CHEATING BEHAVIOR IN HIGH-STAKES TEST

By

TARID WONGVORACHAN

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF EDUCATION IN EDUCATIONAL PSYCHOLOGY

WASHINGTON STATE UNIVERSITY
Department of Kinesiology and Educational Psychology

MAY 2020

Synthesis Statement: Cheating Behavior in High-stakes Test

The epidemic of exam cheating is ever-present and continues to threaten score interpretations and uses (Cizek & Wollack, 2017). Cheating is defined as the violation of testing standards for the false representation of the cheater's score (Cizek & Wollack, 2017; Makarova, 2019). This synthesis statement delineates issues of cheating behavior in high-stakes test to establish my literature foundation and identify researchable areas for my scholarly endeavor. Cheating behaviors were constantly monitored to support test validity, which is constantly at risk from threats on various fronts (Kim et al., 2017; Murdock et al., 2016). Fortunately, several protocols were established to secure integrity of the test (see Ferrara, 2017). Several non-statistical measures such as secure test storage and fraud investigation unit were also established as well (Martineau et al., 2017).

Controversial cause of Assessment misconduct

Makarova (2019) suggests that cheating behavior stems from various sources, including Social Demography, Motivation, and Social context of individuals. While the predictability of demographical characteristic is relatively low to other factors, the interplay between motivational and contextual factors have strong relationships in influencing beliefs of academic misconduct (Effron et al., 2015; Ghanem & Mozahem, 2019; Maloshonok & Shmeleva, 2019). Moreover, the No Child Left Behind (2002) act imposed performance-oriented academic culture to education of the United States, which in turn fosters competitive environment and forces educators to cheat for desired results (Martin, 2015; Menken, 2006; Nichols & Berliner, 2007).

Cheating Behavior and its Detection method

Cheating behavior are classified into three main types. Answer change involves the act of changing answers for more score outside of the test session, while answer copying involves

collusive behavior among examinees to share correct answers, and pre-knowledge is when examinees gain unauthorized access to live test items (Qian et al., 2016; Sinharay & Johnson, 2017; Zopluoglu, 2017). Nevertheless, there are cases that cannot be classified into categories but are still considered as cheating, e.g., answer search (Diedenhofen & Musch, 2017; Tendeiro et al., 2013). To detect cheating behavior, both statistical and collateral data are analyzed through psychometric and/or machine learning technique (Kim et al., 2017; Man et al., 2019; Weinstein, 2017). Some examples of detection method are Response Time analysis, Response Similarity Indices, and XGBoost algorithm (Choe et al., 2018; Zopluoglu, 2017, 2019).

Conclusions and Future Direction

The literature reveals that the field has made a considerable progress to secure test security in the modern age of measurement. For my future direction, one option is to expand the current focus of the field for more coverage. Issues around stakes mitigating is also interesting, and the application of machine learning as a viable addition to detect cheating behavior is also one direction I can proceed to as well. This is an interesting time for measurement, where new technologies are introduced at a rapid rate. As much as the threat to test security can evolve with technology, so does the method to secure integrity of the test as well.

Annotated Bibliography

Cheating Behavior in High-Stakes Test

Choe, E. M., Zhang, J., & Chang, H.-H. (2018). Sequential detection of compromised items using response times in computerized adaptive testing. *Psychometrika, 83*(3), 650–673. https://doi.org/10.1007/s11336-017-9596-3

This article proposes the sequential procedure to detect item pre-knowledge through real time monitoring of Item Response (IR) and Response Time (RT) in Computerized Adaptive Test (CAT). CAT tailors test items to match examinee ability at the time of the test, which discourages test tampering or answer copying. However, CAT is prone to prolonged item administration due to its high cost. The excessive usage of test items and on-demand testing policy increase the discrepancy between the number of examinee and item pool size could potentially compromise test items. Sequential monitoring procedure monitors both IR and RT for maximum detection accuracy. IR analysis monitors unusual increase in the proportion of correct response to test items, while RT analysis monitors the significant decrease in RT for an item that is repeatedly presented in the exam. Simulation results support that the proposed method is an equal alternate to detect item-preknowledge. More empirical studies are needed to reflect the actual scenario.

From this article, I noticed that many articles already point out the existing method to detect cheating behavior. Devising a new detection method is not new to the literature. I might need to find some other way to shape my research interest. I also realized that issues in test security can be viewed from different angles (e.g., protecting test items, detecting the suspected examinees), so there are many ways to contribute to test security other than

detecting cheating behavior, such as investigating the Differential Item Functioning (DIF) or Item Parameter Drift (IPD). I am still trying to figure the common factor that researchers consider when evaluating a cheating detection index to apply it to my systematic review. Maybe the number of data point in the study, type-I error, and precision rate of the method could be feasible.

Cizek, G. J., & Wollack, J. A. (2017). Chapter 1: Exploring cheating on tests: The context, the concern, and the challenges. In *Handbook of quantitative methods for detecting cheating on tests* (1st ed., pp. 3–19). Routledge.

In this chapter, the authors introduce facts about cheating at first to describe the status quo in the field of test security. Operationally defined, cheating is an unauthorized action that was taken before, during, or after the time of the test to achieve unethical advantage and/or inaccurate representation of the cheater's level of construct. The occurrence of cheating ties to stake of the test. Cheating threatens the interpretive accuracy of the test score by making the test unable to determine the true level of the measured construct of the cheating individuals. The inaccurate result could further impact validity of the test by interfering in the comparison between the performance of the test taker, especially in norm-referenced test. There are numerous variations of cheating behavior, e.g., Test tampering, Answer copying, or item pre-knowledge. To detect such behavior, evaluation of evidence, quantitative and qualitative, is necessary. Depending on the stake involved in the test program, it could be considered as professionally irresponsible, especially for those in the field of test development, to leave this issue unattended.

