**Operational Measurement: Quality Assurance 2018**

**Annotated Bibliography**

**Quality Control**

* Leckie, G., & Baird, J. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience.
* Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment.
* Schmidgall, J. E. (2017). *The consistency of TOEIC speaking scores across ratings and tasks.*
* Cauffman, E., & MacIntosh, R. (2006). A Rasch differential item functioning analysis of the Massachusetts youth screening instrument: Identifying race and gender differential item functioning among juvenile offenders.
* Coffman, D.L., & BeLue, R. (2009). Disparities in sense of community: True racial differences or differential item functioning?
* Freedle, R.O. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores.
* Kristjansson, E.A., Desrochers, A., & Zumbo, B. (2003). Translating and adapting measurement instruments for cross-linguistic and cross-cultural research: A guide for practitioners.
* Malau-Aduli, B.S., & Zimitat, C. (2012). Peer review improves the quality of MCQ examinations.
* Van der Merwe, H. (2015). Quality assuring multiple-choice question assessment in higher education.
* Rios, J.A., Liu, O.L., & Bridgeman, B. (2014). Identifying low‐effort examinees on student learning outcomes assessment: A comparison of two approaches.
* Wise, S.L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests.
* Rios, J.A., Guo, H., Mao, L., & Liu, O.L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not?

**Test Security**

* Ferrara, S. (2017). A framework for policies and practices to improve test security programs: prevention, detection, investigation, and resolution (PDIR).
* Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations.
* Sinharay, S. (2017). Which statistic should be used to detect item preknowledge when the set of compromised items is known?
* Chajewski, M., Kim, Y., Antal, J., Sweeney, K. (2014). Macro level systems of statistical evidence indicative of cheating.
* Skorupski, W.P., & Egan, K. (2014). A Bayesian hierarchical linear modeling approach for detecting cheating and aberrance.
* Skorupski, W.P., & Wainer, H. (2016). The case for Bayesian methods when investigating test fraud.
* Fremer, J.J., & Ferrara, S. (2013). Security in large scale, paper, and pencil testing.
* Woodruff, L. (2013). Security issues in classroom testing.
* Lane, S. (2013). Security issues in writing assessment.
* Bush, M.E. (2006). Quality assurance of multiple-choice tests.

**Differential Item Functioning**

* Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests.
* Bjorner, J. B., Kreiner, S., Ware, J. E., Damsgaard, M. T., & Bech, P. (1998). Differential item functioning in the Danish translation of the SF-36.

Leckie, G., & Baird, J. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399-418.

The authors investigated three aspects of rater effects – severity, central tendency and rater experience on essay scoring with data from England’s 2008 national curriculum English writing test for 14-year-olds. Severity drift was examined to see whether raters changed their practice of scoring over time. Central tendency is the propensity to award a restricted range of scores around the mean (or mode or median) and to avoid awarding extreme scores (Saal, Downey, & Lahey, 1980). Rater experience was categorized based on their prior rating experience into three levels – team leaders, experienced raters, and new raters. The research purposes of this paper were (a) to compare the differences in rater severity drift by prior rating experience, and (b) to compute the effects of average central tendency for each of the three groups of raters.

The essay was scored on a 30-point scale with three subscales of (a) sentence structure and punctuation, (b) text structure and organization, (c) and composition and effect. New raters were recruited into the study to participate in face-to-face training meetings with a team leader. The expert committee consisted of senior raters with at least one-year experience of rating and experienced test developers. In total, there were 689 raters: 135 team leaders, 372 experienced raters, and 182 new raters.

The data were conceptualized as hierarchical data, because scores can be viewed as nested within raters, but they also can be viewed as nested within essays. To analyze clustered data, one multilevel model was fitted to measure raters’ mean levels of severity and to measure whether there was a central tendency to their scoring. The other multilevel model was built to establish whether raters’ levels of severity drifted over time. In each model, they examined whether these rater effects differed across the three groups of raters.

In terms of raters’ individual severity trends, the result indicated that raters became more homogenous the more essays they scored. However, each rater group did not, on average, become significantly more or less severe over time. In regard to rater experience and rating severity trend, the experienced and new raters, on average, did not score significantly differently from the consensus scores assigned by the expert committee. However, the team leaders, on average, significantly overscored by half a point higher than the consensus scores. The researchers also suggested that all three rater groups were unstable in terms of rater severity, and the new raters were the most unstable in scoring over time. There was a central tendency to raters’ scoring – raters, on average, overscored low quality essays and underscored high quality essays. This paper ended with a call to investigate the effects of severity on the more experienced raters.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493.

The authors explored the rater bias patterns of native English-speaker raters when they rated EFL essays written by Japanese university students on a topic. The term *bias* refers to rater severity or leniency in scoring. When raters interact with the items in a performance test for a period of time, there might be systematic subpatterns – severity or leniency trends that emerge from the raters’ practices of scoring (Wigglesworth, 1993, p. 309). In this study, the essays were scored against the rating scale containing six categories with equal weights for (a) content, (b) organization, (c) style and quality of expression, (d) language use, (e) mechanics, and (f) fluency. Forty native-speaker raters were recruited to score the essays. Each rater scored 40 essays that were selected from an essay pool.

The research questions included: (a) How reliably does the rating scale function with this particular group of raters? (b) In what ways do the raters differ from each other in their assessments of category characteristics? (c) In what ways do the raters differ from each other? (d) What is the effect of using a large number of raters in discovering systematic bias in rater interactions with categories and writers? (e) Can systematic bias patterns among raters be discovered, and if so, what is their nature?

