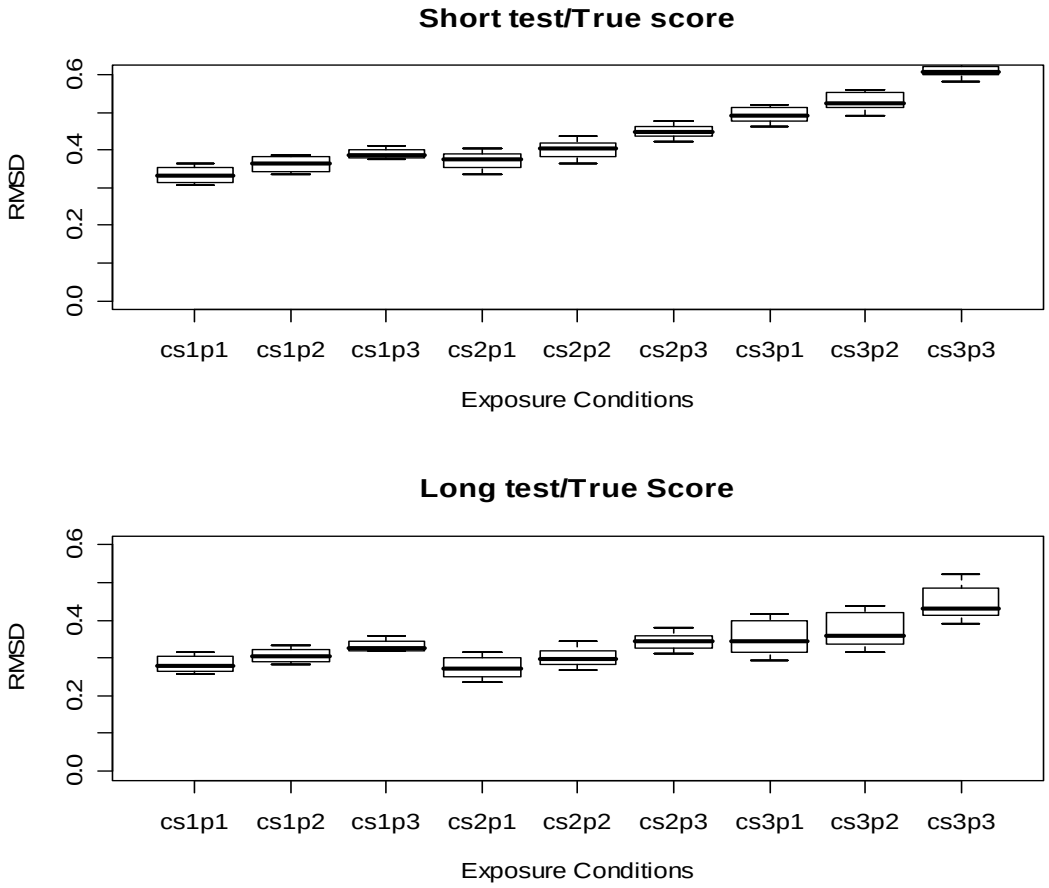


Figure 12. RMSD of True Proficiency. "cs1"=cheating size with 5% cheaters; "cs2"=cheating size with 35% cheaters; "cs3"=cheating size with 70% cheaters. "p1"= 30% items are exposed; "p2"=50% items are exposed; "p3"=70% items are exposed.

Similar to the item difficulty, the RMSD of examinee score increases along with the increasing cheating size. It reaches its highest point when cheating size is 70 percent. Within each cheating size, the RMSD when the proportion of the exposed items is 70 percent reaches its greatest point. Unlike the variance of the RMSD of item difficulty, the variance of the RMSD of the true proficiency in different cases is constantly small in different exposure conditions.

The increasing RMSD of both item difficulty and examinee score along with the increasing cheating size can serve as a strong evidence of the scale shift problem,

which is discussed in the impact of the cheating size. The increasing RMSD partly explains why the sensitivity drops along with the increasing cheating size. The RMSD of both item difficulty and true ability in the long test (the second and fourth row in Table 10) is greater than that of the corresponding cases in the short test (the first and third rows in Table 10), which explains why the sensitivity of the cheating model in the long test is greater than that in the short test.

The RMSD of true ability is greater than that of item difficulty in this simulation study, which is partly impacted by the simulation design. Because 60 percent of the 2000 examinees are purposely selected with a true ability less than -0.5, 30 percent of them are between -0.5 and 0.5 and 10 percent of them are above 0.5, which leads to the examinees being distributed with a mean -0.3. However, the prior of the true ability is set as standardized normal distribution with a mean 0. The gap between examinees' simulated distribution and the prior contributes to the greater level of RMSD of the true ability as well as the cheating ability.

As a complement to describe the model's estimation accuracy, the correlation between the estimated values and their true values are plotted in Figure 13 for both item difficulty and true ability. With respect to item difficulty, the correlation is stably as high as one with small variance in the cases with 5 percent and 35 percent cheating size, however, the correlation sharply drops and variance greatly increases in the case with 70 percent cheating size. With respect to the correlation of true ability, it is less than that of item difficulty with greater variance. In the case with 70 percent cheating

size, the correlation reaches its lowest among all the three cheating sizes. The lowest correlation between the estimated values and true values in the 70 percent cheating size cases is a strong evidence to indicate that the scale shift problem does exist and become increasingly apparent with increasing cheating size.

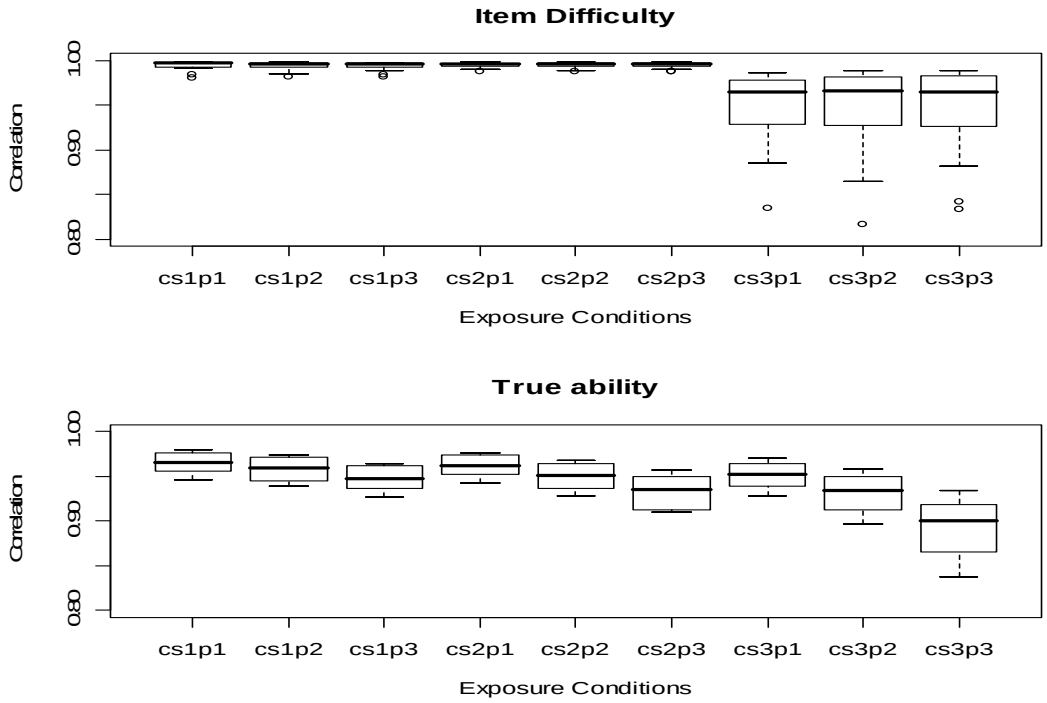


Figure 13. The Correlation between the True and Estimated Values. “cs1”=cheating size with 5% cheaters; “cs2”=cheating size with 35% cheaters; “cs3”=cheating size with 70% cheaters. “p1”= 30% items are exposed; “p2”=50% items are exposed; “p3”=70% items are exposed.