A decent introduction to the issue. This chapter systematically present issues around the subject matter in a convincing manner with examples and statistics from various sources. The proposed definition of cheating behavior is specific and not overly broad. The authors also recapitulate the concept of test validity before diving into the concern about cheating and validity, which is a good refresher for the reader. I might need to look further into the book to understand more about method to identify cheating behavior later on, but before that familiarizing myself with the basic concept of technical aspect such as Person-fit index, or IRT might be the best course of action.

DeMars, C. E. (2018). Chapter 14: Item response theory. In D. L. Bandalos, *Measurement theory and applications for the social sciences* (pp. 403–445). The Guilford Press.

In Classical Test Theory (CTT), the estimation of the test parameter is bound to a certain population group that the data was collected from. However, IRT puts every parameter estimation on the same linear scale to gain independency from the population. Thus, enable a test to capture specific constructs at a more precise level. In IRT, each test item is characterized by Item Response Function (IRF), which depicts relationship between examinees' probability of getting the item correctly (Y-Axis) and their level of construct (θ). Two major types of IRT are Dichotomous IRT model and Polytomous IRT model, and each type has its own sub-model (e.g., 1PL, Graded Response). Each model revolves around the assumed variation of three central parameter of the test. These parameters are A-parameter (discriminability), B-Parameter (Difficulty), and C-parameter (Item guessing or Luck factor). In each implementation, item parameter is estimated through *calibration*, which collects data from a sample group (calibration sample). IRT can be implemented in

various ranges of situation such as licensure, certification, and achievement test through Computerized Adaptive Test.

Truthfully, this theory requires a great deal of focus, but it is essential in order to start treading into the area of test security. I have to constantly revisit the basic concept in psychometric to expand my ground into the IRT. The graph is very helpful in understanding item parameters and model. The explanation about each model is also not too complex to follow, and the formula helps illustrating the relationship between item parameter. However, the concept as a whole may require practice for me to fully understand how its component play out in concert with each other. I hope my knowledge in the theory will be sufficient in order to understand advance statistics in measurement such as Person-fit Statistics or Bayesian Theorem.

Diedenhofen, B., & Musch, J. (2017). Pagefocus: Using paradata to detect and prevent cheating on online achievement tests. *Behavior Research Methods, 49*(4), 1444–1459. https://doi.org/10.3758/s13428-016-0800-7

This article discusses the performance of Pagefocus to prevent cheating behavior in Unproctored Internet testing (UIT). UIT requires low cost and low effort, but it is criticized for its vulnerability to cheating, which inflates the test score and render the test parameter ungeneralizable. Various methods such as verification test and person-fit statistics were devised to detect and prevent cheating behavior, but each method has its own limitation. The proposed method in this paper derives information from paradata, which are participant-generated information from computer-based tests. The program monitors the *focus* state of examinees, which is a questionnaire-generated paradata that monitor the

active window of the test device. A test taker enters a state of *defocus* when they change the test delivering window to something else. Losing *focus* in this particular term could signal behaviors that are irrelevant to the test such as answer searching. However, the program does not know of any other activities of examinees other than the state of *defocusing* and *refocusing*. Results show that delivering a warning message telling examinees to stop cheating could reduce cheating behaviors. Participants tend to cheat more when there is performance-related reward involve in the test. However, the proposed method cannot prevent the usage of other devices or proxy test takers.

This article is another great example of cheating behavior that does not fall into the three categories of answer copying, answer change, and pre-knowledge. Each method has its own limitation, which is the reason of why researchers should use an accumulation body of evidence before drawing conclusions regard the breach in test security. After reading a number of literatures in cheating detection method, I believe it is time to steer my research focus on to something that the area of cheating behavior in highstakes test is yet to explain.

Effron, D. A., Bryan, C. J., & Murnighan, J. K. (2015). Cheating at the end to avoid regret. *Journal of Personality and Social Psychology, 109*(3), 395–414.

The purpose of this article is to test out the concept of anticipatory regret, where individuals commit acts of dishonesty when facing with imminent depletion of the opportunity to cheat. The feeling of frustration intensifies when individuals find themselves at the end of the opportunity to take advantage of a moral dilemma. The authors hypothesize that individuals are likely to cheat on their last opportunity, where the anticipatory regret is at its peak and the opportunity is the most vivid. Alternative explanations to anticipatory regret are moral

self-licensing, ego depletion, and slippery slope attribution, which will be examined in this article as well. Every theory suggests that people tend to cheat at the end of the opportunity, but the anticipatory regret concept could be better in explaining the situation where additional opportunity to cheat is introduced. When facing with extra cheating opportunity, the occurrence rate of cheating would drop if explained by the anticipatory regret, whereas the occurrence rate would rise when explained by the three alternatives. Four studies were conducted to examine hypotheses of every theory. Findings are consistent with the theory of anticipatory regret with minor deviations. Implications of this research could be used to prevent cheating behavior in various settings.