The researchers used multi-faceted Rasch measurement (MFRM), in which the researcher adds the facet of judge severity (or another facet of interest) to person ability and item difficulty and place them on the same logit scale for comparison.

To answer the first RQ, the MFRM analysisreported that the rating scale functioned reliably with this group of relatively inexperienced raters. After training, most of the raters were able to use the rating scale to assess the writing ability of EFL writers to a satisfactory standard – 27 raters achieved acceptable fitting standards, whereas three did not. For RQ2, the analysis revealed the bias for rater by category. Twenty-four out of the 40 raters had significant bias interactions with categories, and there were 57 significant bias terms in all. The bias interactions were relatively balanced for negative (showing leniency), or positive (showing severity). Some patterns in rater-category interactions emerged. If the raters scored more severely for Content and/or Organization, they tend to be more leniently for Language Use and/or Mechanics, and vice versa. To answer the RQ3 and RQ4, the analyses evidenced bias for rater by writer – 329 significant bias interactions between raters and writers. Further analyses showed that it appears that raters are more likely to have severe or lenient bias towards higher ability writers than lower ability writers, and more lenient bias towards the lowest ability writers.

This paper ended with the discussion of the potential use of MFRM to make high-stake performance tests more fair and equitable. Some rater training implications were also discussed.

Schmidgall, J. E. (2017). *The consistency of TOEIC speaking scores across ratings and tasks* (ETS Research Report No. RR-17-46). Princeton, NJ: Educational Testing Service.

In this test report, the author provided evidence to support the consistency of the test scores across tasks and raters in the speaking section of TOEIC, an English proficiency test. The Speaking section includes 11 tasks. The first three tasks were Claim 1 tasks, in which two tasks were designed to ask the test taker to read a text aloud and the third to describe a picture. The Claim 2 tasks included from task 4 to task 9 that elicit responses from test takers to answer questions. Claim 3 tasks included task 10 and 11 that require longer and more comprehensive responses. Each task was judged by a rater and a scale score was assigned against rating rubrics. Tasks 3 to 11 were scored by one rater with one score, but tasks 1 and 2 were given two scores by one rater on pronunciation and intonation, respectively. The author included and rescored 1,390 formerly administered and scored TOEIC Speaking forms. The research questions were about how consistent ratings were at three levels – individual tasks, claim-level performance, and overall scale score, as measured by generalizability and dependability coefficients.

The author employed G-theory as the framework to identify variances associated with test-taker ability (*p*) and two facets of the measurement errors, ratings (r’) and tasks (*t*). The author estimated generalizability coefficients and variance components associated with facets of measurement for scale scores, claim scores, and individual task scores. G-studies were performed to provide estimates that reflect the actual measurement design of a data set and D-studies were estimated for different variations of the original measurement design.

The author provided consistency evidence at three levels, individual task scores, Claim-level performance, and scale scores. Analyses across all three levels of the scoring procedure suggested that differences between ratings had a minimal effect on scores. The analysis of scale scores found that minimal variance was associated with differences between scale scores and different set of ratings (r’) $\hat{ρ}$2 = .89. However, for individual task scores, the generalizability coefficients were moderately high for tasks from 4 to 11 ($\hat{ρ}$2 range is .72 to .92), but relatively low for tasks 1 to 3 ($\hat{ρ}$2 range is .57 to .69). The analysis suggests that ratings across the tasks 1 and 2 that assess test takers’ pronunciation and intonation were not as highly consistent as other tasks that elicit interactive, meaningful responses.

Cauffman, E., & MacIntosh, R. (2006). A Rasch differential item functioning analysis of the Massachusetts youth screening instrument: Identifying race and gender differential item functioning among juvenile offenders. *Educational and Psychological Measurement, 66*, 502-521.

The authors evaluated the Massachusetts Youth Screening Instrument – second version (MAYSI-2) as one assessment tool to screen and assess mental health symptoms among youths in the juvenile justice system. The MAYSI-2 consists of 52 items, categorized in 7 domains: alcohol/drug use, angry-irritable, depressed-anxious, somatic complaints, suicide ideation, thought disturbance, and traumatic experience. However, given the disproportionate number of incarcerated youths from minority backgrounds in the justice system, and how norming samples for most mental health assessments are typically representative of non-Latino White youth populations, there is a need to ensure that the instrument refer youth from various backgrounds for services and treatment without bias. Thus, there is a need to ensure that the MAYSI-2 (and other assessment tools) does not exhibit differential item functioning (DIF), in this context, between youths from different racial and gender backgrounds, as the presence of racial and gender DIF can have profound effects on access and referral to services for some incarcerated youth over others.

The authors address racial and gender DIF in a few ways. The authors applied the Rasch model, as it considers the interaction between individual characteristics and item characteristics from the MAYSI-2. To ensure that the item conforms to the Rasch model, they evaluated the item fit statistics, mainly to evaluate item misfit. Within this scope, they examined the infit and outfit of the items. Infit (information weighted statistic) is sensitive to the misfit of responses from persons whose estimated level on the latent variable is near the location of the item on the latent continuum. Outfit, which is unweighted, is more sensitive to misfit from persons located further away from the item on the latent variable. The authors then, by means of the Rasch model, tested uniform DIF. DIF effects were computed in Winsteps by calculating the difference between item location parameters for the multiple groups. The authors also utilized Winsteps to examine nuisance dimensions, or the potential confounds that may affect measurement and interpretability of the assessment tool.