In summary, the estimation error increases in both the item and true ability when the cheating size increases. When cheating size is less than 70 percent, the RMSD of both item difficulty and true ability is maintained at a low level, however, it sharply goes up when cheating size is at 70 percent level. The correlation similarly has the opposite directionality linked with increasing cheating size. Altogether, the DGIRTM

does not perform as well as expected and has low levels of power to detect cheaters and unacceptable model estimation error when cheating size is at 70 percent level. The increasing estimation error in both item difficulty and true ability confirms that the model's scale drifts away from its original scale of true ability due to the selection of a prior distribution towards to the scale determined by the cheating ability. The increasing estimation error serves as one of key causes of the decreasing sensitivity along with the increasing cheating size. However, a greater test-length can improve the DGIRTM's estimation accuracy resulting in a greater level of sensitivity which somewhat remedies the negative impact of the scale drift. As a note, the scale shift problem can be fixed by improving the DGIRTM's estimation algorithm, which will be briefly discussed in the following section.

Comparison of DGIRTM with *lz* Index and t-test

Comparison with *lz* Index

The *lz* index, as one of the best person-fit indices to detect unusual response patterns, is conducted to serve as a baseline for comparison to illustrate the model's improvement in detecting test cheating. The sensitivity of the *lz* index and the DGIRTM in the set of short test conditions is plotted in Figure 14 and the set of long test conditions is plotted in Figure 15. In Figure 14 and Figure 15, the sensitivity of the cheating model is represented by " Δ " and that of the *lz* index is represented by "+". There are totally 16 lines in these two figures, where eight lines are formed by the cheating model by three full crossed conditions (two levels of standard deviation

of item difficulty, two locations of the mean of item difficulty and two levels of cheating degree for exposed items).

Like the DGIRTM, the sensitivity of the *lz* index is impacted by the cheating effectiveness and cheating size. The sensitivity of the *lz* index increases when the cheating effectiveness grows, and its sensitivity drops down when the cheating size increases. In Figure 14, the *lz* index has its greatest sensitivity in the 5 percent cheating size and high-effective cheating cases. Its sensitivity is almost at 0 in all the three levels of cheating effectiveness when the cheating size is 70 percent, and in the low-effective cheating cases, its sensitivity is close to 0 at all the three levels of cheating size.

Compared to the DGIRTM, the *lz* index seems to be less sensitive than the DGIRTM. The DGIRTM exhibits a greater degree of robustness over the cheating size and cheating effectiveness.

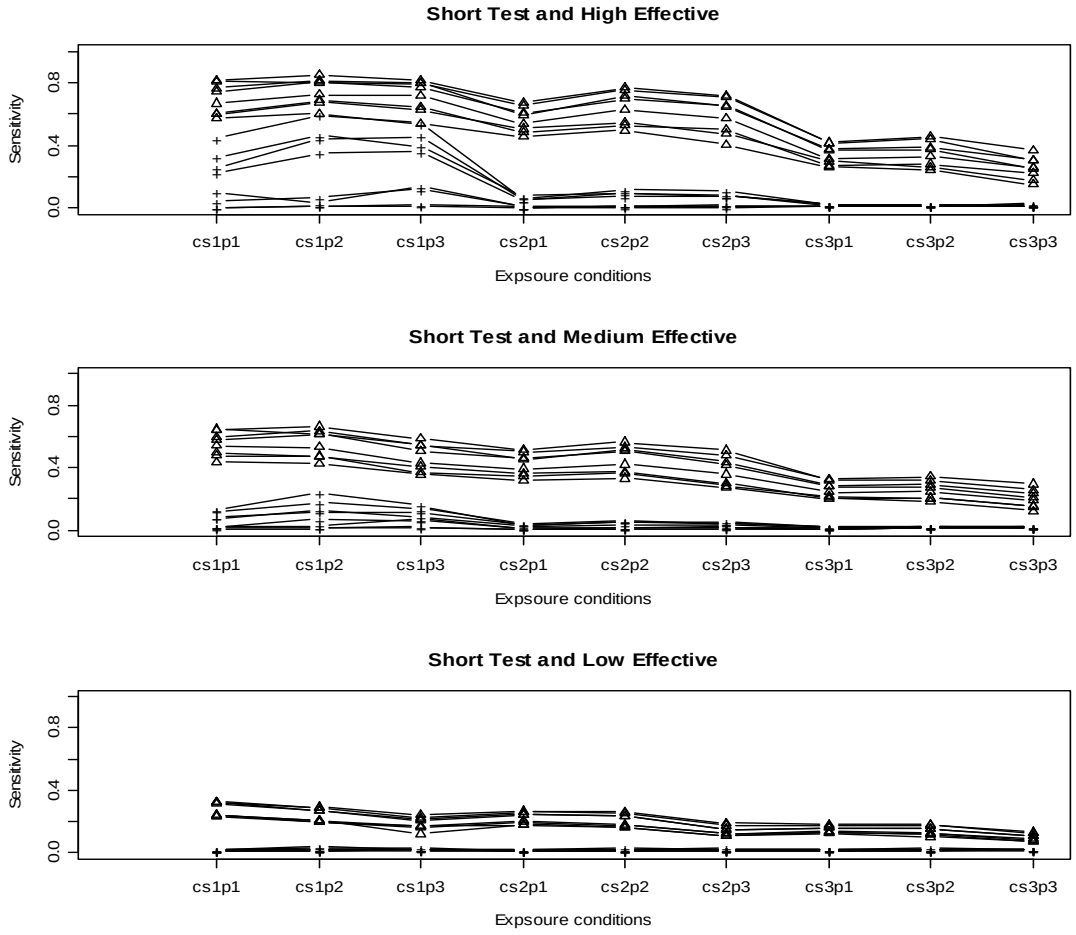


Figure 14. The Sensitivity of the *lz* and the Cheating Model in Short Test. "Δ"=the sensitivity of the cheating model; "+"= the sensitivity of *lz* index; "cs1"=cheating size with 5% cheaters; "cs2"=cheating size with 35% cheaters; "cs3"=cheating size with 70% cheaters. "p1"= 30% items are exposed; "p2"=50% items are exposed; "p3"=70% items are exposed.

The *lz* index in the long test cases exhibits a similar pattern as it exhibits in the short test cases, as shown in Figure 15. Its sensitivity drops when the cheating size decreases or the cheating effectiveness drops. Especially, when the cheating size is 70 percent or when the cheating activities are low effective, the *lz* index has no power to detect cheating. Although the *lz* index achieves a higher level of sensitivity in the long test than it does in the short test, the DGIRTM has a greater sensitivity level than the *lz* index in long test cases.

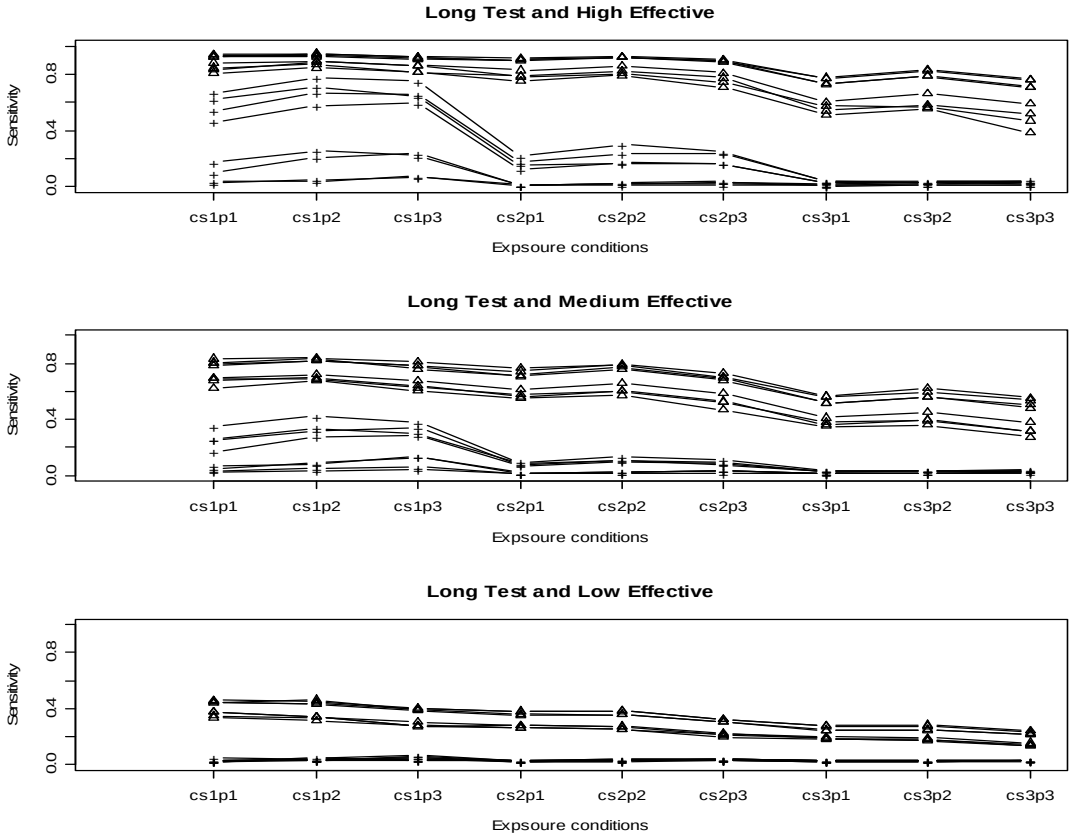


Figure 15. The Sensitivity of the *lz* and the Cheating Model in Long Test of Delta 1. "Δ"=the sensitivity of the cheating model; "+"= the sensitivity of *lz* index; "cs1"=cheating size with 5% cheaters; "cs2"=cheating size with 35% cheaters; "cs3"=cheating size with 70% cheaters. "p1"= 30% items are exposed; "p2"=50% items are exposed; "p3"=70% items are exposed