This article is a helpful to understand the mechanism of cheating behavior. However, the study only covers the scope of cheating in general, not specific to testing context. Maybe my search keyword is off, or the literature on this topic is scarce. I could find out more about situations that induce the anticipatory regret or moral dilemma and start from there to learn more about environment induce cheating. This hypothesis seems to be consistent with the era of NCLB, where educators and students were tempted to engage in unethical means to achieve the educational standard.

Ferrara, S. (2017). A framework for policies and practices to improve test security programs: Prevention, detection, investigation, and resolution (PDIR). *Educational Measurement: Issues and Practice*, *36*(3), 5–23. https://doi.org/10.1111/emip.12151

The PDIR framework presented in this article was proposed as a guideline for test-related professionals to plan and operate a secure test security system. Issues in test security are frequently underestimated, and the cooperation from all test-related party is critical to

minimize the risk of test security breach. The framework consists of four elements that combine into a test security protocol, Prevention, Detection, Investigation, and Resolution. Prevention is the most effective countermeasure to cheating behavior. The procedure is typically carried out after each test session to scan for any potential breach. Detection step usually requires both statistical and non-statistical information to identify the breach. Investigation is then carried out to gather additional data, and resolution phase is taken as the last step to confirm or refute the breach. However, that research on the development of cheating detection method is still scattered and in its emerging phase, which renders the state-of-the-art in detection technique limited in number and quality.

The article offers a clear introduction to the field besides Cizek and Wollack (2017). The framework allows the reader to understand the system of test security in a big picture. The author also raises an interesting point that majority of test security research is conducted through simulation, which causes a literature gap in the distribution theory and a limited scope of application. Methods that work well in a simulation could yield different outcomes when being simulated with different datasets. Another limitation of the currently available detection methods is that they require a large number of examinees to establish a standard, which constrains their application to the large-scale test only. This article could serve as a sound starting point for my study on cheating detection method, as well as opens up an array of research possibilities in the field of test security.

Ghanem, C. M., & Mozahem, N. A. (2019). A study of cheating beliefs, engagement, and perception – the case of business and engineering students. *Journal of Academic Ethics*, *17*(3), 291–312. https://doi.org/10.1007/s10805-019-9325-x

Academic dishonesty is a prevalent threat to academia. Specifically, studies found that business student cheat more than student from other majors due to the external profit-oriented mindset in short-term. However, the support on this statement is still mixed. This paper examines the influence of both individual and contextual factors to cheating behavior between business and engineering student. The authors also investigate the influence of bussiness ethic course and peer influence to the exhibition of cheating behavior as well. Data were collected through a series of survey with both dichotomous and polytomous item. Findings reveal that the prevalence of cheating is spreading at a rapid level. In most cases, cheaters commit and externally justify their academic misconduct despite knowing its consequences. However, few peers would report when facing with unethical behavior in academia. This status quo could be developed into unethical mindset at a larger context such as bussiness and political setting. Contextual factor (e.g., peer perception, norm) also plays a part in inducing academic misconduct, but more than one social actors (and factors) must be considered when predicting such behavior. Surprisingly, the factor of business ethic course was found to positively influence cheating behavior rather than reducing it.

Results from this study provide further support for Makarova (2019)'s study that contextual factor has a significant influence on cheating tendency. The idea could be developed into a notion that regulation of a test center might be a factor that contribute to the breach in test security. Peer influence in cheating is also important as well, which could be lessened by educating people, maybe before the test session, of the repercussion of cheating. Moving

forward, I want to look into the unintended consequence of event-based testing, which could pressurize examinees to cheat due to the limited opportunity to retake the test.

Kim, D., Woo, A., & Dickison, P. (2017). Chapter 4: Identifying and investigating aberrant responses using psychometric-based and machine learning-based approaches. In J. A. Wollack & G. J. Cizek (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (1st ed., pp. 70–97). Routledge.

This chapter explains relationships between test security and test validity. The guideline of the Internal Test Committee (ITC) states that test scores should be secured maintain the integrity of the test score and its entailing decision. Validity argument of a test is created by the combination of content, psychometric, and security of the test. To psychometrically distinguish cheating examinees, aberrant response pattern, aberrant response time, and auxiliary information are considered. This chapter discusses two approach to detect aberrant response, psychometric-based approach and machine learning-based approach. The psychometric approach primarily utilizes statistical model and item response to identify the suspected examinees. This approach is further classified into two categories, person-fit indices, and response time models. Person-fit indices operate based on the agreement between the suspected score pattern and the model of interest to detect cheating in general, whereas the Response Time Model flags examinees with extreme response time to detect pre-knowledge. Machine learning-based approach, however, focuses on the description of auxiliary data through Market Basket Analysis. This technique is used to identify which auxiliary information appears in common with the cheating individuals for additional information and further investigation.

This article confirms my understanding that cheating is under the umbrella of aberrant response behavior. The validity triangle is very helpful in pointing out relationships between validity and test security. The information on psychometric approach to detect cheating and aberrant response is able to further my understanding in how statistical models operate in detecting the suspected. I also become aware that the detection process can be done at any time of the test as well. The machine learning approach is also very interesting in how it collects additional information to complement the limitation of psychometric-based approach. The auxiliary information collected from the model could open an array of possibilities in research.