The authors found that all of the subscales, except for traumatic experience, showed misfit, with *t*-values exceeding +2.0 and -2.0. Additionally, there was evidence that a number of the 52 items exhibited statistically significant levels of racial or gender DIF, within all subscales. Thus, the authors recommend that special care must be taken when making comparisons using the MAYSI-2, as the instrument and subsequent subscales may exhibit different properties depending on respondent’s demographic characteristics. They do recommend eliminating items with large DIF effects, but also caution dropping items that exhibit misfit, as their analyses are limited and lack a baseline to judge the magnitude of the misfit reported.

Finally, the implication for quality control is that, in conducting DIF and item fit analyses for the MAYSI-2 instrument, the authors make an effort to ensure that high quality standards are maintained through validating that the instrument is capturing what it is designed to capture: the mental health behavior and needs of *all* incarcerated youths. DIF effects and item misfit only adds to the extant measurement error. Therefore, performing rigorous analyses and examining the dimensionality of the instrument in this manner can serve to minimize measurement error and potentially improve instrument measurement and score reliability.

Coffman, D.L., & BeLue, R. (2009). Disparities in sense of community: True racial differences or differential item functioning? *Journal of Community Psychology, 37*, 547-558.

The authors evaluated the Sense of Community Index (SCI) as a tool to measure the psychological sense of community (SOC). Specifically, the SCI is used to examine psychological perceptions related to community belonging. However, because different ethnic groups have different cultural and historical experiences that may influence their perceptions and interpretations of the scale items, the authors suspect that differential item functioning (DIF) may exist with how different groups, in this context, White and Black groups, respond to the SCI items. As DIF effects can convolute true between-group differences, and subsequently, result interpretations, the authors investigated the psychometric properties of the SCI to confirm whether the same construct (SOC) is being measured consistently across groups. In addition to examining DIF effects, the authors also evaluated the dimensionality of the SCI, as it has been previously represented as either a single factor or a four-factor instrument.

The authors evaluated the SCI on a diverse sample through computer assisted telephone interviews. They used item response theory (IRT) to address dimensionality and DIF effects, specifically full information item factor analysis with marginal maximum likelihood estimations. The IRT models were fitted using the Akaike weight information criterion (AIC), the Bayesian information criterion (BIC), and the log-likelihood nested tests. The two-parameter logistic (2PL) model was chosen based on these criteria. Lastly, chi-square tests were used to assess racial differences in demographic variables.

For dimensionality, based on a scree plot of the eigenvalues, the authors found that only one factor was retained. Thus, the SCI appears to be a single factor instrument that measures the SOC construct. In examining the information characteristic curve (ICC), the authors found that items in the SCI provided more information, and subsequently more precision, in the below average range for self-reported SOC. Regarding DIF effects, none of the SCI items exhibited statistically significant DIF. Item 5, however, did appear to be a more reliable measure of SOC for White participants than Black participants.

The authors recommend adding more items to the SCI that provide information in the middle and above average range for the SOC construct. Doing so will allow for more precise measurement of the construct across the entire range of scores. Because none of the SCI items exhibited statistically significant DIF effects, the authors have no further recommendations to address DIF.

Evaluating DIF effects is important for quality control. DIF is the statistical analysis that flags items for review. Maintaining high quality standards, thereby, necessitates that researchers examine their scales to ensure that the instruments are accurately measuring what they were designed to measure, particularly when analyses show evidence of DIF effects. If items in a scale are bias, it reduces precision and interpretation capabilities for the overall scale/instrument, which subsequently impacts the stakeholders involved, including the researchers’ ability to make accurate inferences, the participants’ ability to report their beliefs and attitudes, and the policymaker’s ability to act on the researchers’ inferences. Thus, in this context, the authors are ensuring that conclusions drawn from the SCI about individuals’ self-reported SOC are as accurate as possible, across groups, through examining potential DIF effects.

Freedle, R.O. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. *Harvard Education Review, 73*, 1-43.

The author critiqued the SAT, suggesting that the version of the SAT that was administered at the time was both culturally and statistically biased against minority students. Cultural bias occurs when different ethnic groups consistently perform at different rates, for example, as a result of differential interpretation of test items. Typically, this is in reference to Black students performing consistently at lower levels than White students. Statistical bias occurs when two individuals matched on ability levels perform differently on some criterion, for instance, on the subparts of a test. Thus, the author claimed that the SAT, being culturally and statistically biased, exhibited differential item functioning for White students compared to Black students. The author found specifically, by means of previous research, that on the verbal SAT test, minority students on average were getting more hard items correct, whereas White students on average were getting more easy items correct. This is posited to be due to hard items being less ambiguous in linguistic usage than easy items, and thereby absent of cultural familiarity. The author argued that because easy and hard items are scored the same, that the advantages that White students have compared to Black students, relative to the total score, are further exacerbated.

To mitigate this, the author suggested a new method for re-scoring the SAT, called the Revised-SAT (R-SAT). R-SAT scores are calculated from only the hard questions on the original SAT test (e.g., 40 items). The author suggested multiple ways to calculate a R-SAT score, including group-level or individual-level comparisons. However, the most optimal method is a combination of these other methods: the final method for estimating R-SAT. This includes first scoring only the hard half of the original SAT test for every individual. Then, using the individual scores, define a single distribution of scores. The raw scores are then corrected for guessing. Finally, a new scaled score from 200-800 is defined and represents the R-SAT score.