Unlike the sensitivity, the *lz* index has as a high level of specificity as the DGIRTM. The specificity of both the *lz* index and the DGIRTM in the set of short test conditions is plotted in Figure 16 and that in the set of long test conditions is plotted in Figure 17, where " Δ " represents the specificity of the cheating model and "+" represents the specificity of the *lz* index.

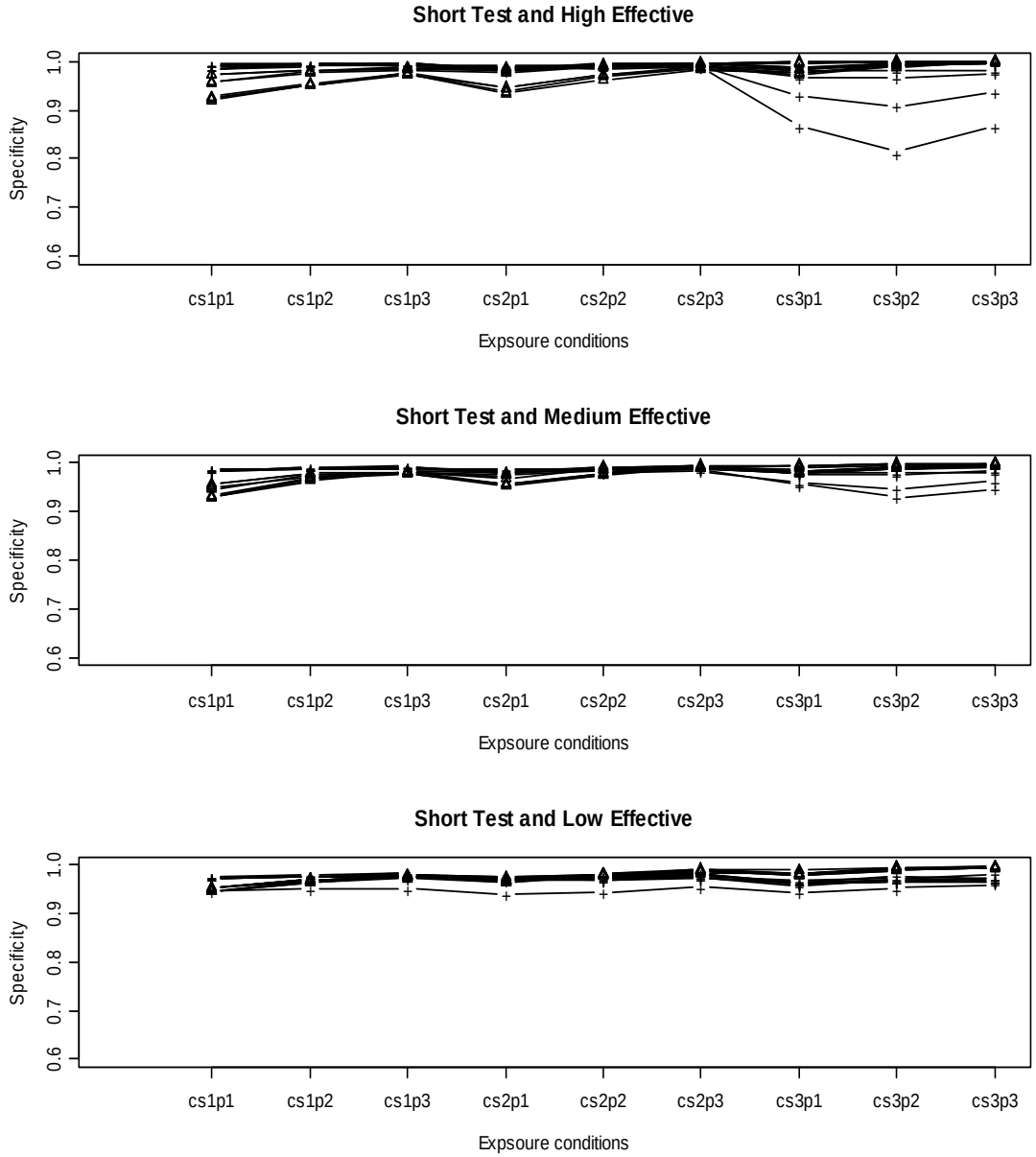


Figure 16. The Specificity of the *lz* and the Cheating Model in Short Test of Delta 1. "Δ"=the sensitivity of the cheating model; "+"= the sensitivity of *lz* index; "cs1"=cheating size with 5% cheaters; "cs2"=cheating size with 35% cheaters; "cs3"=cheating size with 70% cheaters. "p1"= 30% items are exposed; "p2"=50% items are exposed; "p3"=70% items are exposed.

As shown in Figure 13, the *lz* index and the DGIRTM do not exhibit much difference. However, the specificity of the DGIRTM shows a higher level of stability than the *lz* index. For instance, the specificity of the *lz* index has some sharply drops

in the high effective cheating cases when the cheating size is 70 percent.

Similarly, the *lz* index in Figure 17 is able to maintain its specificity at the same level as the DGIRTM does. However, the specificity of the *lz* index exhibits some sharp drops in both high effective and medium effective cheating cases when the cheating size is 70 percent. Compared to specificity in the short test cases, the specificity in the long tests is slightly less stable in the long test cases.

As a summary, the sensitivity of the *lz* index is strongly impacted by the cheating effectiveness and cheating size. It seems to lose its power to detect cheating activities when the cheating size is around or above 35 percent and the cheating activities is at the medium or low level effective. However, the *lz* index is able to maintain a high level of specificity (above 95 percent) in most of the conditions, but its specificity exhibits unusual drops in some cases. Although the test length could increase the power of the *lz* index to detect test cheating, it also undermines the stability of the specificity of the *lz* index. Compared to the *lz* index, the DGIRTM is better in terms of both the sensitivity and specificity.

