**Operational Measurement: Quality Assurance 2016**

**Annotated Bibliography**

**Cheating**

* Cizek, G.J. (1999). Cheating on tests: How to do it, detect it, and prevent it.
* Foster, D., Maynes, D. and Hunt, B. (2008). Using data forensic methods to detect cheating.
* Han, N., & Hambleton, R.K. (2008). Detecting exposed test items in computer-based testing.
* Impara, J.C., & Foster, D. (2006). Item and test development strategies to minimize test fraud.
* Nath, L., & Lovaglia, M. (2009). Cheating on multiple choice exams.
* Aiken, L.R. (1991). Detecting, understanding, and controlling for cheating on tests.
* Fask, A., Englander, F., & Wang, Z. (2014). Do online exams facilitate cheating?
* Yi, Q., Zhang, J., & Chang, H.-H. (2008). Severity of organized item theft in computerized adaptive testing.
* van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing.
* Shu, Z., Henson, R., & Luecht, R. (2013). Using deterministic, gated Item Response Theory model to detect test cheating due to item compromise.

**Evaluating Dimensionality**

* Davenport, E.C., Davison, M.L., Liou, P.Y., & Love, Q.U. (2015). Reliability, dimensionality, and internal consistency as defined by Cronbach: Distinct albeit related concepts.
* Green, S.B., & Yang, Y. (2015). Evaluation of Dimensionality in the assessment of internal consistency reliability: Coefficient alpha and omega coefficients.
* Sijtsma, K. (2015). Delimiting coefficient α from internal consistency and unidimensionality.

**Score Reports**

* Knoetze, J., & Vermoter, C.-L., (2007). Communication of psycho-educational assessment results: Remedial teachers’ reading of assessment reports.
* Trout, D.L., & Hyde, E. (2006). Developing score reports for statewide assessments that are valued and used: Feedback from K-12 stakeholders.
* VanWinkle, W., Vezzu, M., & Zapata-Rivera, D. (2011). Question-based reports for policymakers.

**Supporting IRT Score Interpretation**

* Dorans, N.J. (2004). Using subpopulation invariance to assess test score equity.
* Dorans, N.J., & Holland, P.W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case.
* Kolen, M.J. (2004). Population invariance in equating and linking: Concept and history.
* Bock, D.R., Thissen, D., & Zimowski, M.F. (1997). IRT estimation of domain scores.
* Maydeu-Olivares, A. (2015). Evaluating the fit or IRT models.
* Thissen, D. (2016). Bad questions: An essay involving item response theory.
* van der Linden. (2009). Conceptual issues in response-time modeling.
* Wise & Demars. (2006). An application of item response time: The effort-moderated IRT model.
* Schnipke & Scrams. (1997). Modeling item response time with a two-state mixture model: A new method of measuring speediness.

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

**Chapters 1-3**

In the first chapter, the author defines the concepts of cheating. Specially, in this book, cheating refers to violating the rules of the test. The biggest problem caused by cheating is the threat to validity which means the degree to which evidence supports the inference. The test scores are used to make decisions and inferences about a person’s ability. Cheating will impact the accuracy of inferences in many aspects. For example, the students who cheat admitted by universities may have to withdraw due to low GPA. While the students who do not cheat may not be admitted.

In the second chapter, the author presents the frequency of cheating. The main method used to detect the frequency is survey research. The information gathered by the survey includes the frequency of cheating, who cheats and the behaviors related to cheating. The author recommends that to ensure the accuracy of the results of the survey, people should use an anonymous survey. Other methods include asking people the behavior of their friends, observing the behavior of the students as well as experimental designs and some statistical procedures to compute the probability of two students having similar response patterns. The author describes the trends of cheating over the past decades. There are two findings:

* the frequency of cheating is higher in colleges than other levels of education system;
* since the 1960s, the number of students cheating has increased dramatically.

In the third Chapter, the author lists the common ways of cheating, which include traditional methods such as looking at a classmate’s answer sheet, using forbidden materials and changing the answers after being graded as well as advanced method such as using electronic devices.

The author states the basic concepts and problems related to cheating in the first three chapters. No recommendations are discussed. The detailed recommendations about detecting cheating and prevention of cheating are discussed in chapter 7 and 9.

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

**Chapter 7**

In chapter 7, the author provides the methods for detecting cheating. All of the methods are related to multiple choice formats. Multiple choice items are more objective than open-ended questions and much more efficient. Besides, the use of multiple choice items can increase the validity because this format will cover more content. However, it is very easy for students to copy the answer of others taking the test. Students could take a look at the answer key sheet of their neighbor and copy several responses in a minute.

It is not easy to identify cheating for multiple choice format tests. Because similar response patterns, which are the only sign of cheating, could be due to several reasons. Reviewing the content together is the most common causes of similar response patterns. Particularly, the nature of multiple choice items is related to the high probability of two students having similar answers based on pure guessing.

Thus, two series of procedures are used to detect cheating. One is observational method, and another one is statistical method.

* Observational method refers to the observation of student’s behavior. The disadvantage of this is that it is too subjective because the decisions may be based on inaccurate perception.
* The statistical method is objective. The principle of the method is to determine whether the probability of similar response patterns is due to chance.

The author divides the statistical methods into two categories, chance and empirical. Both of them compare the response patterns to other distribution.

* For the chance method, the response patterns are compared to a known distribution, whereas for the empirical method, the response patterns are compared to the distribution of other pairs of students who have no opportunity to cheat.
* The empirical method can be used in large scale testing since there are enough students to find a pair of students who cannot copy. The chance method can be used when the number of students is not large. For example, it can be used in a test held in classroom.

Although the statistical method is objective and seems better than the observational method, the limitations are very obvious. The statistical methods can only be used for multiple choice items and can only detect cheating within pairs. What is more, it cannot detect all of the cases who cheat.

The author states that both the observational method and statistical method have serious flaws. The alternative way to solve this problem is to use alternative forms of the test to prevent cheating.

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

**Chapter 9**

The topic in chapter 9 is how to deter/prevent cheating. The author discusses the strategies from three levels: preventing cheating in a class room, institutionally, building a system to prevent cheating and preventing cheating caused by educators.

At the classroom level, the strategies can be summarized as follows: management of form and content of the test and management of an examination room. For example, alternative forms of tests should be provided. The essay question is better than multiple choice. Before the testing, the seats of the students should be arranged and the identity of students should be verified. The test should be taken in a smaller classroom rather than a bigger one.

At the institutional level, polices should be set to help prevent cheating. These polices should address the environmental requirement of academic integrity, define the behavior of cheating clearly and develop a series of activities. The relevant policies should be accessible to the students and faculty easily. Particularly, the author addresses the use of an honor code. An honor code refers to the principles and procedures which are used to encourage the development of academic integrity. The studies reveal that the effectiveness of an honor code is doubtful. It may reduce the frequency of self-reported cheating but cannot reduce the actual cheating. The reason is that students and faculty are not willing to report inappropriate behavior in test taking due to the existence of an honor code.