The author found that in using the new R-SAT scoring method instead of the original SAT score, Black students benefited proportionately more than White students. That is, the sample of students that had a lower original SAT score now have a higher R-SAT score. The implications that result from the R-SAT score is that now more minority students will be eligible for their selected colleges, with also an increased likelihood of acceptance. Additionally, the measured difference between White and Black SAT examinees is shown to be substantially reduced when using the R-SAT score. However, R-SAT’s predictive validity is nevertheless limited, as the R-SAT has not been utilized prior to this study, and therefore lacks a comparable baseline.

The author’s recommendation is that using the R-SAT scoring instead of the original SAT scoring has the potential to reduce ethnic bias, and therefore also has the potential to increase dramatically the number of minority students who qualify for admission at colleges and universities. Additionally, the author suggests that a larger number of hard items need to be added to the verbal SAT, particularly hard antonym and analogy items. Doing so will serve to increase the reliability and validity of the SAT, and subsequently the R-SAT.

The implications for quality control are that, although items with large DIF effects may be removed prior to test distribution, there still may be items that elicit small DIF effects, due to respondents’ characteristics. Small DIF effects accumulate, and may have a dramatic effect on a respondent’s estimated true ability. Thus, to ensure quality control, it is important to recognize the pervasiveness of cultural and linguistic differences in order to reduce cultural and statistical bias, and to maintain high test quality standards. In this context, it is important to acknowledge and address consistent DIF effects on the SAT, even if the effect sizes are small.

Ferrara, S. (2017). A framework for policies and practices to improve test security programs: prevention, detection, investigation, and resolution (PDIR). *Educational Measurement: Issues & Practice*, *36*(3), 5-23.

Ferrara proposes a comprehensive framework for test security, prevention, detection, investigation, and resolution. He asserts that test security is not an end in and of itself; instead it is needed to be able to make valid interpretations based on test scores. Taking a system-based approach to test security is required because once a test is released or administered, the security and therefore the validity of the inferences that can be drawn from the test can be compromised by the actions of test managers, administrators and other local staff.

Ferrara advocates using the following framework to look at test: Prevention, Detection, Investigation, Resolution (PDIR). Within these categories, he discusses things that can be done before, during, and after administration. Ferrara argues prevention is one of the most effective countermeasures to cheating and test security threats, but he argues to consider prevention on a wide scale. This includes protocols such as training, chain of custody, observation and even announcements of policy and consequences related to cheating.

In the area of detection, the literature focuses on statistical methods, but indicates reports from testing sites are also very important. Some detection methods also offer some prevention benefits; knowledge that observation occurs tends to limit attempts at cheating. Statistical methods have notable limitations. They may not be appropriate for small sample sizes and information about them tend to be scattered across publications, not distilled into a usable framework or system. Additionally, it is generally viewed as inappropriate to use statistical methods as the only piece of evidence that cheating or dishonesty occurred. Although, this area is developing quick, it is still in its infancy. Approximately only half of state educational testing programs use statistical detection methods.

Ferrara’s key points related to investigation and resolution center around the limited work in this area and the challenges in both conducting investigations and resolving test security issues for school administrators and local staff. He discusses lack of training, and conflicts of interest as challenges for school administrators in addressing these challenges. Additionally, these areas receive less research than the other two in his framework.

To address the challenges in maintaining test score validity through test security, he recommends using his proposed framework and expanding collaboration. Many school districts, states and test administrators do not have adequate resources to address all of these issues. Therefore, further collaboration and resource sharing is needed, such as creating a committee for PDIR for state educational assessments. The U.S. Department of Education has taken steps in this direction by adding test security and systems for protecting data integrity and privacy to their peer review of state assessments criteria. Finally, he advises that the measurement community can have an impact by working to address under-researched threats to test security and continuing to develop statistical methods and collaborating evidence to detect test security threats while minimizing the risk and impact of false positives. Overall, Ferrara seeks to summarize current research in the field, advocate for a unified framework, seek further collaboration and guide future study.

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues & Practice, 35*(1), 38-47.

In “Using Response Time to Detect Item Pre-knowledge in Computer-Based Licensure Examinations,” Qian et al. take one method for using item response times to detect prior item knowledge and apply it to two professional licensure exams. Their primary research questions were: (a) to what extent do examinees have item pre-knowledge and (b) to what extent are there potentially compromised items in each exams item pool.

The exams both had early dates (baseline, because items had no possibility of exposure) and later test dates. One exam used was non-adaptive and the other was adaptive. Given the stakes of the exams (professional licensure), examinees would have motivation to cheat, seek out, or share item information. The researchers used a method proposed by Van der Linden (2006) to model the response times. This used parameters of item intensity and discriminating power for each individual. This model had several assumptions including lognormal distribution of response times, approximately normally distributed residuals and items fitting the Rasch model. There were no major violations of assumptions. This model allows specific item response times for specific examinees to be compared to what is expected or predicted from them.