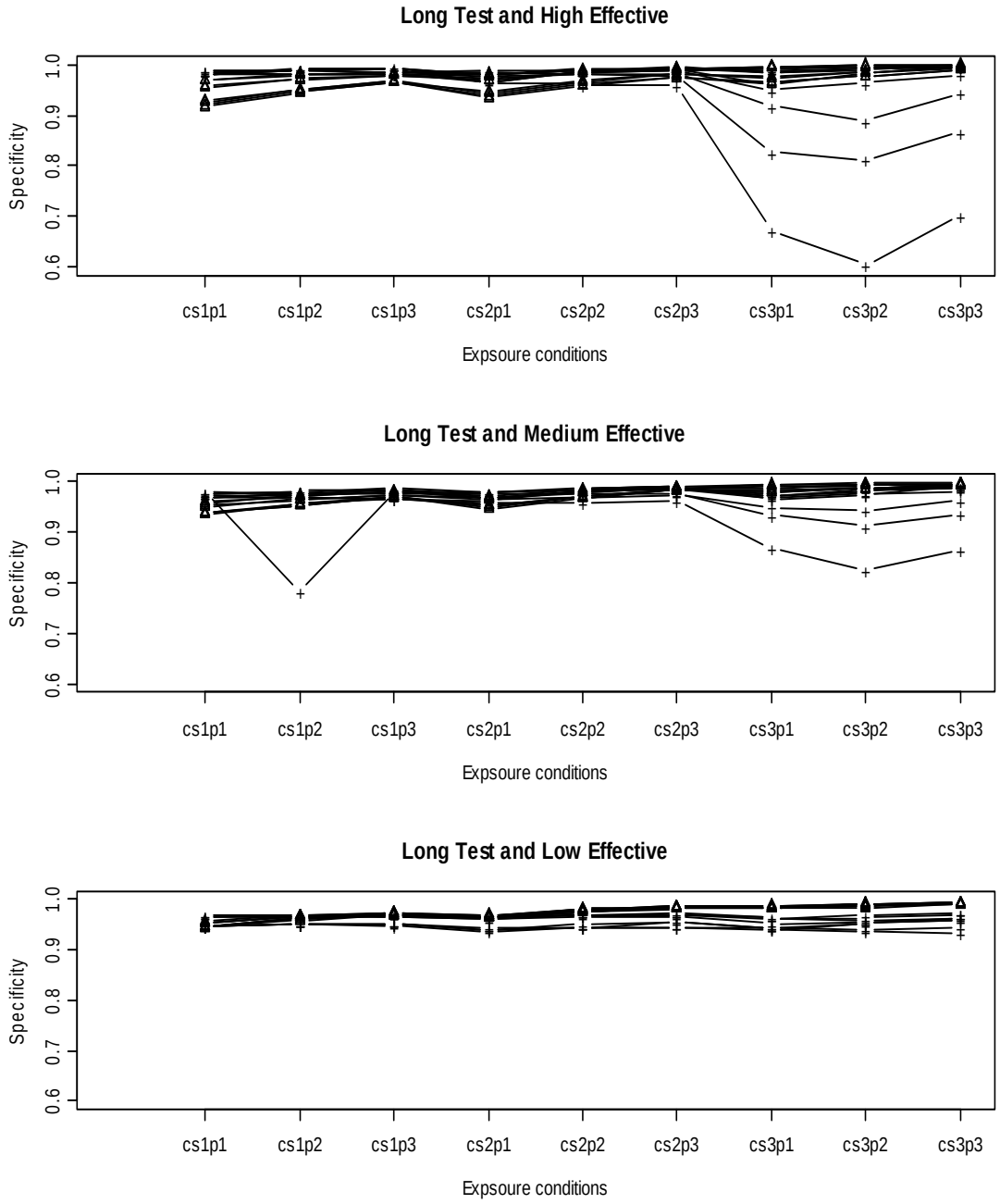


Figure 17. The Specificity of the *lz* and the Cheating Model in Long Test of Delta 1. "Δ"=the sensitivity of the cheating model; "+"= the sensitivity of *lz* index; "cs1"=cheating size with 5% cheaters; "cs2"=cheating size with 35% cheaters; "cs3"=cheating size with 70% cheaters. "p1"= 30% items are exposed; "p2"=50% items are exposed; "p3"=70% items are exposed.

Comparison with a t-test for MLEs (Exposed and Non-exposed Items)

The t-test is designed based on the two sets of items (exposed and unexposed items). As opposed to the cheating model, the t-test is simple and well-accepted by practitioners. If the t-test could derive a better sensitivity and specificity, it would be redundant for us to develop and conduct such a complex cheating model analysis. The sensitivity of the t-test and the cheating model in the set of short test conditions is plotted in Figure 18 and its corresponding specificity is plotted in Figure 19, where " Δ " still represents the specificity of the cheating model and "+" represents the specificity of the t-test.

In Figure 18, the sensitivity of the t-test decreases when the cheating size increases, but its sensitivity does not experience a sharp drop when the cheating effectiveness decreases. The t-test's sensitivity is uniformly greater than that of the DGIRTM in all the conditions. However, the t-test is not as effective as the cheating model in terms of the specificity.

The t-test's specificity is not stable, as shown in Figure 19, across difference cheating sizes. For example, its specificity is approximately 60 percent when the cheating size is 5 percent and it reaches approximately 90 percent when the cheating size is 70 percent. In addition, the cheating effectiveness also greatly impacts its specificity. For example, its specificity is almost as high as that of the cheating model in the 70 percent cheating size cases when the cheating is high effective, but its

specificity is uniformly less than that of the cheating model in all the conditions when the cheating is at low effective level.

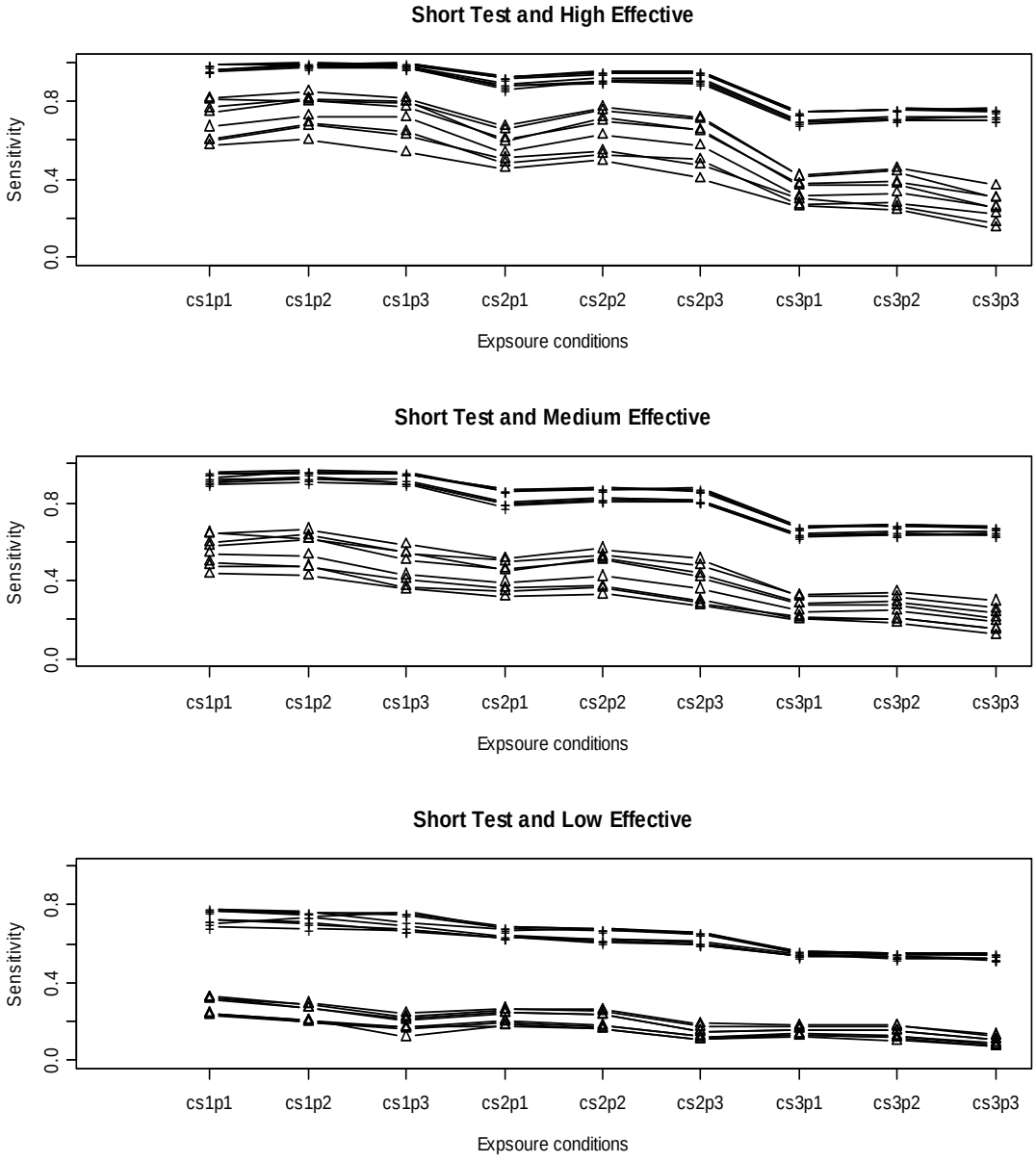


Figure 18. The Sensitivity of the t-test and Cheating Model in Short Test. "Δ"=the sensitivity of the cheating model; "+"= the sensitivity of the t-test; "cs1"=cheating size with 5% cheaters; "cs2"=cheating size with 35% cheaters; "cs3"=cheating size with 70% cheaters. "p1"= 30% items are exposed; "p2"=50% items are exposed; p3"=70% items are exposed.