The author discusses how to prevent cheating caused by educators or people who are giving the test. For example, in state mandated tests, most of the states set laws to ensure the fairness of the test. The purpose is to reduce the activities threatening validity. For example, practices/preparations are forbidden when the only purpose is improving test scores. As another example, reproduction of the test items for the purpose of preparation is not allowed.

The author cites Cannell’s 1989 list of recommendations about preventing cheating in large scale tests. The strategies can be divided into three categories:

1. Organization of the test (e.g., high stake tests should be organized, arranged and monitored by the state and cannot be held by any schools independently).
2. Security of the materials (e.g., the test booklets should be sealed and should not be distributed until the day of the test).
3. Reporting the test scores (e.g., current norms rather than previous ones should be used).

Besides, the author gives the following recommendations:

1. The content of the course and test should be updated.
2. Computer adaptive testing is preferred to paper-based testing.
3. Open ended questions should be used.
4. The use of take-home exams should be avoided.
5. The consequence of cheating should be clarified and academic integrity should be encouraged.

Foster, D., Maynes, D. and Hunt, B. (2008). Using data forensic methods to detect cheating. In C.L. Wild and R. Ramaswamy (Eds.), *Improving testing. Applying process tools and techniques to assure quality* (pp. 305–321).London, UK: Routledge.

As testing has become the predominant “gatekeeper” to professional achievement in industry and commerce, the motivation to cheat has increased. In fact, the authors acknowledge the following: “cheating and test piracy are prevalent, and usually successful” (p. 306).

The article focuses on the types of cheating behavior that can be addressed by data forensic methods, which as defined by the authors is “an investigative methodology based on the statistical analysis of test and item responses for indications of cheating and piracy” (p. 308). The authors call out the susceptibility of these methods to both type I (falsely detecting cheating) and type II (failing to detect cheating) errors. However, other detection methods share the same susceptibility.

Specifically, data forensic methods were found to be useful in addressing erasures and answer changing, collusion, gain scores and volatile retakes, response aberrance, and retake analysis. Due to the fact that each data forensic statistic is designed to measure independent aspects of cheating, the detection power can be increased by combining statistics, and thus, increasing support for the determination of cheating.

Findings from data forensics evaluations are expressed in probabilities that an examinee’s test response agrees with one or more statistical models of cheating. These findings have been used in courts and have been deemed “credible evidence” to demonstrate that a testing irregularity is present. The presence of this testing irregularity provides testing programs with a reasonable basis to doubt the validity of a given test result (examinee score). This allows a testing organization to cancel an examinee’s score and protect the value of the examination program.

The authors also discuss the fact that testing programs have lagged in using technology to make the tests more secure; however, cheaters have embraced technology to further methods of cheating. This can have serious consequences for testing programs, whereas it costs money to replenish their test items, the damage to an organization’s reputation may be irreparable.

A goal for data forensic analysis is to move the point in time of its occurrence closer to the testing event so that cheating can be detected and stopped as it is happening. This would be extremely effective in deterring cheating.

Han, N., & Hambleton, R.K. (2008). Detecting exposed test items in computer-based testing. In C.L. Wild and R. Ramaswamy (Eds.), *Improving testing. Applying process tools and techniques to assure quality* (pp. 323–348).London, UK: Routledge.

Han and Hambleton begin the article by noting that exposed test items are a major threat to the validity of computer-based testing. Most of the research in this area has focused on finding ways to minimize item usage. The researchers propose that instead, the focus should be on the generation and investigation of item statistics that can reveal whether test items have been exposed to examinees.

Han and Hambleton designed the current study to extend the research into item exposure detection statistics by also taking into account the presence of ability shifts over time. Previous research had been based on the assumption that examinees’ ability over time was stationary. The researchers also examined these detection statistics over various item performance levels (difficult, easy, and low and high discrimination).

Using simulated data sets, Han and Hambleton constructed a linear test of 75 items with item parameters consistent with item statistics in a national credentialing exam. The number of examinees used was 5,000 with three different ability levels assumed. The item exposure detection statistics examined were moving p-values, moving averages of item residuals and moving averages of standardized item residuals.

The researchers found that with fixed normal ability each of the three item exposure detection statistics were stable. When the ability distribution was shifted in the simulated data the p-value statistics shifted as well, therefore it was determined that p-value shifts are confounded with shifts in ability distributions and do not reflect item exposure because there was no item exposure in that particular simulation.

Overall, an advantage was found for the two IRT-based statistics. As noted above, item p-value was not applicable in all situations. The IRT-based statistics typically identified exposed items with the exception of when only a small amount of exposure was simulated. The researchers also discovered that it was easiest to detect more difficult items that were exposed and also the more discriminating items.

Han and Hambleton are encouraged that the item exposure detection statistics are applicable across all forms of computer-based testing (linear, linear-on-the-fly, multi-stage, and fully adaptive). The researchers recommend that future research focus on just one of the detection statistics—item standardized residuals—and investigate additional item exposure models.

Impara, J.C., & Foster, D. (2006). Item and test development strategies to minimize test fraud. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 91–114). Hillsdale, NJ: Lawrence Erlbaum.

Test fraud is a serious threat to the validity of test scores. The scores of cheaters do not reflect what these individuals know and can do. This is not only a measurement concern, but can also cause damage to the reputation of the testing organization.

Impara and Foster define the two classes of test fraud, cheating (obtaining a score through inappropriate means), and piracy (stealing items, item pools, or tests). They then extend the discussion to a review of item writing and test development strategies from the perspective of how these strategies can help to minimize test fraud.

To avoid memorization and potential piracy of items the authors recommend using novel material to test higher-level concepts and the avoidance of overly specific and overly general content. These recommendations are in line with general item-writing strategies. However, there were a few strategies with which Impara and Foster did not agree. Specifically, minimizing the amount of reading in each item, and avoiding “window dressing” or excess verbiage. The authors argue that following these guidelines can lead to the item being easier to memorize. They do acknowledge that not following these guidelines can cause potential validity problems with the content not reflecting the construct of interest.

In addition to the traditional multiple-choice item types, the authors also discuss cheating and piracy in relation to performance-based items. Impara and Foster caution against using performance-based items that can be easy to remember this can be especially difficult for an organization having limited funds and unable to rotate in new items on a regular basis.

The authors also reviewed test assembly strategies that can assist in combating test fraud. Among the recommendations is the randomization of response choices, the use of multiple forms, using LOTF (linear-on-the-fly) or adaptive models, and using item variants or clones on alternate forms.

The authors openly acknowledge that sometimes there is a conflict between “good psychometrics and good security.” Test developers are encouraged to give careful thought to item writing and test design in order to protect the investment of the organization and to ensure a fair testing environment for the examinees.

Nath, L., & Lovaglia, M. (2009). Cheating on multiple choice exams: Monitoring, assessment, and an optional assignment. *College Teaching*, *57*(1), 3-8.