Makarova, M. (2019). Factors of academic misconduct in a cross-cultural perspective and the role of integrity systems. *Journal of Academic Ethics*, *17*, 51–71. https://doi.org/10.1007/s10805-019-9323-z

Academic dishonesty, particularly exam cheating and plagiarism, is detrimental to education. This study utilizes mixed method approach to investigate academic integrity from the three factors of Individual, Motivational, and Contextual in cross-cultural perspective. Individual factor comprises of socio-demographic features such as social status and personal achievement, GPA included. Motivational factor is psychological attribute such as self-regulation or self-justification. Lastly, contextual factor involves many levels of societal factor from peer influence, culture, and institutional code. This study investigates universities from USA, Latvia, Poland, and Russia, with the US as a sample of country with a well-founded integrity system, and the remainder as the developing country. Quantitative survey was distributed with convenient sampling, as well as document analysis and semi-

structured interview were conducted. Results found that individual factor is ambiguous in detecting academic misconduct, while contextual factor of the implementation of academic integrity policy and teachers' control is strong in predicting academic misconduct. Results in contextual factor is supported by findings in motivational factor, which posits that collectivism plays a role to predict collaborative cheating. In a place with weak integrity system, students tend to neutralize and externalize their integrity violation.

This study is a very helpful starter to learn about ecological factors of cheating behavior in tests, which is a part of academic misconduct. Results in this study is consistent with the idea that cheating is situational (Nichols & Berliner, 2007). A strong, but not oppressive policy in accountability testing could prevent cheating behavior in test. An observation I would like to make is that participants in this study are person in academic institution such as students, administrators, and teachers, but person in highstakes testing could be non-academic professionals such as nurse or lawyer. As a result, findings in this study might have limited applicability outside educational setting.

Maloshonok, N., & Shmeleva, E. (2019). Factors influencing academic dishonesty among undergraduate students at Russian universities. *Journal of Academic Ethics*, *17*(3), 313–329. https://doi.org/10.1007/s10805-019-9324-y

With the increasing epidemic of academic dishonesty, researchers have been contributing efforts to investigate its causal factor. This paper proposes to identify causal factors to academic dishonesty in Russian student. Academic dishonesty was born through the interplay of several factors (e.g., individual, psychological, and societal). The model of

Planned Behavior Theory (TPB) was found to be an effective predictor of academic dishonesty among college student. TPB suggests that the performance and intention of cheating behavior are determined by behavioral belief, normative belief, and control belief. Findings reveal that for Russian student, perception of subjective norm to cheating behavior has greater influence on academic dishonesty than individual attribute due to collectivistic social norm. The factor of behavior control, however, rests on the level of regulation enforcement imposed on the student. If there is weak or no enforcement at all, the factor of whether cheating behavior can easily be performed would have less effect than the other two factors. Despite the counter intuitive results in behavioral control, regulation on academic integrity is still an effective measure to reduce academic dishonesty.

From all paper I have read about academic dishonesty, an idea that emerged is that characteristic of cheating behavior varies from one region to another. Focusing on using contextual factor might be able to reduce frequency of the breach in test security. The topic of pressure in event-based testing to cheating behavior sounds promising. Some notable examples are Gaokao exam in China, CSAT exam in Korea, or Nyugaku Shiken in Japan. One commonality that the mentioned examples share is that they are all located in Asia, where education is different than the western country. Research in this area could inform the improvement of college admission test to balance the stake attached to the test.

Man, K., Harring, J. R., & Sinharay, S. (2019). Use of data mining methods to detect test fraud. *Journal of Educational Measurement, 56*(2), 251–279. https://doi.org/10.1111/jedm.12208

Cheating behavior in tests not only put validity of the score interpretation at risk, but also negatively impacts the standard in score comparison with the questionable result, especially in high-stake competitive assessment. The literature then describes classifications of cheating behavior and methods to detect them (e.g., Answer Copying, Pre-knowledge, and Answer changing). Methods to detect cheating behavior are categorized into two major groups, Item Response Theory (IRT) model and Response Time (RT) model. Each method has its own limitation. Data-mining method is proposed to close the limitation by utilizing data from both models. The technological advancement in past decades resulted in migration from paper and pencil test to Computerized Adaptive Testing (CAT), and the migration also led to the possibility of data mining during testing scenario as well. Datamining technique has two types that yield different results, Unsupervised Machine Learning method and Supervised Machine Learning method. The authors implement datamining on a secondary dataset with results showing that datamining have high sensitivity in flagging cheating individuals than traditional RT and IRT methods.

I decided to approach my systematic review topic by starting from the most recent article, then tracing back to the basics. This article is a good beginning in familiarizing myself with issues around test security and cheating detection. The authors explain subject matter in an understandable manner that allows me to grasp the current situation of test security without much struggling. However, several technical topics that need more clarification are mentioned in the paper, and I might need to do more literature search for that. Such topics are Item Response Theory (IRT) and Person-Fit statistics. Also, I might need to investigate

deeper into the issue of cheating and its ramification, and each type of cheating behavior in detail as well to gain more ground in this area.

Martin, A. J. (2015). Chapter 5: Are these testing times, or is it a time to test? Considering the place of tests in students' academic development. In H. Proctor, P. Brownlee, & P. Freebody (Eds.), *Controversies in education orthodoxy and heresy in policy and practice* (Vol. 3, pp. 55–62). Springer.