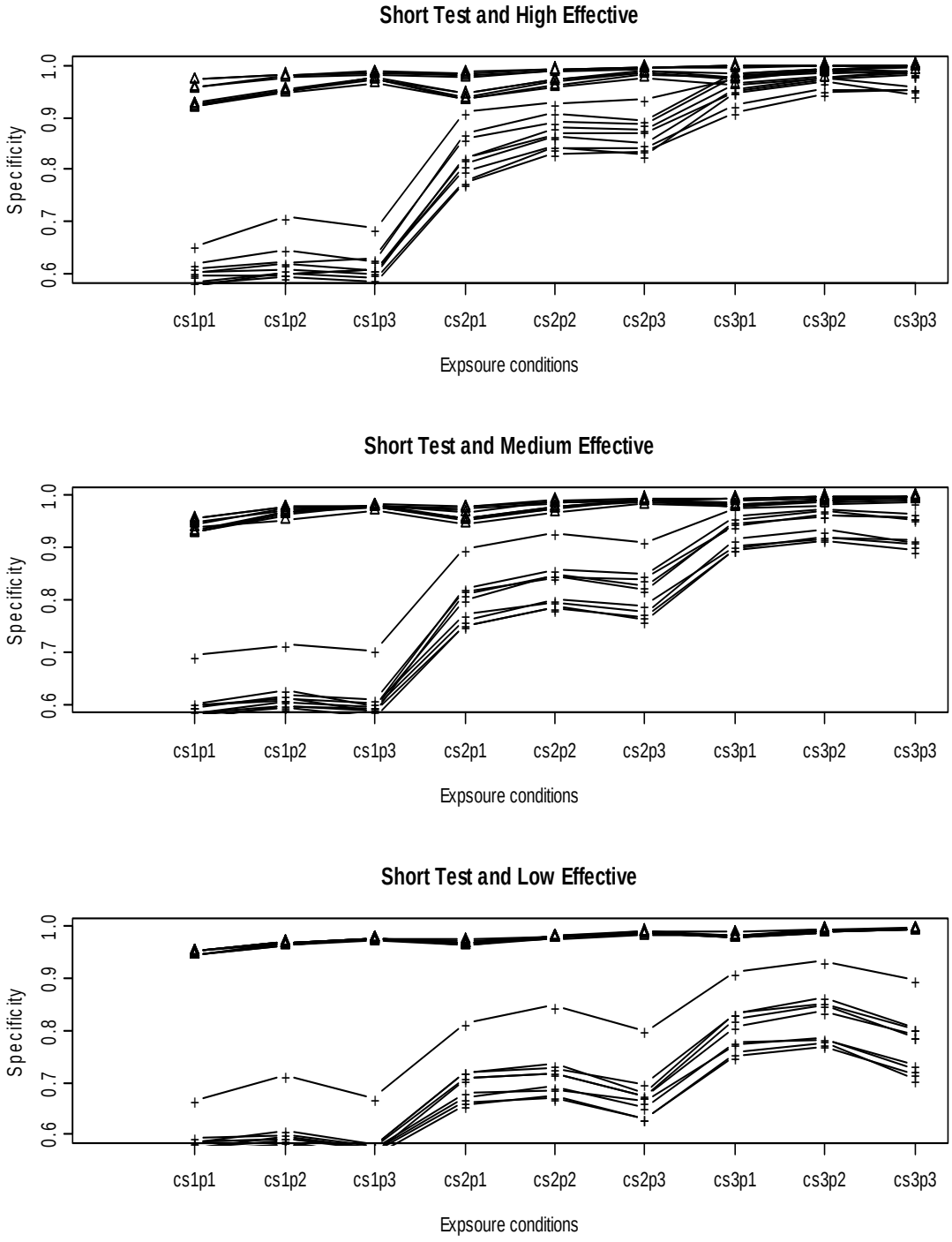


Figure 19. The Specificity of the t-test and Cheating Model in Short Test. "Δ"=the specificity of the cheating model; "+"= the specificity of the t-test; "cs1"=cheating size with 5% cheaters; "cs2"=cheating size with 35% cheaters; "cs3"=cheating size with 70% cheaters. "p1"= 30% items are exposed; "p2"=50% items are exposed; "p3"=70% items are exposed.

The sensitivity of the t-test and the cheating model is plotted in Figure 20 for the set of conditions in long test and the corresponding specificity is shown in Figure 21. The t-test still outperforms the cheating model in terms of the sensitivity, and the degree of outperformance is much less than that in the short test. Similarly, the cheating model is still better than the t-test in terms of the specificity, and the degree of difference between the cheating model's specificity and t-test's specificity is less than that in the short test cases.

Although the t-test is a simple significance test, it is effective at detecting test cheaters. However, it is unable to maintain a constant high level of specificity. Specifically, its sensitivity is also impacted by the cheating size and cheating effectiveness. An increasing cheating size or a decreasing cheating effectiveness results in a less sensitive t-test. The t-test is able to derive a higher level of specificity when cheating size grows, but a lower level of specificity when cheating effectiveness drops.

As compared to the DGIRTM, the t-test is better in terms of the sensitivity, but worse in terms of the specificity. From an ethical and practical standpoint, having a low rate of false positives should rank as the first priority and should be the first principle for all practitioners in preventing or detecting test cheating. Based on this principle, the cheating detection model shows extreme promise for application in real settings across many conditions.

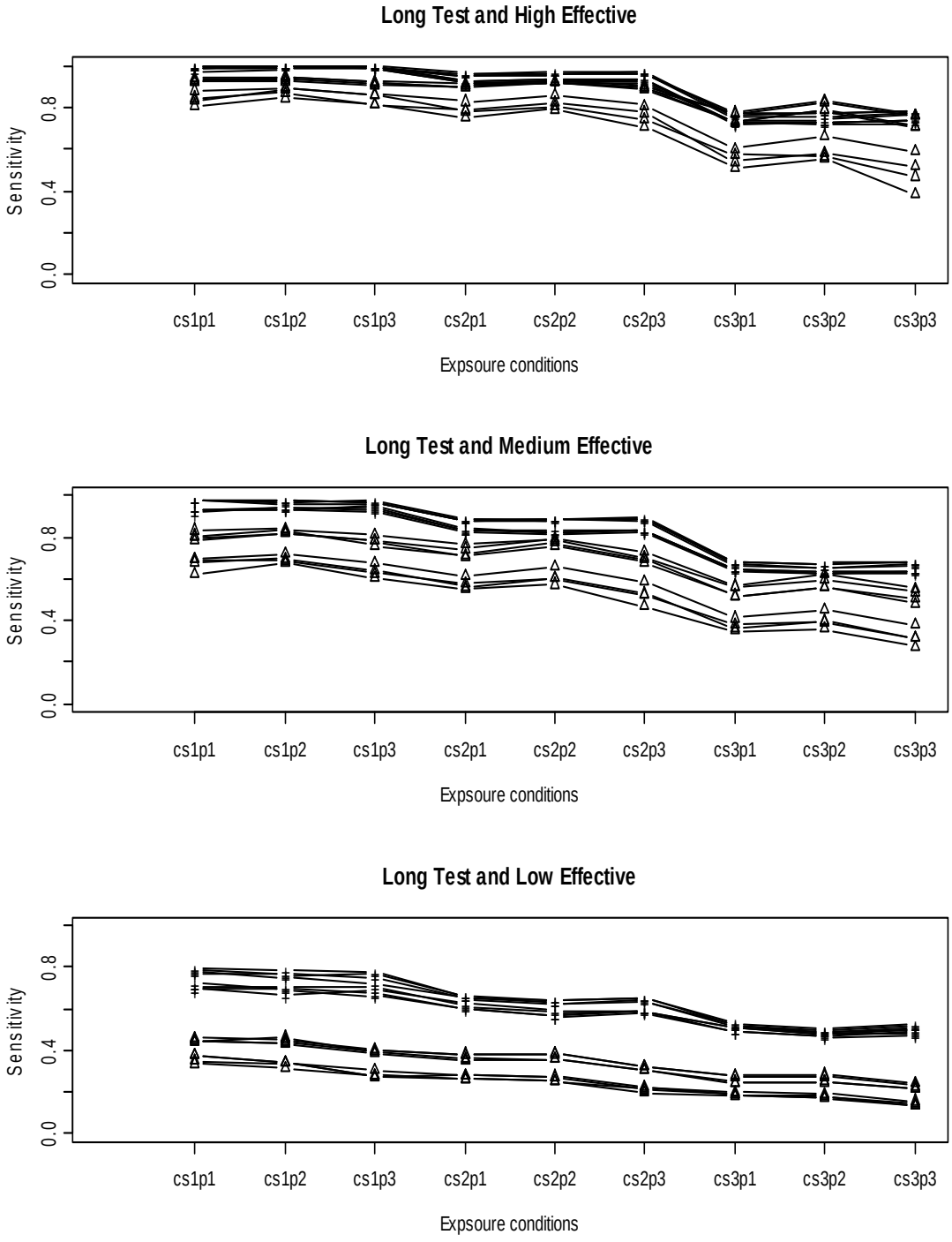


Figure 20. The Sensitivity of the t-test and Cheating Model in Long Test. "Δ"=the sensitivity of the cheating model; "+"= the sensitivity of the t-test; "cs1"=cheating size with 5% cheaters; "cs2"=cheating size with 35% cheaters; "cs3"=cheating size with 70% cheaters. "p1"= 30% items are exposed; "p2"=50% items are exposed; "p3"=70% items are exposed.