According to Nath and Lovaglia (2009), a large proportion of cheating behavior exists in academic institution, but only a small percentage of cheaters are identified. In this paper, the authors provided methods to detect cheating and prevent students cheating. They listed several factors may have impact on students’ cheating behavior:

* Younger students cheat more than older students
* Non-married students tend to cheat more than married students
* Cheating may have relation with majors.

The authors modified Harpp and his colleagues’ program (Harpp & Hogan 1993; Harpp, Hogan, & Jenings 1996), which calculates the ratio of choosing the identical wrong answers divided by missing the same question. If the ratio is larger than .75, students were identified as cheating. Instead of comparing all students’ answer sheet, the authors only estimate the probability of two students sitting next to each other. This method would reduce the mistake of false accusations. Identified student will receive an interview with teacher. Most of students would deny their cheating on the exam until teachers showed the statistical evidence.

Across five semesters of applying interview method, cheating cases were significantly dropped to zero. For students who admitted cheating, they provided an optional assignment, which is writing a research paper on the topic of “Ethics and Social Relations: The Function of Honesty and Responsibility in a Complex Society”. In their study, almost all of the students completed the optional assignment to make up the lost exam points. Only one student failed to complete and left the university.

The authors mentioned that a desired normative atmosphere which encourages honesty is essential to prevent cheating. It’s important to avoid false accusation or false confessions. The authors recommended the use of their methods to redirect students, but not for use in legal situations.

Aiken, L.R. (1991). Detecting, understanding, and controlling for cheating on tests. *Research in Higher Education*, *32*(6), 725-736.

Aiken(1991) used questionnaires to investigate the occurrence of cheating by students. Also, the author talked about cheating by teachers and the reasons they cheat. In this paper, the author provided two programs to detect cheating and a method to prevent cheating.

The author distributed the questionnaire on test cheating to 200 students enrolled in two universities. One is a relative small church-affiliated university, and another one is a large state university. The analysis was conducted by chi-square analysis and reported:

* Males tend to cheat more than females.
* Students with strong religious convictions were less likely to cheat on tests.
* Students are less like to cheat if the instructors are friendly and considerate.

Two computer programs to detect cheating were designed by the author. The first program, Cheat-1, used a chi-square test to decide if two students answered items similarly. The value of chi-square in a 2\*2 table includes the number of students who answered the first and second items right or wrong, the first right and the second wrong, the first wrong and the second right, or both wrong. A much larger chi-square value than those obtained from other pairs of students would be considered as cheating and involved further investigation.

Cheat-2 used the information obtained from Cheat-1 to determine the probability of two examinees selecting similar answers by the binomial distribution. The author believed that different test forms would be problematic in equating test forms. Instead of using different test forms, the author suggested using multiple answer sheet forms for one test. This would prevent two adjacent students from cheating from each other.

At last, the author mentioned the fact of tampering by teachers of students’ answer sheets to increase test scores. The reasons may be the social pressure, such as evaluations of schools, media report, principles and so on.

Fask, A., Englander, F., & Wang, Z. (2014). Do online exams facilitate cheating? An experiment designed to separate possible cheating from the effect of the online test taking environment. *Journal of Academic Ethics*, *12*(2), 101-112.

The authors conducted an experimental design to investigate students cheating in an online testing environment versus traditional test environment. The motivation of this paper is the growth of online courses and online exams in academic institutions. Some faculties perceived online testing as a greater opportunity for cheating than traditional classroom environment. The authors used two regression models to analyze the separate effect of testing environment on exam performance from the distinct effect of a possible greater opportunity to cheat on unproctored, online exams on exam performance.

44 participants were from two elementary statistic classes. Both classes had the same topics, lectures, materials, midterm exam and assignments in a traditional class environment. The only difference is that one class of students took the practice test and the final exam online, whereas the other class took the tests in a traditional classroom environment.

Students were told that the practice test was similar to the final exam and did not count in the final scores. Therefore, the practice test should be relatively free of cheating.

The first regression model only investigated the impact of the exam environment on performance, where the scores of the practice test were predicted by testing environment and students’ GPA. The result showed that students in the classroom-environment testing situation performed better than those in the online testing environment.

The second regression model studied the cheating behavior on the final exam, where the scores of the final exam were predicted by testing environment, mid-term score, practice test score, and accumulative credit. The result showed that online testing scores were much higher than traditional classroom test scores, which indicated that the final exam was likely the result of cheating in class with the online testing environment.

The author suggested that professors and deans should suppress student cheating in those online courses and assure students’ grade are reflective of their learning.

Yi, Q., Zhang, J., & Chang, H.-H. (2008). Severity of organized item theft in computerized adaptive testing: A Simulation Study. *Applied Psychological Measurement*, *32*(7), 543–558. doi:10.1177/0146621607311336

One of the main security concerns for computerized adaptive testing (CAT) is the issue of item exposure and item sharing. In this paper, the authors investigated the detrimental effect of organized item theft to CAT through a simulation study.

Organized item theft means some test takers (called “thieves” in this scenario) would memorize test items from a CAT administration and then share these items with future test takers, causing these items to be compromised. For different CAT item selection and exposure control methods, the effect of organized item theft may be vastly different. There has been research that derives theoretical results based on the randomized item selection method, which is rarely used in real-world CAT operations.

This study compared the performance of two realistic item selection methods with either systematically or randomly occurring thieves, and the damage of item theft was measured by the total number of compromised items in the item pool, the average number of compromised item a test taker may encounter, and the measurement precision for IRT person parameters.

Their results showed:

* out of 1,000 test takers, 30 thieves who memorized one quarter of the test items was enough to compromise 40-50% of the entire CAT item bank
* half of the items encountered by future test takers were compromised
* the estimated person parameters were larger with compromised items, especially for low-achieving test takers
* -stratified item selection was more robust with regard to organized item theft than the commonly-used maximal item information method.

In conclusion, the authors suggested using methods that select items more evenly and moderately increasing the size of item bank as safeguards against organized item theft.

van der Linden, W.J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*(3), 365–384. doi:10.1007/s11336-007-9046-8

Aberrant response behavior from test takers may indicate cheating. Conventional methods focus on developing person-fit indices to check for unusual pattern in test takers’ responses. Computerized Adaptive Testing (CAT) makes it possible to record the time each test taker spends on an item, providing additional information to help detect various cheating behavior.

In this study, the authors used Bayesian statistical procedure to derive a posterior predictive distribution of the response time of a test taker on any arbitrary item given the response time on all the other items. This was done by specifying one parameter for test taker’s speed, one parameter for the time intensity of an item, and a third parameter similar to the discrimination parameter in the item response theory models. Based on this model, it is possible to calculate the mean and variance of the predicted response time (in this case, the logarithm of response time) on each item by the same test taker and conduct further statistical analysis, for instance, constructing prediction intervals, accordingly. The authors also derived a joint distribution that synthesizes test takers’ responses as well as RT, although they did not illustrate its application in detail in the latter section of the paper.

The use of response time in combination with responses can be a promising approach in checking for test security violation. The authors identified two forms of cheating that are amenable to the new method. Erratic response and random response time may suggest a test taker is memorizing instead of answering the item; and unreasonably short response time and correct response points to the possibility of item preknowledge.