The model was then used to detect specific items that may have been exposed and specific examinees that may have item pre-knowledge of multiple items. The first testing period was used to control for type one error. In the non-adaptive exam, they were able to flag two items (of 111) as potentially exposed and two individuals (of 1172) as showing evidence of item pre-knowledge. On the adaptive exam, they were not able to identify any items that they believe were exposed nor were able to identify any suspect of preknowledge; however, they did identify an interesting response pattern that was flagged in 4 of 4675 examinees. Originally, these looked like preknowledge, but when comparing response time data with fit data and the position of the specific items, it became clear that these aberrant patterns were due to motivational changes. The different findings between the two exams are not too surprising given the differences in how long the tests were out prior to the second administration (nearly a year for the non-adaptive and a couple of months for the adaptive) and the number of questions in each question pool. The adaptive test had a much larger item pool, reducing the chances that any two examinees were displayed the same questions. Additionally, the large question pool and adaptive nature of the test reduced examinees’ motivation to discuss questions with each other.

Overall, it appears that using response time data with the model provided by Van der Linden (2006) can provide useful additional information particularly when paired with traditional fit data and item position data. An additional, notable limitation is that examinees with preknowledge could map response times if they are aware that response times are being monitored. Response time does appear to be a useful way to identify and eliminate items that may have been exposed. The authors provide some evidence of the usefulness of response times to identify preknowledge but also highlight some of the concerns and constraints with its use.

Sinharay, S. (2017). Which statistic should be used to detect item preknowledge when the set of compromised items is known? *Applied Psychological Measurement, 41*(6), 403-421.

The author only addresses the situation in which test managers or test producers know that an item or given set of items is compromised. This tends to occur after the testing has occurred and information comes to light indicating possible items being compromised. The statistics discussed in this article relate to identifying examinees who may have benefited from these compromised items.

The author addresses the performance of two existing suggested statistics for this purpose: the posterior shift statistic (PSS; Belov, 2016) and a statistic based on the likelihood ratio test (LRT; Sinharay, 2017). Belov’s PSS was compared with seven other statistics for the same purpose and showed the best performance. Sinharay’s LRT was not included in this analysis so this current article undertakes comparing the PSS and LRT in both simulation and actual data sets. Detailed descriptions are available in this article as well as Belov (2016) and Sinharay (2017).

Limitations of PSS include not having a known distribution under the null hypothesis, increasing the difficulty of identifying critical values and the maximum value of PSS depends on the choice of ability levels. For LRT, the null hypothesis should be a normal distribution given large numbers of compromised and uncompromised items. This situation is generally true in real world situations, but LRT showed reasonably normal distributions for compromised items numbers of four or greater.

The simulation study was based on 100 questions for non-adaptive and 50 for adaptive using item parameters from several disclosed items from the Law School Admission Test (LSAT). The simulation results indicate that LRT shows similar power to PSS when looking at a few compromised items. The study continued on to compare the two statistics, using real data as well. For this study, they used item response data from two forms of a non-adaptive licensure test including 170 items 87 of which were shared between them. The sample included approximately 1600 examinees for each version. Both statistics showed some success in identifying possible aberrant individuals in the real data set. Within the real data set, the two statistics showed similar performance. Given the similar performance, the authors argue that practitioners may find LRT more attractive for use. This is due to the fact that critical values can be calculated directly (without simulation) due to the known distribution.

In concluding the author reinforces that a statistical measure alone is not enough to confirm aberrant behavior or item pre-knowledge, but he does suggest that it may be possible to use multiple statistical measures together to provide convincing evidence. Further investigation into detecting item pre-knowledge would be beneficial, including comparing these methods with recently described response time methods. Note that the article includes R code needed to compute either of these statistics for future use or research.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment*, *14*(3-4), 160-179.

In her 2009 article, Martiniello explored linguistic complexity and nonlinguistic representation as sources of differential item functioning (DIF) for English language learners (ELLs) on mathematics assessments. The author examined a potential source inequity in current assessment practices – math assessments may not accurately measure the desired construct in ELLs, a group that in 2002 made up 10% of K-12 enrollment. On the 2009 National Assessment of Educational Progress, 43% of ELLs scored below proficient in math, compared to 16% of non-ELLs. Other authors previously found that longer items on math assessments were more difficult for ELLs than non-ELLs. Conversely, nonlinguistic representations were critical for ELLs when interpreting and communicating mathematically, suggesting a potential practice for offsetting linguistic difficulties.

An important feature in understanding mathematics is recognizing construct-relevant vocabulary in context. For this study, the author considered only natural language in the analysis of linguistic complexity. The sample of 68,839 fourth graders included 3,179 ELLs and 65,660 non-ELLs. A panel of experts reviewed grammatical and linguistic complexity in 29 multiple choice and 5 short answer questions on a state 4th-grade math assessment using a 1 to 5 rubric with benchmark examples. Think-alouds administered to ELLs and expert reviews were used to determine which items were primarily schematic (as opposed to text only or pictorial). Schematic representation referred only to those diagrams that visually or symbolically showed relationships between objects or variables. The results of the assessment were modeled using a unidimensional IRT model, with the DIF statistic being the difference in difficulty estimates for ELLs and non-ELLs.

The author found significant (*p* < .01) correlations between the DIF statistic and linguistic complexity (.575) and schematic representations (-.553), suggesting that linguistically complex items favored non-ELLs compared to other items, and items with schematic representations favored ELLs compared to other items. Linguistic complexity (*p* < .001) and its interaction with schematic representation (*p* < .001) accounted for 66.3% of the variation in DIF in an ordinary least squares regression model fitted to the data. A second DIF measure, the standardization method, confirmed the previous finding (*r* = .93, *p* < .001). The results suggested that use of schematic representations when designing mathematics assessments could mitigate some of the difficulties ELL students face due to linguistic complexity.