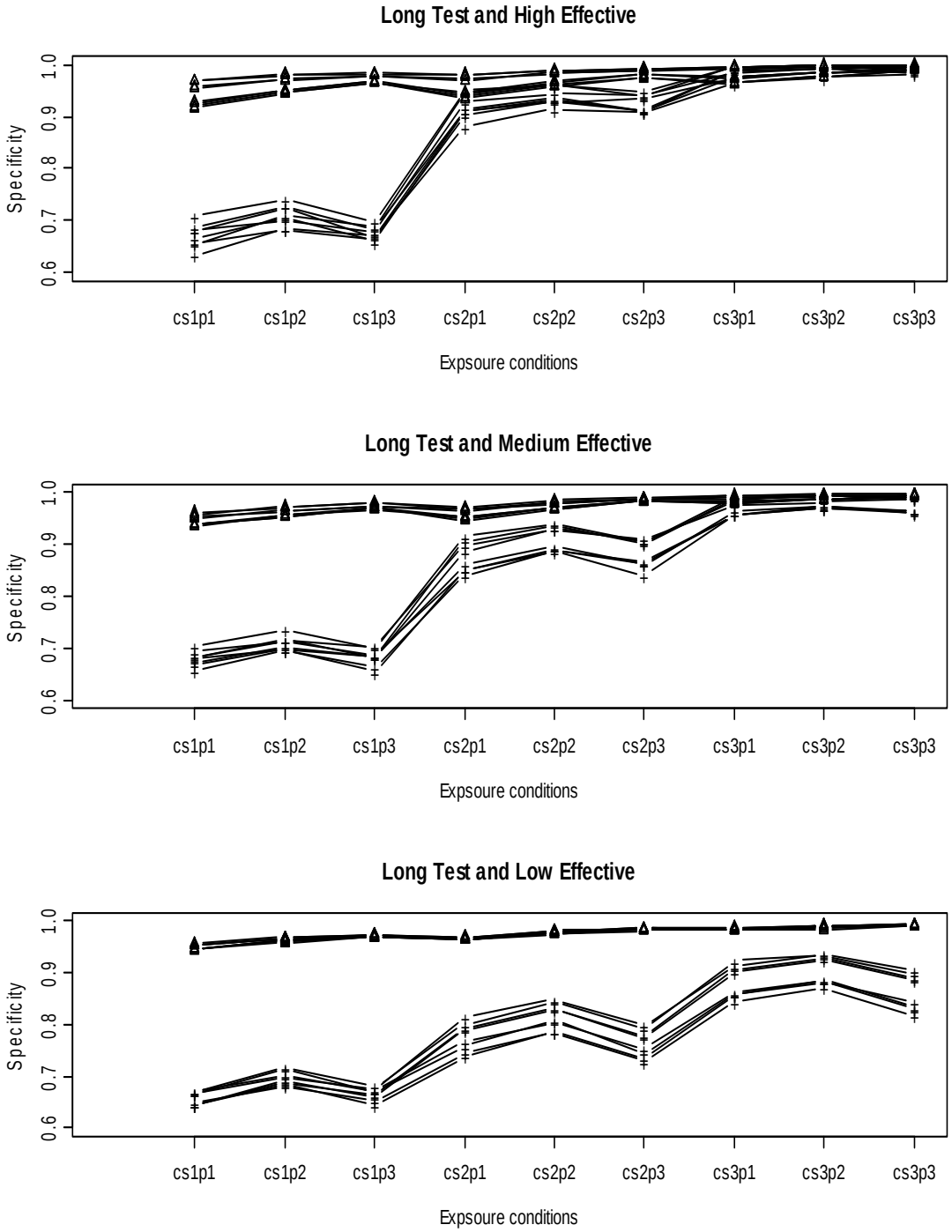


Figure 21. The Specificity of the t-test and Cheating Model in Long Test. "Δ"=the specificity of the cheating model; "+"= the specificity of the t-test; "cs1"=cheating size with 5% cheaters; "cs2"=cheating size with 35% cheaters; "cs3"=cheating size with 70% cheaters. "p1"= 30% items are exposed; "p2"=50% items are exposed; "p3"=70% items are exposed.

As a note, the scale shift problem faced by the cheating model, to a large degree, can be fixed by adjusting the estimation algorithm, and the model's sensitivity level could also be increased using this new estimation algorithm (which will be discussed in the final chapter). The cheating model is a promising and attractive tool that practitioners could use to respond proactively to individual and organized test cheating.

Conclusion

In the simulated settings, the *Deterministic, Gated IRT model* seems to be a promising tool for detecting test cheating with a small degree of mistakes. One of the valuable characteristics of the model is that its specificity is maintained at a high level in every joint condition considered. The practical implication is that the model generally makes a small degree of mistakes in different applied settings. Such characteristics ensure that innocent test takers are treated with fairness in the cheating analysis.

As compared to the model's specificity, the model's sensitivity is greatly impacted by test cheating effectiveness. Generally speaking, cheating effectiveness and cheating degree are two aspects measuring cheating severity that determine the model's sensitivity. As shown in the study, the sensitivity in the high-effective cheating cases is the highest, the sensitivity of the low-effective cases is the lowest, and the sensitivity of the medium-effective cases is between that of the high-effective

and medium-effective cases. Given a certain cheater's increasing cheating effectiveness, his/her probability of being detected grows. Other than cheating effectiveness, an increasing cheating degree (the degree of each exposed items have been cheated by cheaters) can increase the probability of cheaters being detected. In other words, the probability of cheaters being detected grows when he/she cheats on a greater number of items.

Test length, as a critical factor determining test information, exhibits its power to impact model's sensitivity and estimation accuracy. An increasing test length results in an increasing model sensitivity and a greater level of accuracy to characterize cheaters' real knowledge level and cheating severity. More specifically, the sensitivity is impacted by the number of exposed items and unexposed items. Given a certain total number of items, if the item number of exposed items is less than that of unexposed items, the model tends to be conservative and makes less degree of mistakes (which means a higher degree of specificity and a lower degree of sensitivity). However, if the number of exposed items is greater than that of unexposed items, the model tends to be more literal and makes a slightly higher degree of mistakes (which means a higher degree of sensitivity and a lower degree of sensitivity). When the number of exposed items is equal to that of unexposed items, the model derives the highest level of sensitivity and specificity. In other words, the model requires a decent number of exposed items and unexposed items to derive a high level of sensitivity and specificity.

In the two comparisons, the DG-IRT model greatly outperforms the *lz* index in terms of sensitivity and performs as well as the *lz* index in terms of specificity. The lz index seems to lose its power when the cheating size is large or cheating effectiveness is low. The t-test seems to be more powerful to detect test cheating due to its higher level of sensitivity. As study shows that it could almost identify every cheater, but the *lz* is too liberal. A lot of non-cheaters are identified as cheaters, especially in the cases where cheating size is low (e.g., 5% and 35% cheating size cases). Although the model is not as sensitive as the t-test, it maintains a high degree of specificity in every case.

As a special note, the DG-IRT model becomes less and less sensitive when the cheating size becomes greater and greater. In practice, it means that the model might not be an efficient tool to detect and respond to concert or organized test cheating. This limitation is referred to as a scale shift problem of the model in the previous section. Apparently, the scale of the estimated item difficulty and examinee score drifts away from its true ability scale to a scale of the combination of true ability scale and cheating ability scale. Essentially, the model's basic assumption that the item does not change whether the examinee cheats or not is violated. However, it could be fixed by changing the prior of cheating ability, by allowing a free estimation of the mean and standard deviation of the normal distribution of the cheating ability. Such new estimation algorithm in R, which fixes the scale shift problem, is provided in the index at the end of this dissertation.