As an illustration, the authors applied the procedures to a case study based on the real data from the GMAT®. After estimating the model parameters, the predicted mean and variance were computed and used to standardize the observed log response time for a given test taker. Any item with a standardized residual lager than 1.96 or smaller than -1.96 would be flagged as problematic. This enabled us to pinpoint the test takers with too many unusual response times as well as monitor item performance over time. By examining the actual response and response time pattern of the flagged test taker, we may come to some tentative conclusions regarding the cause of the aberrant performances. In the example provided by the authors, the response time showed a clear time effect, strongly suggesting time mismanagement rather than cheating.

It should be noted that response time must not be used as the sole determining factor in cheating detection. As is shown in the above example, reasons other than cheating may lead to unexpected response time. Therefore, the authors recommended using response time as a routine check only to flag potential questionable cases for further review, and other information and evidence are always needed to actually prove or disprove the occurrence of cheating.

Shu, Z., Henson, R., & Luecht, R. (2013). Using Deterministic, Gated Item Response Theory Model to detect test cheating due to item compromise. *Psychometrika, 78*(3), 481–97. doi:10.1007/s11336-012-9311-3

In this paper, the authors proposed a new model (Deterministic, Gate Item Response Theory Model, DGM) that is able to quantify two separate abilities from the test takers: their true abilities and their abilities gained from cheating in a test. This was done by first classifying items as being exposed or unexposed. An unexposed item was assumed to be safe even if a test taker had preknowledge of the some test items from the item bank. Another distinction made in the model was cheater vs. non-cheater, and cheater is defined as having a cheating ability higher than the true ability.

Each classification was represented in the model by a dichotomous indicator variable. Only under the condition where a compromised item is encountered by a cheater should the cheating ability be used to model the success probability of the test taker on that item. The success probability was based on the Rasch model. Using Markov Chain Monte Carlo (MCMC), it is possible to obtain an estimate of cheating indicator variable (a value between 0 and 1), which can be read as a probability statement on how likely a test taker has a score gain on the exposed items. A cut-point may also be chosen to force a decision of cheater vs. non-cheater.

As an illustration, the authors first conducted a simulation study in which the level of score gains from cheating, the proportion of exposed items, and the proportion of cheaters were varied. The sensitivity (true positive) and specificity (true negative) of decisions were calculated and compared with the l\_z index, a previously developed statistic for detecting item preknowledge. The results showed that DGM had similar specificity but better sensitivity, and it could correctly identify as many as 80% of cheaters when both exposed items and cheater proportion was low.

A real data illustration was provided next as proof of feasibility of DGM in real-world testing situations. The model converged properly and about 9% of the test takers from the dataset were flagged as cheaters. It is impossible to judge its validity, but the authors did show that most of the cheaters are low or medium ability students, an observation that is realistic lends support to the model.

The DGM does not assume cheaters to have a 100% probability of correctly answering a compromised item, and it allows practitioners to obtain a “cheating probability” for each test taker. The use of DGM still involves some subjective judgements, including specifying the exposed status of each item and selecting cut-points for classifying cheater, either of which would determine the validity of the conclusion drawn from model estimates. Another problem with DMG is that its performance drops when most of the test takers are cheaters, hence it may not be effective for a testing program that already suffers from severe test security issues.

Davenport, E.C., Davison, M.L., Liou, P.Y., & Love, Q.U. (2015). Reliability, dimensionality, and internal consistency as defined by Cronbach: Distinct albeit related concepts. *Educational Measurement: Issues and Practice*, *34*(4), 4-9.

What is the problem or issue being addressed by the authors?

* **Cronbach's (1951) coefficient alpha is** the most commonly used *index* of reliability
* The biggest misconception: "alpha serves also as indexes of internal consistency and unidimensionality"

***Cronbach's Definition***

* Reliability: the accuracy or dependability of measurements
* Internal consistency: the degree to which items measure the same thing
* Dimensionality: "for a test to be interpretable, the items need to have a large first principal factor saturation, but the common factor structure accounting for the item covariances need not be unidimensional" (Cronbach, 1951).

How did they investigate or go about solving it?

* Uses definitions of coefficient alpha provided by Cronbach in his seminal paper
* Show the concepts of reliability, dimensionality, and internal consistency are distinct BUT interrelated
* Begins with the critique of the definition of reliability
* Explores mathematical properties of Cronbach's alpha
* Then internal consistency and dimensionality are discussed based on the Cronbach's definition
* Functional relationships are given: relate reliability, internal consistency, and dimensionality
* Demonstration of the utility of these concepts as defined

What are their findings or recommendations?

* Reliability, internal consistency, and dimensionality should each be quantified with separate indices.
* Internal consistency is a fundamental property of the items that does not depend on the number of items. Internal consistency affects alpha in two ways:
  + First, internal consistency must be positive
  + Second, if positive the level of internal consistency determines the number of items needed for any desired level of alpha
* BUT it is important to recognize that those three concepts are interrelated.
* High levels of unidimensionality and internal consistency are not necessary for reliability as measured by alpha NOR for interpretability of test scores.
* Alpha, computed using multidimensional items as the units of analysis, will likely underestimate score reliability, because items are not true score–equivalent. Alpha computed using true score–equivalent parcels as the unit of analysis or using coefficients based on structural equation modeling can provide a more accurate estimate of score reliability.

Green, S.B., & Yang, Y. (2015). Evaluation of Dimensionality in the assessment of internal consistency reliability: Coefficient alpha and omega coefficients. *Educational Measurement: Issues and Practice*, *34*(4), 14-20.

This article is a commentary article to the Davenport, Davison, Liou, & Love (2015) article.

What is the problem or issue being addressed by the authors?

* Argue that factor analysis should be conducted before calculating internal consistency estimates or reliability.
* If factor analysis indicates the assumptions underlying coefficient alpha are met, then it can be reported as a reliability coefficient. However, to the extent that items are multidimensional, alternative internal consistency reliability coefficients should be computed based on the parameter estimates of the factor model.
* The authors also agree with the main issue addressed by Davenport et al. (2015) that we encounter misinterpretations of coefficient alpha as an index of internal consistency or of unidimensionality (i.e., homogeneity) in the educational and psychological literature.

How did they investigate or go about solving it?

* Choose to view coefficient alpha within a modeling framework (i.e., viewing reliability within a modeling process).
* Prefer to model the relationships among items and choose an internal consistency estimate of reliability that is consistent with the chosen model and the intended purpose of the measure.
* Provide formulas for computing reliability coefficients based on parameter estimates from a factor analytic model. These estimates can be determined using either confirmatory factor analysis (CFA) or exploratory factor analysis (EFA).

What are their findings or recommendations?