This paper discusses an alternative approach to high-stake accountability testing to reduce concerns of unbalanced stake in education. Majority of concerns about high-stake accountability testing revolve around the policy of accountability testing, which places its emphasis on test results while neglecting surrounding contexts of students. Furthermore, the enforcement of the policy causes educators to spend their time preparing students for tests rather than teaching for the sake of learning. However, there are four criteria that can be referred to when using a high-stake test to ensure educational validity for test takers and test-relevant person. First, the test must be able to enhance student achievement with its result. Second, the test has to demonstrate student development, which can be done by measuring students across time with their previous achievement, not the whole normative curve. Third, the test must be able to address and respond to concerns of accountability testing by adopting the perspective of unique growth for each student. Fourth, test results must be used for formative purposes to inform educators and students of rooms for improvement while providing anonymity for each school at the national level to avoid stigmatization.

This paper offers a relatively short, but rich, introduction to a new perspective to high-stake testing. Many high-stake tests would not exist without their merit; however, their stake and associated costs can be intimidating if mistreated. The problem is that even through if we are able to adhere to the discussed criteria, problems at test content level of the test still exist. Consider GRE, for example, even native English speakers have a hard time in preparing for this test, let alone those who use English as their second language. This points to the area of test fairness and DIF, which is not in the scope of my project topic.

Martineau, J. A., Jurich, D., Hauger, J. B., & Huff, K. (2017). Chapter 15: Security vulnerabilities facing next generation accountability testing. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (1st ed., pp. 283–307). Routledge.

This chapter discusses the need for the improvement of test security in next-generation accountability testing. The transition from paper-based testing (PBT) to computer-based testing (CBT) gives rise to new challenges in the test security issue. The passing of No Child Left Behind act drastically increase the demand for the high-stakes standardized tests. The pressure of high-stake testing was found to be associated with the escalation of cheating behavior in other the students and the educators. The associated cost in implementing measures in test security can also be costly. However, it is necessary to maintain the test security in order to secure the confidence of the stakeholders in the test score. Various aspects of concern in test security exist, such as the assessment type, sources of the threat, and even the variation of the threat itself. As a result, each testing agency or institution must tailor the measure in test security in accordance with their unique context. Even so, the

confirmation of a breach must be decided with utmost consideration and sufficient data, for it could create a lasting impact on the confidence of the stakeholders.

I took a step away from the technical detail of cheating detection to read about the conceptual aspect of test security. Understanding the underlying status quo could provide rationales to support the importance of test security research. This article greatly addresses the gravity of the issues in test security in a convincing manner, and the examples presented in the chapter gives me a clear picture of the field of test security as well. The authors present a good point that computer-based test or internet-based testing is becoming more prevalent every passing year, which indicates the need of development in the field of test security as well.

Maynes, D. D. (2017). Chapter 3: Detecting potential collusion among individual examinees using similarity analysis. In J. A. Wollack & G. J. Cizek (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (1st ed., pp. 47–69). Routledge.

This chapter explains cheating detection method of similarity statistics, which is a complement to answer-copying statistics to detect response similarity in tests. Answer-copying statistics only cover the scenario of one source and one copier in the same test setting. Similarity statistics take all other scenarios such as the collusion between two or more examinees, the usage of hidden electronic device to communicate with external answer sources, or even the usage of proxy test takers into account. Additionally, the similarity statistics are able to detect test violators at the group level and have no need to define the copied and the copier due to its symmetry attribute. The statistical similarities operate based on the three assumption; first, response between two response vectors are

independent of each other, second, the response of a single test taker depends on his/her performance of the test, and third, item responses are only independent at local level. Some limitations of this approach are its requirement of large sample size to calculate the matching probability and its inability to detect individuals who cheat independently without cooperation with conspirators.

This chapter is a great addition to further my understanding in answer-copying detection. The two methods of answer-copying indices and similarity response statistics could be used in tandem to complement each other. While reading this chapter, an insight came to me that the advancement of technology may alter the nature of both testing and cheating, as well as give rise to the need of a new method to keep up with the change. A solid example I got is that paper-and-pencil test used to be prevalent in educational testing until the rising of computerized adaptive testing, which makes the examinees unable to copy the answer of each other due to the adaptability of test items.

Meijer, R. R., Niessen, A. S. M., & Tendeiro, J. N. (2015). A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics: Examples and a computer program. *Assessment*, *23*, 52–62. https://doi.org/10.1177/1073191115577800

This paper was written with an aim to delineate person-fit statistic for non-specialist or those who are new to the field. Inconsistent score pattern could undermine the overall validity of the test if left unchecked. Person-fit statistics work based on the assumption that person with high-level of construct (θ) should be able to correctly answer the low-level test

item. Patterns that do not adhere to the assumption is called *Guttman error*, which signifies aberrant response pattern. There are two existing models for this statistics, Person-fit for dichotomous items and Person-fit for polytomous items. The more aberrant the score pattern is, the higher the person-fit score will be. Extreme person-fit score contributes to the evidence that the score pattern of the individual conflicts with the *Guttman assumption*. This conflict connotes the lack of reading skill of the respondent, deviant response behavior, lack of motivation, or even pathology personality. The limitation, however, is that person-fit statistics are only sensitive to inconsistent score pattern, and power of the model largely depends on test length and item characteristics.

This paper did well in introducing the person-fit statistic, which is an advance topic in measurement that builds up on the concept of Item Response Theory. The provided example is also easy to follow. However, the statistics mechanism of the model is relatively arduous and need to be visualized in order for me to keep up with the concept. My prior knowledge in data management and visualization, R studio in particular, is very helpful in understanding the paper. Still, I can only grasp the conceptual aspect of the model for now, but not the practical part. R program is already intricate in itself, let alone implementing the package discussed in this paper. Further reading is needed for me to proceed further.