The DGIRTM is a promising tool to detect test cheating due to item preview, item memorization or internet collaboration on the basis of its performance in the simulated and real setting

CHAPTER V

SIGNIFICANCE, LIMITATION AND FUTURE RESEARCH

Significance

Test cheating, as a negative factor invalidating the inference made based on tests, is commonly recognized by stake-holders of various tests. Testing agencies or other stakeholders have suffered great pain from diffusedly existing test cheating, especially in high-stake tests. Statistical cheating detection methods, as one of important methodologies to detect test cheating, have been proposed and researched by intelligent precedents, however, they re frail in terms of their methodology design, sensitivity, specificity and information they provide. Unlike traditional approaches (e.g., copying indices, person fit indices) which rely on aggregation of individual statistics (Segall, 2002), the *Deterministic, Gated IRT model* derives test cheating or test compromise summaries by response matrix. This cheating model seems to be an attractive and promising tool for practitioners to respond to test cheating by both individual and organized cheating, given its modeling design, the provided information, sensitivity and specificity level

This cheating model is designed to detect the score gain (the score gain from the cheating activities). In real testing settings, the effective cheaters (e.g., those who cheat make significant score gains) jeopardize the validity of tests. The more effective

cheaters are, the greater the impact on the validity of tests grows. From a practical standpoint, "effective cheaters" must be identified, so that the impact on test validity can be examined and appropriate actions can be taken to reduce the impact of cheating (when deemed necessary). As shown by the study, under this cheating model, high effective cheaters are most likely to be detected, medium effective cheaters are less likely to be detected and low effective cheaters are least likely to be identified. Although this cheating model is not effective to identify low-effective cheating, it is the low-effective cheaters who have no or little impact on the validity of tests. What makes this cheating model even more attractive is that it is able to maintain a high level of specificity in every joint condition considered in this dissertation. This cheating detection model is able to detect effective cheaters with a small degree of error.

The other nice feature of this cheating model is the information it provides. One of the key information provided by the cheating model is the cheating probability for each single examinee ($\tilde{T}$). Given higher value of $\tilde{T}$ for a certain test taker, the probability of being a cheater for that test taker grows. Such probability represents the degree of severity of cheating tendency. Not only providing a single cheating probability, this model also characterizes cheaters' real knowledge level by (θ_t) and the severity of their cheating activities (θ_c), which enable practitioners to have deeper inside analysis on the characteristics of cheaters and non-cheaters.

Based on the current study, we believe that the *Deterministic, Gated IRT* model is an effective model that can be used to detect test cheaters. The stability and high accuracy in specificity, capability to characterize cheaters' real knowledge level and cheating degree, and its sensitiveness in detecting effective cheaters, make it a promising tool to proactively respond to the individual or organized test cheating.

Limitation and Future Study

This model is designed to mainly detect test cheating by item-preview, item memorization or internet-collaboration. Its application in other cheating settings might be limited and demand further research and investigation. Based on current study, this model proves to be useful in large-scale tests where a large number of examinees and items are available, but its application feasibility in tests with small sample sizes (e.g., classroom level formative assessment) might be limited. The reason for this is that MCMC is used as the model's estimation algorithm. MCMC estimation requires a large sample size to obtain reliable and accurate estimation. The model is Rash-Model based which does not take guessing and item discrimination into account. However, guessing does exist when examinees answer the questions, and a lot of practitioners believe that items' different levels of discrimination should be estimated. Therefore, the impact of guessing on the model's sensitivity and specificity should be further researched. As a further research direction, this model may be extended by incorporating both pseudo-guessing and item discrimination parameters.

Generally speaking, the *Deterministic, Gated IRT* model is not only a cheating model which can only be applied in detecting test cheating. It can be seen as a general model to distinguish two groups of examinees and be able to characterize the two groups of examinees with two latent traits. For instance, if the model input is defined as item status relative exposure status, it is the model that is used to detect test cheating caused by item exposure (the case discussed in this dissertation). The group of cheaters who have significant score gain is distinguished from the group of non-cheaters who have no significant cheaters. The cheating ability (a combination of true ability and score gain) is the latent trait to characterize the group of cheaters and the true ability is the latent trait to characterize the group of non-cheaters. If the model input is used to group items by two different administration time, it becomes a model that can be used to monitor the examinee growth made between the two testing time. Specifically, it could distinguish a group of students who have significant growth from a group of students who have not significant growth. Therefore, the model's appropriate application in other areas, such as Differential Item Functioning, Student Growth, Scaling Shift, is another research direction which deserves attention.

Although the DGIRTM successfully overcomes some limitations of the previous cheating detection techniques, this new model is certainly not immune to abuse or probative misuse. "There remains an enormous amount of empirical research left to be done before its real application." (Luecht, 2010). The successful and appropriate

application in different real settings is the ultimate purpose for us to develop, and the final standard to evaluate the *Deterministic, Gated IRT model*.

REFERENCES

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

Bellezza, F. & Bellezza,S. (1995). Detection of copying on multiple-choice tests: An update. *Teaching of Psychology*. *22(3), pp.180-182.*

Cizek, G. J. (1999). *Cheating on Tests: How to Do It, Detect It, and Prevent It*. Mahway, NJ: Lawrence Erlbaum Associates.

Drasgow, Levine,& Williams (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of mathematical & statistical psychology. 38 (1), pp.67-86.*

Drasgow, F., & Levine, M. V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement, 10, pp.59–67.*

Drasgow, F.; Luecht, R. M.; & Bennett, R. (2006). Technology and Testing. In R. L. Brennan (Ed.), *Educational Measurement, 4th Edition, pp. 471-515*. Washington, DC: American Council on Education/Praeger Publishers.

Dwyer, D.J., & Hecht, J.B., (1996) Using statistics to catch cheaters: Methodological and legal issues for students personnel administrators. *NASPA Journal*, 33(2), *pp.125-135.*

Frary, R. Tideman, N. &Watts, T. (1977). Indices of cheating on multiple choice tests. *Journal of Educational Statistics. 2(4), pp.235-256*.

Fu,J. (2005). A polytomous extension of fusion model and its Bayesian parameter estimation. Unpublished doctoral dissertation. The University of Wisconsin-Madison.

Hanson, B. Harris, D. & Brennan, R. (1987). A comparison of several methods for examining allegations of copying. ACT research report NO 87-15. Iowa City, IA: American College Testing.

Hambleton,R,K. & van de Linden,W.J.(1982). Advances in Item Response Theory and Applications: An Introduction. *Applied Psychological Measurement, 6(4), pp.373-378*

Holand, P.(1996). Assessing unusual agreement between the incorrect answers of two examinees using the *K*-index: Statistical theory and empirical support ETS Technical Report No.96-4. Princeton, NJ: Educational Testing Service.

Henson,R. Templin, J. & Willse, J. (2009). Defining a Family of Cognitive Diagnosis Models Using Log-Linear Models with Latent Variables. *Psychometrika. 74(2), pp 191-210.*

Iwamoto, C.K., Nungester, R.J., & Luecht, R.M. (1998). *Use of response similarity methods in multistage computer adaptive testing*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25, pp.258–272.*

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4, pp.269–290.*

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

Luecht, R.M. (1998). A framework for exploring and controlling risks associated with test item exposure over time. Paper presented at the National Council on Measurement in Education Annual Meeting, San Diego.

Lewis,C., & Thayer, D.,T.,(1998). The power of K-index to detect test copying. Research report 08541. Princetion, NJ: Educational Testing Service.

Luecht, R. M. (2005). Some Useful Cost-Benefit Criteria for Evaluating Computer-based Test Delivery Models and Systems. *Association of Test Publishers Journal*. (www.testpublishers.org/journal.htm)

Lewis,C.,(2006). Note on unconditional and conditional hypothesis test: a discussion raised by van den Linden and Sotaridona. *Journal of Educational and Behavioral Statistics, 31(3), pp.305-309.*

Messick,S.,(1990). Validity of test interpretation and use. Princeton, NJ: Educational Testing Service.

Meijer, R. R. (Ed.). (1996). Person-fit research: Theory and applications [Special issue]. *Applied Measurement in Education, 9(1), pp.9–18.*

Nering, M. L. (1996). *The effects of person misfit in computerized adaptive testing.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Nering, M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21, pp.115–127.*

Mcdonald, R.P. (1999). Test theory: a unified treatment. Mahwah, NJ: Lawrence Erlbaum Associates.

Mcleod,L., & Lewis,C., (1999). Detecting item memorization in CAT environment. *Applied Psychological Measurement. 23(2, pp.147-159.*

Mcleod, L., Lewis,C.,& Thissen,D.,(2003). A Bayesian Method for the Detection of Item Pre-knowledge in Computerized Adaptive Testing. *Applied Psychological Measurement. 27(2), pp.121-137.*

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K. B. Laskey & H. Prade (Eds.), *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (*pp. 437-446*). San Francisco, CA: Morgan Kaufmann.