Martiniello observed that DIF offered evidence of bias, but proving that bias existed was outside the scope of her study. A thorough review of evidence and an exploration of counter-hypotheses would have to be performed to conclude bias. She suggested many other linguistic patterns as possible sources of DIF, including syntax, text layout, and references to mainstream culture. Lastly, the interaction between construct-relevant vocabulary and linguistic complexity is another area she believed should be explored further.

Kristjansson, E.A., Desrochers, A., & Zumbo, B. (2003). Translating and adapting measurement instruments for cross-linguistic and cross-cultural research: A guide for practitioners. *CJNR (Canadian Journal of Nursing Research)*, *35*(2), 127-142.

Measurement plays a critical role in medical diagnosis. Different instruments are used for different purposes - the flexibility of these instruments allows nurses to assess a wide range of conditions. Many steps are taken to ensure the reliability of the instruments and the validity of diagnoses. Bias is a large concern, referring to a specific population subgroup systematically performing differently due to a characteristic not relevant to the instrument. Kristjansson, Desrocher, and Bruno (2003) explored the many methodological pitfalls that could introduce bias when translating an instrument. The authors also described steps to mitigate bias in translations.

Problems:

* **Lack of Conceptual Equivalence.** Psychological concepts are not necessarily universal. Different psychological conditions manifest themselves differently across cultures.
* **Differences in Cultural Norms.** Societal norms shape thoughts and actions. Culture could explain reluctance to report pain or weakness, or tendency to provide socially desirable responses.
* **Lack of Semantic Equivalence.** Words and phrases are often how meaning is conveyed in psychometric instruments. The following are specific examples of challenges that arise in translation.
	+ **Problems of Lexical Mapping.** Words and idioms do not always translate perfectly, which can change the way items are interpreted. Additionally, cultural concepts such as units of measurement can change question benchmarks.
	+ **Problems of Grammatical or Syntactic Equivalence.** Sentence structures differ between languages, and proper translation may require changing syntax.
	+ **Experiential Equivalence.** Interpreting instruments often requires general knowledge, and general knowledge is not the same everywhere.

Steps to take:

1. **Verification of Focal Concept Equivalence.** Use ethnographic research to determine if the construct of interest is relevant to both cultures. If it is not, either abandon the research or find a similar construct that is relevant.
2. **Translation of the Instrument and Development of Preliminary Versions.** Multiple professional translators with experience in both languages, cultures, and the relevant construct should collaborate with the research team to do a double/back translation.
3. **Committee Review and Evaluation of Preliminary Version.** A committee of experts and translators should review the preliminary version and resolve any issues they find.
4. **Pre-testing the Instrument.** Pre-test the adapted version on a small representative sample of the targeted population. Focus groups and think-alouds can be helpful. Revise.
5. **Pilot Test.** Pilot test in multiple phases, first using a test-retest method with a small sample (*n* = 20) of bilingual people to determine concurrent validity and reliability. Next, do a much larger pilot (*n* > 100) for item analysis and a final revision.

Bjorner, J. B., Kreiner, S., Ware, J. E., Damsgaard, M. T., & Bech, P. (1998). Differential item functioning in the Danish translation of the SF-36. *Journal of Clinical Epidemiology*, *51*(11), 1189-1202.

Rigorous techniques to reduce bias in the translation phase of a health survey do exist but are insufficient for detecting DIF in the translated forms. Previous authors concluded that the Danish version of the SF-36 health survey has a comparable internal consistency and factor structure to the English version, and a Rasch analysis found the item difficulties to be comparable. Bjorner, Kreiner, Ware, Damsgaard and Bech (1998) sought to determine if differential item functioning (DIF) techniques could be useful tool in evaluating translations of health surveys.

The analyzed data were drawn from a 1994 Danish health survey (*n* = 3,950) and a 1990 U.S. health survey (*n* = 1,506). Because of the rigorous assumptions of an item-response approach, the authors opted to use a contingency table method to test for DIF. Magnitude of DIF was detected using a Mantel-Haenszel odds ratio modified for polytomous items. A secondary analysis confirmed DIF by using items that had not shown it as an anchor scale. The impact of DIF was calculated by comparing (a) scales with DIF items included, (b) scales with DIF items excluded, and (c) individual items. The authors controlled for ethnicity, age, and sex.

Primary and secondary analysis confirmed 12 of the 35 items showed signs of DIF. Five of those items had severe DIF, six had moderate DIF, and one had negligible DIF. Removing items with DIF did not change the conclusions drawn from the survey about overall mean differences between American and Danish populations. Both the Danish and American populations were relatively healthy, with skewed distributions containing many fewer people with poor health. Because of this, DIF may have been more pronounced in those populations since they had less of an influence on the overall scale. Additionally, comparing individual items cross-nationally showed serious bias.

The cause of DIF could not be explained via the results of the study and required a post hoc explanation. Fortunately, the authors could draw on translation quality control issues to explain many of the items with severe DIF. Several items had poor conceptual equivalence with no exact translation. Two of the items which related to distance changed the unit from mile to kilometer (0.62 miles) which resulted in Danes rating it as easier. For the other items, the authors hypothesized that despite a good translation, cultural differences could possibly account for the DIF.