Menken, K. (2006). Teaching to the test: How no child left behind impacts language policy, curriculum, and instruction for English language learners. *Bilingual Research Journal*, *30*(2), 521–546. https://doi.org/10.1080/15235882.2006.10162888

This article discusses the detrimental impact of high-stake accountability testing regulation from the passing of No child left behind act (NCLB) to English Language Learners (ELL). NCLB changes the pedagogical policy into performance-focused curriculum, as well as making students in the United States responsible for their score in excessively high-stakes tests. As a result, the overreliance on English proficiency test as an indicator of student ability widens the performance gap between native English speakers and ELLs, thus raising concerns for psychometric issues as well as test fairness. Further, the mentioned gap acts as a major contributing factor to student dropout rate, causing schools to change their teaching policy for the purpose of test taking rather than language mastering. The teaching policy was altered to prepare students for the test by emphasizing literary elements than aiming for communication competence. Educational gatekeepers must consider differences between English speakers and ELLs when making important decisions on their profile.

This paper raises important points that I can incorporate into my final project. The emphasis on accountability testing impacts how schools teach their students in English class, and I can relate to that with my experience as a test taker as well. Aside from cheating behavior, cheating detection methods, and the causal factor of cheating, test fairness might be the area I would want to explore in gaining perspectives of high-stake assessment. If the test is fair to begin with, the need for examinees to cheat or engage in intensive test preparations could be lowered. Additionally, if we could assess students without right or wrong answer,

pressure of high-stake testing could be significantly reduced as well. As far as I recall, Dr. French is working on projects of this scope at the moment.

Murdock, T. B., Stephens, J. M., & Grotewiel, M. M. (2016). Chapter 11: Student dishonesty in the face of assessment who, why, and what we can do about it. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (1st ed., pp. 186–203). Routledge.

The purpose of this chapter is to review cheating behavior from various perspectives, educational, personality, achievement motivation, moral, and classroom context. Cheating, or in this chapter, academic dishonesty, is an act that can be done both digitally and physically for unethical advantages over other peers. It hampers the purpose of assessment as well as putting test results at a risk of losing public confidence. Personality is one of many factors that contributes to the occurrence of cheating. Fortunately, various evidence from personality researches support that personality is malleable through time and experience, thus students who engage in cheating can be educated to lessen their behavior. External factors such as teachers' pedagogy, orientation of goal in the class, and social norm could determine academic dishonesty as well. Teachers can promote academic honesty in students through education, psychometricians can focus on cheating detection methods and cheating-proof assessment, and policymakers could advocate for appropriate stakes in accountability testing.

This chapter provides a very broad perspective on the causal issue of academic dishonesty or cheating behavior. Cheating in high-stake test is a term under an umbrella concept of academic dishonesty, so there should be no problem in applying the explanation of this

chapter to cheating behavior in other articles. After reading a number of articles on cheating detection method, I realized that there is already an array of options to detect cheating. However, efforts should be made to tackle the problem at its root as well, not just its leaves. We might need to consider other related factors to cheating behavior such as the assessment stake, the policy on testing, and contextual element of the behavior. A bigger perspective could be useful to understand the circumstance of cheating in high-stake test, which allow us to approach the problem with efficiency.

Nichols, S. L., & Berliner, D. C. (2007). Chapter 12: The pressure to cheat in a high-stakes testing environment. In E. M. Anderman & T. B. Murdock (Eds.), *Psychology of academic cheating* (1st ed., pp. 289–311). Academic Press.

This chapter discusses consequences of high-stake accountability testing and the how it contributes to cheating behavior. In contrast to the common belief, the inclination to cheat of an individual varies across time and situation. Whether people cheat or not could be partly attributed to the test itself. The No Child Left Behind act (NCLB) puts immense pressures on students, teachers, and school administrators in aiming for the standard, which forces people to cheat to achieve the desired score. The pressure intensifies when the stake is exceedingly high, resulting in a binary belief that a person can only be either "remarkable" or "incompetent" with their "pass" or "fail" grade. The linkage between high-stake testing and cheating behavior can be explained by Campbell's law, which states that the higher the stake of the test, the easier the test takers and the system will be corrupted. High-stake testing also alters the educational focus into performance-oriented approach, which commodifies the meaning of mastery into a mere number in a score report. Thus,

developing a testing culture where cheating is justified. However, cheating can be either clearly wrong or morally ambiguous. This fact many times put educators in an ethical dilemma of whether to break the rule.

This chapter provides me a great understanding in undesirable results of the high-stake testing and its sources. My thought is that I need understand both sides of high-stake testing in order to fully grasp its situation. While I agree that high-stake test provides solid information to the test takers about their ability, the associated cost of taking a test is high and stressful. For that reason, I aim to understand issues in high-stake testing in a broad perspective to identify rooms to improve in both the literature coverage and the practice of using the high-stake assessment.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, *63*, 1–11. https://doi.org/10.1016/j.jrp.2016.04.010

This paper tests out the effectiveness of multi-test extension of person-fit statistics to detect inconsistent score pattern against random responses in web-based questionnaires. Since person-fit statistics cannot be applied to a test with multiple short subscale due to its dependency on A-parameter, B-parameter, and test-length, multi-test extension of person-fit was created in order to increase its detection sensitivity. Multi-test extension pools information on person-fit statistics across multiple short subscales by summarize the number of Guttman Error per subscale (G_m^p) or the standard log-likelihood of statistics of the subscale (l_{zm}^p) together to create a normative data. The authors then investigate the

sensitivity and specificity of multi-test extension and other statistical techniques on both empirical data and simulated data of International Personality Item Pool (IPIP). Results suggest that person-fit multi-test extension yields better result comparing to other consistency-check methods. The first key-takeaway is that all random-proof methods do not have high sensitivity in detecting random responses maybe due to the fact that normal and careless response are similar to each other. The second takeaway is simple techniques (e.g., Response time analysis, Long string analysis) seems to be more efficient in moderate-length questionnaire than other complex technique, multi-test person-fit included.