* Assuming a bifactor model evidenced good fit, and the measure was designed to assess a single construct, omega hierarchical –the proportion of variance of the total scores due to the general factor –should be presented.
* Omega –the proportion of variance of the total score due to all factors –also should be reported in that it presents a more traditional view of reliability, although it is computed within a factor analytic framework.
* By presenting both these coefficients and potentially other omega coefficient, the reliability results are less likely to be misinterpreted.

Sijtsma, K. (2015). Delimiting coefficient α from internal consistency and unidimensionality. *Educational Measurement: Issues and Practice*, *34*(4), 10-13.

This article is a commentary article to the Davenport, Davison, Liou, & Love (2015) article (DDLL).

The author argues that there is no convincing reason to connect the three concepts that are reliability, internal consistency, and unidimensionality.

***Internal Consistency***

* The definition of internal consistency as the mean interitem correlation, , may be historically correct, but the definition ignores that highly different factor structure may underlie the same mean interitem correlation whereas different one-faction structures may underlie highly different mean interitem correlation.
* The variation among the interitem correlations and the patterns they constitute are precisely what might be interesting for understanding what the test measures, more than the central tendency of the distribution of the inter-item correlations.

***Reliability and Coefficient alpha***

* For reliability to be high, items do not have to correlate highly or even positively. What is necessary is that random measurement error is small and items thus provide stable scores across administrations, so that test scores based on the items are stable.

***Unidimensionality***

* The one-factor solution reflects perfect unidimensionality that is largely hidden from the eye by a heavy load of systematic error and random measurement error –that is, a weak but unidimensional signal and loud noise.

The author says that as DDLL illustrated in their paper, adding meaning to math can help students to avoid abstract thinking and help them to feel more comfortable by gaining a frame of reference. However, he also expresses concerns on mixing mathematical quantities and real-world phenomena since it is difficult to be aware of the relationship all the time. The author suggests that we shouldn’t ignore psychometrics so that it becomes meaningless for measurement of attributes and points out that this is where validity comes in.

Finally, the author suggests three take-away message:

* Use coefficient alpha as a lower bound to the reliability and do not relate alpha to internal consistency and unidimensionality.
* If a theory of the attribute is available and implies that items measuring the attribute have high internal consistency or are unidimensional then use these requirements to select items for the test; else use other requirements the theory implies.
* If a theory is unavailable, by studying the items selected or rejected based on high internal consistency or unidimensionality the researcher may learn about the attribute and its measurement.

Knoetze, J., & Vermoter, C.-L., (2007). Communication of psycho-educational assessment results: Remedial teachers’ reading of assessment reports. *South African Journal of Education, 27*(1), 1-18.

**Purpose**

Through focus groups and in-depth interviews, the study sought to understand remedial teachers’ interpretation and use of psycho-educational assessment reports.

**Background**

In order to determine whether learning supports are needed for particular students, teachers ask psychologists to conduct a diagnostic assessment and make recommendations for instruction.

**Participants and Methods**

Ten remedial teachers participated in the focus groups with two of the teachers then participating in an in-depth follow-up interview. The full process entailed:

* 1. Participants read a typical psycho-educational report prior to the focus group
  2. Focus groups conducted and common themes identified
  3. In-depth interviews provided depth and clarification of the common themes
  4. Transcripts of the focus groups and interviews analyzed with Grounded Theory
  5. Themes and analysis verified by participants

**Themes from the focus groups and interviews**

1. Reports are often too vague and lack specificity for improving classroom instruction
2. Inclusion of IQ scores are useful for revealing if students’ difficulties are subject specific
3. Reports often used as confirmation of teachers’ intuitions about a student
4. The language is often filled with psycho-jargon instead of educational language
5. Due to time constraints, teachers prefer more information on what should be done next and less on how the results were derived
6. There is a disconnect between educators and psychologists

**Recommendations to improve appropriate interpretation and use of reports**

* Improving teacher and remedial teacher training in psychometrics to increase their ability to understand the assessment process and results
* Providing an avenue for teachers to be certified administrators of the assessments, thus removing the need to seek an outside evaluator (i.e. psychologist). This could improve teachers’ interpretation, decrease the number of requests for formal assessment by increasing teachers’ confidence in their initial judgments, and decrease teachers’ dependence on IQ scores as their only useful source of information.
* Improve psychologist training in educational contexts, thus enabling psychologists to produce recommendations with greater specificity and relevance to classroom instruction.

The authors seemed to interpret teachers’ comments as an indictment of the broader teacher-psychologist relationship rather than focus more exclusively on improving score reports. Thus, while the authors recommendations do impact how reports are constructed and interpreted, the changes are due to alterations in teacher and/or psychologist training.

Trout, D.L., & Hyde, E. (2006, April). *Developing score reports for statewide assessments that are valued and used: Feedback from K-12 stakeholders*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. Retrieved from personal correspondence.

**Purpose**

Conducted focus groups and interviews to determine how statewide standardized assessment results are used by various stakeholders and what information is most valued on score reports.

**Participants and Methods**

* Phase I asked participants to identify tasks and workflow for which state assessment results would be used and feedback on score report design. Phase II used the initial feedback to create a paper and web-based report prototype and test its usability.
* Conducted focus groups and in-person or phone interviews with teachers, principals, state and district testing administrators, and parents from multiple states and representative of school size, ethnic diversity, and urbanicity.

**Results**

* *State and district administration* – need results fast with clear information in order to respond accurately to media/public requests, as well as make recommendations to schools
* *Principals* – use for comparing class results in order to offer assistance to teachers and build environment of cooperation
* *Teachers* – want scores as soon as possible after testing, but also emphasized their limited time to read and interpret results. Thus, reports need to be easily accessible. Also want large pools of practice items released to help with prep and correcting student mistakes. Additionally, need help understanding scale scores so they can communicate results clearly to parents and students, preferably through graphics
* *Parents* – concerned primarily with student scores and want more specific strategies for improving weaknesses. Also liked graphical comparison information.

**Recommendations**

* Given the variety of uses, reports should be tailored for each stakeholder group
* Color is a valuable tool for highlighting positive and negative areas, but choice of color should be color-blind friendly and convert well to black and white printouts
* Use of terminology should be minimized, and when necessary should be accompanied by definitions
* Teachers valued distractor analysis so they know which students, and the class as a whole, were making certain kinds of mistakes
* Teachers also willing to have smaller font to allow for more information than other stakeholders
* Information should be presented in a way that is actionable
* Web-based report allows for specific information to be found quicker; however, stakeholders also want paper reports available.
* Separate parts of the website should be for younger students, older students, and parents and include links to learning activities and resources.

VanWinkle, W., Vezzu, M., & Zapata-Rivera, D. (2011, September). *Question-based reports for policymakers* [Research Memorandum]*.* Princeton, NJ: Educational Testing Service.

**Purpose**

Conducted a usability study of a question-based online score report with seven administrators to provide feedback on further development of the reporting system.