Despite its shortcomings, the authors recommended keeping the Danish SF-36 in its current form. Removing items on short instruments is not feasible. Although revising items was a possibility, survey translation is a balance between conceptual equivalence and common language use. Making items more readable would result in reducing equivalence and the psychometric evidence suggested that the English and Danish instruments performed similarly overall. They suggested that fitting a model that corrected for DIF was a potential future direction for research. The authors concluded that DIF analysis is a powerful tool in a cross-national evaluators arsenal and should be used whenever surveys are translated.

Chajewski, M., Kim, Y., Antal, J., Sweeney, K. (2014). Macro level systems of statistical evidence indicative of cheating. In N. Kingston, & A. Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 101-120). New York, NY: Routledge.

The authors analyze the ways in which we can use statistical tests to examine high levels of irregular responses that are indicative of cheating. In the beginning however it goes over the context for where this cheating happens. Cheating, as is described, takes on three forms: (a) taking, giving, or receiving information from others; (b) using forbidden materials or information; and (c) circumventing the assessment process. Most large testing facilities have Test Security Offices and these attempt to prevent, detect, and investigate issues with test security. Testing and cheating has become much more advanced because of technological advances as well as globalization, so the detection needs to become advanced to keep up. There are sophisticated methods for detection but first you need to address prevention.

Although the main focus of this chapter is the detection stage, it makes note about how important the initial prevention is. We need to learn the techniques students use to cheat and ways to combat those. Randomized test form administration and within form section scrambling are often used to prevent collaborative cheating as one example. In addition to the students, we need to prevent fraud from the educators or administrators. Changes in policy have led to some of this fraud, especially from No Child Left Behind when it made teachers be judged based on students’ test scores. Teachers then, as reported from one who was interviewed, will “do what is necessary to avoid low scores”. This is especially problematic because these educators used to be the primary agents for prevention.

Detection of cheating can be done in the micro level with human coders and experts but this is time consuming and expensive. Routine screening administrations can identify student pairs, aggregate this into macro levels, and are low cost when using statistical methods. These regular screenings can build our knowledge about what an acceptable rejection rate would be and what irregular test responses would look like. Security screenings then should be relevant in evidence to a variety of sources and flexible enough to globally evaluate test administration. These methods then should not be separated from other evidence such as traditional cheating investigations and a macro system should incorporate both.

A study was performed to verify the usefulness of combination. Over 400,000 students between two groups were analyzed using a 100 item multiple choice exam. Two statistical indexes were used to identify levels of similarity between respondents, which would indicate cheating. One was based on classical test theory where matching items were conditioned on number incorrect. Another was based on IRT where students responses were analyzed based on predicted matches by chance. The groups were separated into schools where it was noted if that school was flagged for a higher than normal level of suspected cheating. This study found that 2% of the schools were flagged, which would be 160 schools. This is consistent with what would be expected, which gives us evidence that these tests are accurately detecting what was intended. In addition, both indexes flagged mostly similar schools which helps self-corroborate. This study also found issues with higher rates based on demographic data and issues with longitudinally. Any results found from testing such as this would then need to be brought back to the micro level and investigated in a case by case basis to fully fulfill the ideals of the macro system.

Skorupski, W.P., & Egan, K. (2014). A Bayesian hierarchical linear modeling approach for detecting cheating and aberrance. In N. Kingston, & A. Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 121-133) New York, NY: Routledge

These authors propose a new method for analyzing fraud in statewide assessments (SWA). There are many methods used to statistically detect evidence for cheating such as similar response string patterns, and analyzing person-fit data. Many of these however focus on the student level, yet this is inappropriate for SWAs because the students have no incentives to cheat. Educators however need to make adequate yearly progress as mandated by policies such as No Child Left Behind. This can be a strong incentive to commit fraud because the results from the SWAs can affect teacher, school and district accountability and can even affect pay. School and student level analysis has been done, But classroom and teacher level analysis has been limited. This is important because that is where the fraud is hypothesized to happen.

The authors presented a model previously to address the classroom level. This method would detect group level cheating using a Bayesian hierarchical linear model (HLM). It was performed with real data and modeled change in individual scores, nested within schools over three years. External evidence of cheating was available which corroborated with the findings from the model. This was encouraging to support this model; however, with real data you cannot examine the accuracy of the method itself so a simulated data set was used for this study.

The data for the current study was simulated and built into 60 groups over three administrations. The data were made to be similar to the previous real data, with representative sample sizes, group means, and longitudinal effects. These were constrained to normal distributions; however, the real data may not have been perfectly normal. From these data 5% of the groups (9 groups) were randomly selected to demonstrate “aberrance”. This 5% rate across the longitudinal data was constituent with what was seen in the real data. This same process was repeated 50 times and each had the Bayesian HLM analysis performed to ensure reliability.

The model that was used had individuals within groups within time point. This is done to analyze the effects at each level. The main effect for group can be analyzed as well as average rate of growth for the time effect. In addition an interaction effect for group by time accounts for any unique effects that cannot be explained by the main effects. The model would predict aberrance and flag any group that had a predicted score that was above a certain threshold.

The results from the analyses were very consistent with the true data, with the model predicting aberrance for groups that had been chosen. All nine groups that were selected to be aberrant had predicted values very close to or above the threshold, whereas the rest had values that were lower. Across the replications, incorrect flags appeared between 0% - 2% of the time. The strong detection power from this model gives great support for using this model in the future to detect possible cheating and fraud in more real world applications.

Skorupski, W.P., & Wainer, H. (2016). The case for Bayesian methods when investigating test fraud. In G.J. Cizek, & J.A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 346-357). New York, NY: Routledge.