Rost, J,(1990).Rash models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement,14(3), pp.271-282.*

Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24, pp.146–178.*

Patz, R. J., & Junker, B. W. (1999b). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), *pp.146-178.*

Rupp,A., Templin,J.,&Hesnon,R.(2009). *Diagnostic Measurement: Theory, Methods, and Applications.* New York: Guildford Press.

Stocking, M. L., Ward, W. C., & Potenza, M. T. (1998). Simulating the use of disclosed items in computerized adaptive testing. *Journal of Educational Measurement, 35, pp.48–68.*

Segall, D,(2002).An item response model for characterizing test comprise. *Journal of Educational and Behavioral Statistics*, *27(2), pp.163-179.*

Segall,D,(2004).A sharing item response theory model for computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *29(4), pp.439-460.*

Sotaridona,L .S & Meijet, R.R (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement, 40(1), pp.53-69.*

Sotaridona,L .S (2003). Statistical methods for the detection of answer copying on achievement tests. Twente University Press, Netherland: AE Enschede.

Tatsuoka,K. (1996). Use of generalized person-fit indexes, zetas for statistical pattern classification. *Applied Measurement in Education. 9(1, pp.65-75.*

Templin, J. & Henson, R. (2006). Measurement of Psychological Disorders Using Cognitive Diag-nosis Models. *Psychological Methods*, 11 (3)*, pp.287-305.*

Templin,J, Henson,R., Templin,S., & Rousso,L.,(2008). Robustness of Hierarchical Modeling of Skill Association in Cognitive Diagnosis Models. *Applied Psychological Measurement (OnlineFirst)*

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

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(3), pp.283-304*

Watson, S.A., Iwamoto, C.K., Nungester, R.J., & Luecht, R.M. (1998). *The use of response similarity statistics to study examinees who benefit from copying.* Paper presented at the National Council on Measurement in Education Annual Meeting, San Diego.

Wollack,J.A(1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*. *21(4), pp.307-320.*

Wollack, A. & Cohen, A. (1998). Detection of answer copying with unknown item and trait parameters. *Applied measurement in education. 22(2), pp.144-152.*

Wollack,A. Cohen,A. & Serlin,R. (2001). Defining error rate and power for detecting answer copying. *Applied Psychological Measurement.25(4), pp.385-404.*

Wollack, J. A. (2006). Simultaneous Use of Multiple Answer Copying Indexes to Improve Detection Rates. *Applied Measurement in Education*, *19(4), pp.265-288*

APPENDIX A: THE CODE OF THE DETERMINISTIC, GATED IRT MODEL

The Deterministic, Gated IRT model estimation is provided in this appendix. This code is built within R static language by following the specification of the prior distributions listed in the model estimation section.

```
# Rash Model
Rash<-function(thetas,betas,alphas){
p<-matrix(NA,length(thetas),length(betas))
for(i in 1:length(betas)){
p[,i]<- 1/(1+exp(-1.7*alphas[i]*(thetas-betas[i])))}
p}

#The Deterministic, Gated IRT Model
mix<-function(true,cheating,betas,alphas,expose,T){
tr<-Rash(true,betas,alphas)
ch<-Rash(cheating,betas,alphas)
p<-tr^(1-T)*(t((1-expose)*t(tr)+expose*t(ch)))^T}

# Module for estimating the mean and standard deviation of cheating ability
draw.mu<-function(cheat0,T0,mu0,sd0){
mu1<-runif(1,max(0,mu0-0.1),min(3,mu0+0.1))
sd1<-runif(1,max(0,sd0-0.1),max(3,sd0+0.1))
theta<-cheat0[T0*(1:N)]
temp<-exp(-(theta-mu0)^2/sd0^2)/sd0
likehood0<- sum(log(temp))
temp<-exp(-(theta-mu1)^2/sd1^2)/sd1
likehood1<- sum(log(temp))
accept<-likehood1-likehood0
accept<-ifelse(accept>0,1,exp(accept))
accept<-ifelse(accept>1,1,accept)
accept<-ifelse(runif(1)<accept,1,0)
mu<-ifelse(accept,mu1,mu0)
sd<-ifelse(accept,sd1,sd0)
output<-c(mu,sd)
output
}
```

```
#Module for estimating the examinee's abilities and cheating status
draw.person<-function(cheat0,true0,T0,betas0,mu0,sd0,N,X,expose){
true1<-rnorm(N,0,1)
cheat1<-ifelse(T0==1, rnorm(N,mu0,sd0),rnorm(N,0,1))
T1<-rep(0,N)
T1[cheat1>true1]<-1
temp<-mix(true0,cheat0,betas0,rep(1,length(betas0)),expose,T0)
temp<-ifelse(X,temp,1-temp)
likehood0<-apply(log(temp),1,sum)+log(dnorm(cheat0,mu0,sd0))*T0+log(dnorm(true0))*(1-T0)

temp<-mix(true1,cheat1,betas0,rep(1,length(betas0)),expose,T1)
temp<-ifelse(X,temp,1-temp)
likehood1<-apply(log(temp),1,sum)+log(dnorm(cheat1,mu0,sd0))*T1+log(dnorm(true1))*(1-T1)

accept<-likehood1-likehood0
accept<-ifelse(accept>0,1,exp(accept))
accept<-ifelse(accept>1,1,accept)
accept<-ifelse(runif(N)<accept,1,0)
tr<-ifelse(accept,true1,true0)
delta<-ifelse(accept,cheat1-true1,cheat0-true0)
ch<-ifelse(accept,cheat1,cheat0)
ten<-ifelse(accept,T1,T0)
output<-cbind(tr,delta,ten,ch)
output
}

# Module for estimating the item difficulty
draw.item<-function(cheat0,true0,T0,betas0,N,X,expose){
betas1<-rnorm(length(betas0),0,1)

temp<-mix(true0,cheat0,betas0,rep(1,length(betas0)),expose,T0)
temp<-ifelse(X,temp,1-temp)
likehood0<-apply(log(temp),2,sum)

temp<-mix(true0,cheat0,betas1,rep(1,length(betas0)),expose,T0)
temp<-ifelse(X,temp,1-temp)
likehood1<-apply(log(temp),2,sum)

accept<-likehood1-likehood0
accept<-ifelse(accept>0,1,exp(accept))
accept<-ifelse(accept>1,1,accept)
```

```
accept<-ifelse(runif(length(betas0))<accept,1,0)
dif<-ifelse(accept,betas1,betas0)
dif
}

#Sampling procedures
MCMC_MIX<-function(iter,N,J,X,expose){
betas0<-rep(1,J)
true0<-rnorm(N)
cheat0<-true0
T0<-rep(1,N)
mu0=0
sd0=1
a=0
delta<-matrix(NA,N,iter,dimnames=list(c(paste("C",c(1:N),sep="")),c(paste("iteration",c(1:iter))))
)
tre<-matrix(NA,N,iter,dimnames=list(c(paste("C",c(1:N),sep="")),c(paste("iteration",c(1:iter)))))
be<-matrix(NA,J,iter,dimnames=list(c(paste("Item",c(1:J),sep="")),c(paste("iteration",c(1:iter)))))
Ten<-matrix(NA,N,iter,dimnames=list(c(paste("T",c(1:N),sep="")),c(paste("iteration",c(1:iter)))))
mu<-matrix(NA,1,iter)
sd<-matrix(NA,1,iter)
while(a<iter){
a=a+1
print(a)
tem<-draw.person(cheat0,true0,T0,betas0,mu0,sd0,N,X,expose)
true1<-tem[,1]
T1<-tem[,3]
cheat1<-tem[,4]
ipar<-draw.mu(cheat1,T1,mu0,sd0)
betas1<-draw.item(cheat1,true1,T1,betas0,N,X,expose)
be[,a]<-betas1
Ten[,a]<-T1
delta[,a]<-tem[,2]
tre[,a]<-true1
mu[,a]<-ipar[1]
sd[,a]<-ipar[2]
betas0<-betas1
cheat0<-cheat1
T0<-T1
true0<-true1
mu0<-ipar[1]
sd0<-ipar[2]}
```

```
output<-list(Delta=delta,true=tre,T=Ten,betas=be,mu=mu,sd=sd)
output
}
```

In this set of code, N is the number of examinees, J is the number of items. X is the response pattern, Expose is the model input to label the item exposure status. Iter is the number of iterations for MCMC.