**Initial Report Design**

The authors used suggestions from Goodman & Hambelton (2004), Hattie (2009), and Mayer (2001):

* Minimize scrolling by including navigation panes
* Use color in a purposeful way, such as displaying performance levels
* Present most important and most general information first, followed by more details
* Pair related graphics and text in a manner that maintains spatial contiguity and helps users make appropriate interpretations and use of scores
* Include comparisons between different groups
* Include a Key Findings and Results section that clearly and concisely summarizes data
* Provide “What’s this?” rollovers (mouse hovering over item) to define terms and provide additional information for those interested
* Include a section that clearly documents the purpose, use, and interpretation of the test

Along with the above suggestions, the authors designed the report around questions administrators typically want answered by score reports. For instance, “What are the results for my district for the 8th grade test in all subjects?” Administrators can then change the group, grade, and subject to fit their specific question.

**Usability Study Participants and Methods**

Seven administrators were guided through six different reports produced from the online score report system based on six unique questions. Four assessment experts were then asked for additional feedback and recommendations.

**Results and Recommendations**

* Participants praised the “What’s this?” rollovers, graphic legends, and Key Findings and Results sections.
* Some terminology needed to be adjusted and more definitions needed to be added (such as what a boxplot represented)
* Increase interactivity, for instance, by receiving exact numbers when clicking on sections of a graphic
* Bar graphs should be vertical rather than horizontal
* Colors should be used in shades for proficiency and in contrast for highlighting strengths and weaknesses
* More information regarding the context of scores should be provided, including attendance.
* More emphasis should be placed on standard errors, especially when comparing groups

Dorans, N.J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*(1), 43-68.

It is important to get fair measurement and also fair use. Test fairness has a long history. People who are evaluated want to be evaluated fairly. Test results have indicated that there are noticeable differences between different groups (e.g., males and females on math test). Tests are criticized because they are used inappropriately, or measured the wrong things. Achieving subpopulation invariance is very important to test fairness. The violation of subpopulation invariance can be account for selection effects, misspecification errors, and criterion issues.

Several efforts have been made to achieve invariance. First, differential prediction studies examine whether the same prediction models hold across different groups. To assess fair selection, the Regression or Cleary model (Cleary, 1968) was a preferred model that fair selection of applicants is consistent with fair rejection of applications. It also examines whether the regression of the criterion on the predictor space is invariant across subpopulations.

Fair prediction is difficult to achieve. The prediction model must be the appropriate model; misspecification of the model will introduce statistical bias. The prediction model is appropriate if all the predictors needed to predict criterion scores are included (e.g., high school grades and test scores to predict college grades), and the functional form used to combine the predictors is the correct one. One prerequisite for this is that the criterion itself should be reliably and fairly measured. Even though the correct equation is correctly specified in the full population, invariance may not hold in subpopulations because of selection effects. One must keep confounding influences in mind to assess whether a prediction equation is invariant across subpopulations.

Second, differential item functioning (DIF) brings the fair assessment to item level. The regression of item score onto the matching variable needs to be invariant. DIF examines how performance on item varies across total score; lack of DIF means the relationship is the same across subpopulations.

Subpopulation invariance is the most critical requirement to score equity assessment (SEA). Score equating is a statistical process that produces scores that are considered comparable enough across test forms to be used interchangeably. SEA focuses on invariance at the reported score level. Checking SEA via subpopulation invariance of equating functions could serve as a quality-control check to ensure that well-developed test assemblies remain within acceptable tolerance levels with respect to equitability. If a SEA determines that a test and score production process is out of control and not producing the degree of exchangeability expected, then follow-up analyses, such as DIF, should be taken into consideration.

For subpopulations, we can examine SEA for gender, ethnic/race groups, and also language proficiency groups. Once one begins to all these groups, and add in others, such as grade levels, the number of subpopulations becomes large. Comparing linkings in each of these subpopulations to the total group linking function can be done, but the interpretation would be difficult. One could simply do a single linking for the total group and assess whether that linking holds for each of the subgroups. It can be used to identify which subpopulations may differ from the total population.

Overall, SEA examines only one aspect of fairness. DIF and differential prediction are other aspects that merit attention. All three aspects of fairness should be addressed.

Dorans, N.J., & Holland, P.W. (2000). *Population invariance and the equatability of tests: Basic theory and the linear case*. Princeton, NJ: Educational Testing Service.

Test equating is designed to remove the effect of unintended differences in the relative difficulty of test versions and makes across-version comparisons of examinee results meaningful. Test equating has received a wide audience because of a desire to give wider interpretability to test results. For test equating, there are five requirements: equal construct, equal reliability, symmetry, equity, and population invariance.

The most common way to check the equal construct requirement is the inspection of the content and wording of the test questions. In another word, examine the two tests to see if they appear to competent judges to be measuring the same thing. One test of the equal construct requirement is that the two tests line up different subgroups of examinees in exactly the same way in terms of their mean test scores.

The equal reliability requirement is often violated. Practical circumstances often lead to tests of high but clearly different reliabilities being equated with no evidence of problems. However, we believe it is not a fundamental requirement of test equating. It can be violated, while equating can still be satisfactory. In addition to be concerned about equal reliability of tests being equated, attention should be focused on the amount of reliability – more reliability is better for equating.

Regarding to the symmetry requirement, the usual definitions of the linear and equipercentile equating functions automatically insure that it is satisfied. This requirement does not appear to be very fundamental, and rarely provides useful guidance in selecting one linking function over another.

The equity requirement has two parts. One is concerned with average or expected test performance, and the other is concerned with aspects of performance that go beyond means and expected performance. Examinee ought to expect the same score on either one of two equated tests. Besides, an examinee ought to have the same expected distribution of performance on either one of two equated tests.

If the tests that measure different things or that are equally reliable are linked using standard linear or equipercentile methods, the results will not be invariant across certain subpopulations of examinees. This makes us regard population invariance as one of the most important of the five requirements. The most important thing about the population invariance requirement is that by computing linking functions on subpopulations and comparing them, we can quantify this invariance condition. When population invariance fails to hold to a sufficient degree, it weakens the claim about the link between two tests. A single linking function is inappropriate when population invariance fails to hold.

Much more needs to be done to specify the choice of subpopulations to use in computing our measures. Discovering and describing these subpopulations can be an interesting research area. The important subpopulations are those that reveal the two tests as differentially aligned with the relevant educational experiences of subpopulations. An important question is how to identify such groups of examinees, and how to test opinions in a convincing way.

Kolen, M.J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement, 41*(1), 3-14.

The goal of test form equating is to be able to use scores from alternate forms interchangeably. To achieve this goal, the function used to convert performance on each alternate form to the score scale should be the same for various populations or subpopulations of individuals. Populations or subpopulations are defined by characteristics of examinees, such as gender, race, or region of the country. If the equation functions for populations or subpopulations differ systematically, the interchangeability of test forms is questionable. We are also concerned about the property of invariance in linking, which refers to relating scores on tests when the test are not built to the same content or statistical specifications.

As early as 1951, Flanagan believed that it was important to conduct equating on examinees closely representing the examinees taking the test operationally. Angoff (1971) argued that an equating function should be thought of as a conversion from one set of units to another in physical measurement. The perspective of unique conversion led him to the position that equating should be population invariant. Both Flanagan and Angoff agreed that these linkings are expected to be population dependent. There was empirical evidence that the more two tests whose scores were to be linked differed in content, the greater the population dependence of the linking relationships.