An informative read in exposing myself to methodology to detect aberrant response behavior. However, the technicality in person-fit statistic part is relatively hard to grasp and requires multiple read to be understood. Nevertheless, there is still some part that I'm struggling with, G_m^p and l_{zm}^p in particular. I might to consult with faculty members for clarification. Aside the mentioned topic, other part is clear enough, and the experiment is really helpful in understanding the effectiveness of multi-test person fit technique. The concept of sensitivity and specificity is also easy to follow in assessing how detection method operates.

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice*, *35*, 38–47. https://doi.org/10.1111/emip.12102

This article investigates the usage of Response Time analysis to detect item pre-knowledge. Pre-knowledge occurs when the examinees have unauthorized access to test item pool,

which enables them to correctly answer the questions they do not have the required level of construct (θ). The memorized test items are called *compromised items*. Both pre-knowledge and compromised test items could undermine validity of the test score. RT analysis operates on the assumption that the examinees cannot fake their RT, and both excessively short and long RT could indicate pre-knowledge and item memorizing respectively. The base idea behind models used in this study is that if a test item is time-intensive and the examinee has low (θ), the RT is expected to be long. Results shown that RT analysis could detect pre-knowledge and compromised item. However, psychometric evidence alone should not be used as an evidence to invalidate a test score, but it is a justified point for further investigation. The fault could rest on the item itself; Having compromised items reviewed by content experts is also a good alternative to check if the item is memorable. Detecting compromised items can improve validity of the test score, as well as strengthening the item generation in various testing contexts.

I became aware that Cheating and Aberrant Response pattern share a considerable amount of characteristic in common. Both are purposeful behavior aside from responding to the test item with the true level of construct. However, not all test takers with pre-knowledge have the intention to cheat. They might be subjected to unintentional item exposure, retesting, or practice effect. I agree with the recommendation that additional data should be used to indicate cheating behavior. It takes a lot more than just data to justify the invalidation of a test score.

Sinharay, S., & Johnson, M. S. (2017). Three new methods for analysis of answer changes. *Educational and Psychological Measurement, 77,* 54–81. https://doi.org/10.1177/0013164416632287

It is well-known that cheating endangers validity of the test score, especially in high-stake testing environment. Analysis of Answer Change are a set of analyses to detect answer erasing patterns that tampers with the test integrity. Erasure Detection Index (EDI) is one of many effective countermeasures to Answer Change (AC). Higher EDI value indicates a higher potential of Answer Changing. However, EDI without continuity correction has inflated type-I error (false positive), and EDI with continuity correction has lower statistical power despite having no Type-I error problem. The authors propose three new alternatives to detect test tampering, which requires only the final answers of the examinee and applicable to any IRT model. The proposed methods are Generalized Binomial Model, Exact probabilities and Score Patterns, and Posterior Predictive Model Checking. All methods have better Type-I error rate and statistical power comparing to the EDI. However, further evidence is needed for the confirmation, and examinees with AC should be given an opportunity to re-take the test or defend themselves.

This article is very helpful in exposing myself to issues in Answer change. There are some technical parts that I'm struggling with, but I manage to extract the conceptual part out of it. My plan is to read through the article to understand as many forms of cheating behavior and how to counter them as exhaustive as possible. Then, I will proceed to read each method in-detail, as well as overarching theory and practice of test security to understand the bigger picture of the field. I became aware that even though each detection method addresses each cheating behavior specifically, the aberrant response pattern itself affect test validity as a

whole. Test developers should invest efforts to detect any potential threat to the test item and keep the test up to date.

Tendeiro, J. N., Meijer, R. R., Schakel, L., & Maij-de Meij, A. M. (2013). Using cumulative sum statistics to detect inconsistencies in unproctored internet testing. *Educational and Psychological Measurement*, *73*, 143–161. https://doi.org/10.1177/0013164412444787

This article introduces the application of Cumulative Sum Statistics (CUSUM) to detect cheating behavior in Unproctored Internet Test (UIT). UIT is well known for its low cost, low time required, and minimum human supervision. However, the test is criticized for its lack of standardized condition. For validity reason, test takers are usually required to take a verification test of similar nature to assess consistency of result. The proposed CUSUM technique compares ability estimates from the UIT with score patterns from the verification test to identify potential over- or underperformance of the examinee. CUSUM is able to use item-level information instead of overall ability estimates to detect signs of cheating, and the result can be visually displayed in charts as well. CUSUM operates by the process of continuous parameter monitoring. Traditional l_z- and Z statistics are also applied to a real data set to validate the proposed method as well. As a result, CUSUM was found to have similar performance to alternative methods in detecting cheating behavior. The authors recommend further empirical applications of cheating detection statistics to gain more generalizability for the field.