Braun and Holland (1982) explicitly incorporate the population into the definition of equating. Observed score equating methods (e.g., linear and equipercentile) must be defined with regard to a specific population. Lord and Wingersky (1984) indicated that under an IRT model, true score equating is invariant for subpopulations within a population. Even though the true scores are unknown and there is no theoretical reason to apply the true score relationship to observed score, it is often done in practice. Van der Linden (2000) suggested that population differences in observed score equating functions are expected to be minimal if the test forms to be equated are every similar to one another.

Population invariance can be used as a criterion for evaluating how well equating performs under different equating conditions. If the test forms are constructed to be very similar in content and statistical characteristics, then equating functions can be expected to be similar across populations. This seems to be most likely to hold when test forms are built to well-developed tables of content and statistical specifications, the forms are long enough to provide a reasonable and complete sampling of the test content, and the development process leads to alternate forms that are very similar in content and statistical properties.

Linking studies often are very difficult to conduct, and there are a host of factor that can affect the results. Data for the linking studies are typically based on students who have taken both instruments. Issues that arise during the interpretation of results include the representativeness of the examinees that take both tests, the amount of time that has elapsed between taking the two tests, and the order in which the tests are taken. These factors complicate the interpretation. Doran and Holland (2000) proposed that the greater the differences in test content between the tests to be linked, the more the links are population dependent.

Standard errors of equating and linking have been used to judge whether the function differ across populations. It was also proposed to use significance tests to test for differences in distributions of scale scores.

Bock, D.R., Thissen, D., & Zimowski, M.F. (1997). IRT estimation of domain scores. Journal of Educational Measurement, 34, 197-211.

In this article the authors present a way to transform the IRT scale score to obtain an IRT domain score. The idea is to relate the IRT scale score to the percent-correct score from classical test theory. In classical test theory a test is thought of as a random sample of items from a specific domain. Hence, the interpretation of the total score or percent-correct score as the proportion of the domain mastered. However, in the IRT framework we do not have that concept of random sampling items from a domain. Instead we have a set of calibrated items and an IRT scale score.

The authors argue that especially in educational measurement it makes sense to have IRT scores that can be interpreted as domain scores just as in classical test theory. A linear transformation is used to report the IRT scale score as IRT domain percent-correct score or IRT domain score.

Three advantages of using these scores are:

1. the items of a test do not need to be a random sample from the domain,
2. the method is generalizable to multidimensional IRT, and
3. the IRT domain score is more accurate than the test score from classical test theory.

The authors present a series of examples to empirically demonstrate the applicability of the IRT estimated domain scores to diverse type of tests. The paper finalizes with a discussion of when reporting IRT domain scores would be pertinent.

The situations in which the authors recommend the use of IRT domain scores are “in the context of student qualification, and also in accountability assessment, because of their more straightforward interpretation” (pp. 209). Additionally, the authors discourage the use of domain scores in college selection testing (since the aim is just to rank people) and for empirical research and data analysis.

Maydeu-Olivares, A. (2015). Evaluating the fit or IRT models. In S.P. Reise & D.A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 111-127). New York, NY: Taylor & Francis (Routledge).

Maydeu-Olivares talks about the importance of goodness of fit (GOF) indices in a statistical model in general, and for IRT models in particular. GOF statistics are used to measure how well the model fits the data. Before jumping into any conclusion, Maydeu-Olivares argues that we should always take a look to the GOF of the model first. He reviews some classical approaches to measure GOF in categorical models in general (IRT is a special case of these type of models), and also some new procedures.

GOF statistics usually used for categorical models (such as traditional chi-square test) cannot be used for IRT models because of the large number of degrees of freedom. Maydeu-Olivares defines limited and full information GOF statistics and discusses their use in IRT models. He recommends GOF statistics for IRT models, large and sparse ordinal data, and a way to approximate GOF for tests of patient-reported outcomes. Additionally, piecewise GOF assessment is also a good practice. This is done usually after analyzing an overall GOF index. Piecewise GOF considers pieces or parts of the whole models and looks at the GOF of each part separately. Maydeu-Olivares argues that this is important since the overall model might have a good fit but it can have terrible fit in some parts.

The author provides an empirical illustration of several GOF procedures using data from the PROMIS depression short form. In summary, this article presents several GOF indices, discusses their applicability for different scenarios in IRT modeling, advantages and disadvantages.

Thissen, D. (2016). Bad questions: An essay involving item response theory. *Journal of Educational and Behavioral Statistics*, *41*, 81-89.

In this article, Thissen discusses several questions that he argues are bad questions to ask in the context of IRT models. He sees the need of addressing them since these bad questions “are asked by intelligent, well-trained persons” (pp. 81), and that clearly calls for better measurement training. What these questions have in common is that they are usually thought of as yes/no (black/white) questions, when in reality we will have shades of grey. Here are the questions in the article and a brief summary of the discussion.

* Does the IRT model fit? This alludes to a yes/no answer; therefore we always should answer it with a no. Thissen says that even if we get good model fit, this would mean that we do not have enough information and that the result is a way to say that with the data in hand no more complex models are supported. He also argues that under some context we shouldn’t worry about goodness of fit, for example, when testing items or computing comparable scores.
* Is this test unidimensional? Similarly to the previous question, this is a yes/no, black/white question and usually reality could accommodate several answers.
* Are IRT scores estimates on an interval or ordinal scale of measurement? Usually in IRT models, the latent variable is on an interval scale. But notice that this is a property given by the model. It doesn’t mean that in reality our latent variable is on that scale.
* Is <insert name of student here> proficient? This is again a yes/no question. Thissen quotes Thorndike’s line: “Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality” and calls for changing the way we report scores. He argues we should adopt a probabilistic way to report them and take advantage of that opportunity to educate test consumers.
* Achievement levels as black, white, with (perhaps) two shades of gray. Thissen expresses his disagreement with creating achievement levels. He argues the arbitrariness of the process (the standard setting procedure, the judges, the wording of the questions, etc.) as a factor to not reporting achievement levels.
* Is <insert name of student here> an “effective teacher”? The author is clearly nor in favor of statistical models that categorize teachers as effective/non-effective. Teacher effectiveness is multifaceted, and many other factors (such as non-school factors, and previous student experience) should be considered to evaluate teacher effectiveness if at all.

van der Linden, W.J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement,* *46*(3), 247-232.

*Background*

Thurstone (1937) first addressed the relation between responses and response times (RTs) from a perspective now known as item response theory with the intention of analyzing the notions of ability and speed. In his graphical example of a response surface, the main features are (a) a decrease of the probability of success with the difficulty of the item but (b) an increase of the probability with time. However, several important conceptual questions about response time (RT) remain unanswered. For instance: Are RTs independent of the features of the items? Does local independence also hold for the responses and RTs?