This paper discusses the detection of cheating behaviors that are not covered in the three categories of Answer Copying, Answer Change, and Pre-knowledge. While numerous

statistical indices exist to detect cheating behavior, there is no limit to how cheating pattern varies as well. After covering a little bit more ground on cheating detection method, my plan is to read about ecological context of cheating behavior to identify any contextual characteristic of examinees that could contribute to the occurrence of cheating behavior. Some examples of ecological factor that I could think of are socio-economic status of examinee population, educational system, or even system level problems such as educational policy that could influence curriculum of schools.

Weinstein, M. J. (2017). Chapter 19: When numbers are not enough: Collection and use of collateral evidence to assess the ethics and professionalism of examinees suspected of test fraud. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (1st ed., pp. 358–369). Routledge.

This chapter discusses the rationale to identify, acquire, and preserve collateral evidence of the breach in test security. In many cases, the desired ethical and professional characteristic of the test takers are considered as equal as the test score itself. Some examples of the mentioned cases are the medical field, which maintains the security of the test on high-vigilance majority of the time through the routine update of item pool or multi-stage certification process. To ensure the integrity of the test, a vertically integrated test security program is needed. The program operates by assigning every test-responsible sector such as test development, executive, and information technology with their respective protocol to maintain the test security as a whole. Examinees are expected to acknowledge and sign the comprehensive agreement that the test security protocol is based on. Lastly, the test provider

should make conclusions regard validity of the suspected examinees' score based on the collective body of evidence.

This chapter is essential to understand the context of test security and cheating as whole. Earlier in the reading, Cizek and Wollack (2017) explain the nature and the definition of cheating behavior. Majority of the articles already explain why cheating is considered as a threat to the test validity. Martineau, Jurich, Hauger and Huff (2017) also describe why the improvement of test security and cheating detection measures should be reinforced and maintained, as well as its associated cost and the potential vulnerabilities. This article serves as the end of the circle by explaining the significance of collateral evidence, as well as the advisable actions taken after the detection of the breach aside from "plugging" it. In short, all articles I mentioned above act as a foundational rationale for articles on specific cheating detection method, which will be the body of the whole project.

Zopluoglu, C. (2017). Chapter 2: Similarity, answer copying, and aberrance: Understanding the status quo. In J. A. Wollack & G. J. Cizek (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (1st ed., pp. 25–46). Routledge.

Answer copying or unusual response similarity is a cheating behavior that indicates test fraud. Two statistical methods were developed to detect answer copying, one of them is Response similarity indices, another, Person-fit indices. Person-fit statistical examines the alignment between a single suspected response vector with a calibrated response normative model. The Person-fit models are able to detect the aberrant response pattern in general, but they are considered less effective and under-researched when being compared to response similarity indices. The reason is that every copier has aberrant response pattern, but not

every aberrant respondent has an indication of answer copying. Response similarity indices, on the other hand, are techniques that revolve around the assumption of independent responding and focus on the likelihood of agreement between two suspected response vectors. The article suggests two-stage approach to detect answer copying, using Person-fit technique as a screening tool for a suspected pair of examinees, and using Response similarity indices to calculate the degree of agreement between the two response vectors. However, the discussed methods were tested in paper-and-pencil test only. Results could be different in computer-based test.

With this article, I am able to familiarize myself with all three variations of cheating. During the read, I realized that there is no right or wrong model to use to detect cheating behavior. I was also able to draw the big picture on the topic of methods to detect aberrant response pattern, that each type of measurement requires different kind of aberrance-proof measure. Non-cognitive test may only require methods to detect Aberrant Response pattern such as long-string analysis, but cognitive test requires cheating-proof methods that are more specific comparing to the non-cognitive test. Nevertheless, all method belongs under the umbrella concept of Aberrant Response Detection, which could be used to indicate potential threat to validity.

Zopluoglu, C. (2019). Detecting examinees with item preknowledge in large-scale testing using extreme gradient boosting (XGBoost). *Educational and Psychological Measurement*, *79*(5), 931–961. https://doi.org/10.1177/0013164419839439

This article explores the effectiveness of Extreme Gradient Boosting (XGBoost) algorithm in detecting pre-knowledge. Many psychometric-based methods were devised to detect item pre-knowledge through the analysis of item response and response time (RT). However, literature on ML algorithm to detect cheating behavior is still lack in number. XGBoost operates by predicting an outcome through the iteration of classifying a score pattern of a test taker. XGBoost is tested with a real dataset and compared with two traditional psychometric methods to detect pre-knowledge for confirmation. The algorithm uses Tree Ensemble decision model to predict the occurrence of pre-knowledge. Results indicate that XGBoost is comparable to the traditional psychometric-based method. Additionally, the incorporation of RT in the analysis greatly improves the detection rate, and the nominal item response-based model proves to be more powerful than the dichotomous item-based model. The limitation of XGBoost is that the greater the number of decision stage, the more complex the model will be to the point of becoming a 'black box'. In an actual implementation, more information could be added to make the model more realistic.

After I finished with the study selection phase in my systematic review assignment, I can more or less recognize what the author is saying about when referring to the literature in the field. This article raises an important point that literatures ML-Based approach is still scattered. However, I am aware that I need more background in machine learning or how algorithm operates. I will try to read for conceptual understanding for now. Still, I am able to see a vague direction of the field. After the migration of paper-and-pencil test into computerized adaptive test (CAT), literature in pre-knowledge seems to be more popular than the other two cheating behaviors. This might be due to the hard-to-cheat nature of CAT.