*Purpose*

The article aims to resolve several conceptual issues in response time by analyzing a selection of RT models that are typical of the different conceptual choices that have been made in the literature. Specifically, three main categories of RT models, distinct models for responses and RTs, response models that incorporate RTs and response models that incorporate RTs were discussed.

*Clarifications on conceptual issues in response time modeling*

1. RTs on test items should be treated as realizations of random variables.
2. For any type of test, RTs, item completions, and responses should be treated as realizations of distinct random variables.
3. Time and speed are different concepts but are related. RT models with speed as a person parameter should also have an item parameter for their time intensity.
4. Speed and ability are related through a distinct function , which requires fixed parameters for the effective speed and ability of the test takers.
5. RT models require item parameters for their time intensity but difficulty parameters belong in response models.
6. It is reasonable to assume conditional independence between RTs and responses on the same items as well as between RTs on different items.

*Recommendations for incorporating response time in measurement models*

Three different assumptions of local independence relevant to the modeling of the response time were made. However, in real-world administration, the response time and person ability will fluctuate during the test, and minor violations of the assumptions will be observed. When the violations are random and minor, there is no problem. But when larger and more systematic violation is observed, it should be treated as a signal of the result of inadequate instructions to the test, fatigue due to the test length, and the impact of the time limit.

Practical testing problem can be solved by incorporating RT in the measurement models. For instance, to assess the degree of speediness of a new test form, RT models with an explicit parameter for the speed of the test takers will provides us rich information on that aspect. Specifically, it is possible to fit a RT model and check the residual RTs for any changes in speed during the test.

RT models can be implemented in computerized adaptive test where more precise and constrained item selection algorithm can be devised to achieve higher test administration precision and estimate accuracy.

Additionally, RTs can also be used in computer based test to detect aberrant behavior of test takers, for example, due to preknowledge of some of the items, pure guessing during the test or attempts to memorize item during the tests.

Wise, S.L., & Demars, C.E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement,* *43*(1), 19-38.

*Background*

Wise and Kong (2005) noted that there are at least three situations in which examinee effort should be a concern to measurement professionals. First, there are a number of assessment programs that have serious potential consequences for institutions but little personal consequences for individual examinees. Second, high-stake testing programs sometimes administer test items in low-stakes settings, particularly in the early stages of the program. Third, a substantial amount of measurement research has been conducted in low-stakes settings at colleges and universities where the subject pools mainly include volunteer students.

*Purpose*

The general goal of the current study was to develop an IRT model that incorporates examinee effort and to evaluate its usefulness when rapid-guessing behavior is present.

*Incorporating effort into the measurement process*

1. Schnipke and Scrams (1997) identified two types of examinee behaviors: solution behavior and rapid-guessing behavior.
2. Wise and Kong (2005) hypothesized that rapid-guessing behaviors on low-stakes tests reflected a lack of examinee effort and developed an index, termed response time effort, for measuring an examinee’s overall test-taking effort.
3. In each encounter, the examinee makes a choice to engage in either solution or rapid-guessing behavior. The choice is reflected by the time the examinee takes to respond to the item.
4. The adequacy of the chosen time thresholds can be evaluated by (a) comparing it with a combination of the amount of reading and cognitive processing required to identify the correct answer under solution behavior and (b) comparing the response accuracy on thresholds with the expected accuracy of rapid guess which equals to the response accuracy under random responding.
5. The effort-moderated model specifies two-item response functions – one for solution behaviors and one for rapid-guessing behaviors. Depending on the time used by examinee *j* to answer item *i*, one or the other of the response functions would be used to model the response.

*Recommendations of effort-moderated IRT model in measurement practice*

In real-world testing situations, whenever test performance has little or no consequences for examinees, some examinees will devote little effort toward the test. This is a direct threat to test score validity. The introduction of effort-moderated modeling captures low examinee effort by the parameterization of rapid-guessing behaviors, which will potentially help measurement professionals to manage the problem.

Relative to the standard 3PL model, the effort-moderated model (a) fits the response patterns for far more examinees, (b) yielded substantially different item parameter estimate, and (c) generated proficiency estimates with higher convergent validity. The presence of rapid-guessing behavior also explains the differences observed in the item difficulty parameters between the standard IRT model and effort-moderated model. The accuracy of rapid guesses tends to be far lower than responses based on solution behavior, which implies that a standard 2PL model will yield item difficulty parameters that are too high.

According to the real data and simulated data analyses result, the effort-moderated model would be more appropriate to use than standard IRT models when a modest amount of rapid-guessing behaviors is present (e.g., 2% or more). This also represents one of the prerequisites of the effort-moderated model which is the use of computer based tests where the response time can be automatically recorded.

Schnipke, D.L., & Scrams, D.J. (1997). Modeling item response time with a two-state mixture model: A new method of measuring speedness, *Journal of Educational Measurement, 34*(3), 213-232.

*Background*

Yamamoto (1995) was concerned with the guessing component of speededness. His primary goal was to develop a method of estimating item parameters that eliminates the effects of random responding. He assumed that examinees begin a test with solution behavior, the active attempt to determine the solution to every item. This strategy might switch to a random-response strategy as time elapses, and this behavior is believed to result in a lower response accuracy.

*Purpose*

The aim of the article is similar to Yamamoto’s, but the focus is on response time, rather than accuracy, to categorize examinee behavior. The author emphasize that rapid random responses are made rapidly, consistent with the assumption that guessing behavior is due to lack of time.

*Hypotheses and result*

1. Is rapid-guessing behavior the same across time? It appears that the rapid-guessing response-time distribution is essentially the same across items.
2. Does rapid-guessing behavior increase toward the end of the test? The proportion of rapid guessing obtained from the fit of the standard mixture and common-guessing mixture models tend to increase across item position.
3. Is accuracy an increasing function of response time? Accuracy starts near chance (or below) for fast response times, then increases as more time is spent processing an item, up to a plateau that varies by item. It can be concluded that accuracy does increase as a function of response time up to a point, after which accuracy is fairly constant.

*Recommendations on the use of the model in practical measurement*

The two-state mixture model achieved the best performance in nonadaptive computer-administered test design, where the techniques takes advantage of the fact that all examinees received the same items in the same order. Additionally, the two-state mixture model could also be used when blocks of items are administered together.

Even though the two-state mixture model might not be implemented for the standard CAT design since individual items are selected and administered to examinees and the items that are affected by rapid-guessing will appear in various positions. However, this model may play an additional role in CAT. Because items will be used over time in CAT design, some items will become known to at least some of the examinees, and a fast correct response may potentially be produced by the examinees. Two-state mixture modeling may be helpful in detecting such behaviors.

The mixture model result serves as an indicator for anomalous item performance. In the example, the mixture model could not be estimated because of nonconvergence. This suggests a low proportion of rapid-guessing behavior. This item was the second most difficult item on the first half of the test which suggests that something other than guessing behavior is affecting the responses. Although the estimation results for this item are disturbing, they do demonstrate the importance of checking model fits with multiple methods when using this technique.