These authors examine how detection for statistical testing is done specifically relating to the type I error rates that come up when using multiple approaches. Traditional approaches have an inflated error rates and incorrectly interpreted *p*-values. Commonly used *p*-values lead to confusing and misleading interpretations, specifically when predicting the probability that a person is a “cheater”. Considering Bayesian reasoning for cheating detection can help solve these issues. The authors argue that it is important that we are accurate in our decisions because the cost of false positives and false negatives is too great for high stakes situations such as these. A falsely accused student could be barred from testing and lose out on an important certification and a false negative could allow a student to not be caught and damage the validity of the test.

The authors explain the theory behind Bayesian methods with a similar example to Savage and Wainer (2008). It explains a scenario that there is statistic that identifies cheaters which is “99% accurate” meaning that the type I error rate is set at 1%. Although this would lead you to believe that this would identify 99% of the non-cheaters correctly (to be 99% accurate with its specificity), this is a misinterpretation. The actual accuracy would be lower because there is a much higher rate of non-cheaters than cheaters. This leads more non-cheaters to be identified as cheaters than vice versa. Bayes’ rule solves this issue by factoring in how likely a person is to be a cheater in general. This requires previous knowledge but it allows for much more confident results.

Bayes’ rule is then explained both as a concept and within this context. This is an equation that shows that the probability of θ given x (in this case that one is a cheater given that they were flagged for being a cheater) is a function of the probability of x given θ, multiplied by the marginal probability of θ divided by the marginal probability of x. For this to be done, you need a prior marginal probability for θ. Because of this, the equation is then rearranged to instead use the probability that one is not a cheater. This can be done because we know that the probability of being a cheater plus the probability of being a non-cheater equals one. Now we can use our previous “accuracy” as our marginal probability. Using this with the proportion of flagged “cheaters” as the marginal probability for x, we can use the equation to give a more confident probability for detecting cheaters.

Using this in a data set with known cheaters shows that our predictions are fairly accurate. The higher the power for each null hypothesis from the traditional hypothesis testing, the more probable the estimate could be. This gives evidence that this method fits in with the testing that has been done, and can fit into methods used going forward. This method of predicting the probabilities of cheating in a population will be useful with the more confident detection for high stakes assessments.

Fremer, J.J., & Ferrara, S. (2013). Security in large scale, paper, and pencil testing. In J.A. Wollack, & J.J. Fremer (Eds.), *Handbook of test security* (pp. 17–37). New York, NY: Routledge.

The authors mention that after the No Child Left Behind Act, the total amount of school cheating increased. Although they mention that just 1% or 2% of school staff were involved in inappropriate testing behaviors, they recognized that no one really knows what is the exact amount of cheating and even that 1% can represent thousands of school staff. The authors focus on large-scale paper and pencil testing, explain what the security violations look like, present a conceptual framework for test security, and give recommendations for test security managers about how to deter cheating.

Regarding how people violate test security, the authors mention some examples of cheating such as teachers failing to return secure test booklets of state assessments and then use them in the classroom, teachers preparing students for a comprehension test by reading students a story identical to a passage in the test, teachers dictating responses to constructed-response items, and students paying a “smart” student to take the test for them. It is important to deter cheating because it undermines the decisions based on those scores. Also, cheating behaviors can damage students’ ethical sensibility.

The authors propose a conceptual framework for test security. According to them, a rigorous test security system is necessary but not sufficient; the most important part depends on the professional ethical thinking and behavior of teachers and schools staff in general. Without any rules and guidelines, the risk of making a bad decision is high but even if the school has rules, if teachers do not have a culture of professional ethics they might not follow the rules. In paper and pencil tests, the most common ways to cheat include sharing information about test content before its administration and helping students during the test. The use of multiple-choice or true/false items can increase the probability of cheating in paper-tests; also providing accommodations to students with disabilities and English-language learners might generate opportunities to cheat in subtle ways.

In order to maintain the security in paper and pencil testing programs, the authors recommend to expect cheating, to anticipate it and to create strategies to minimize it; implement a culture that promotes security in testing program which should include rules, guidelines, and a security plan. Once schools have that, it is crucial to train administrators to follow guidelines. This training must focus on how to assist school staff in understanding rules and guidelines, on making them feel supported and comfortable, and on teaching them how to set a tone for a calm organization especially when they have to deal with large-scale tests in big classrooms. It is important to involve all participants in the discussion about security in testing programs (i.e. parents, teachers, and other school personnel). Finally, authors suggest that managers need to put into practice all the recommendations above, so cheating on tests can be prevented.

Woodruff, L. (2013). Security issues in classroom testing. In J.A. Wollack, & J.J. Fremer (Eds.), *Handbook of test security* (pp. 85–100). New York, NY: Routledge.

Cheating is a problem that occurs at all education levels. The motivation to cheat can increase due to parental expectations, peer pressure, students desire to succeed, and testing consequences. However, it is not only a concern for standardized testing but also for low-stakes activities. Given this context, it is crucial to prevent cheating and implement security policies. The purpose of this chapter is to review the scope of cheating, the common cheating techniques, and the recommendations to secure test administration, all focused in a classroom setting.

Regarding the scope of cheating in classroom testing, the author argues that in order to prevent and include test security in classroom testing, it is important that teachers understand that cheating is a problem that still affects schools. Some teachers have the old stereotype that cheaters are evil-doers; however, this no longer applies. Along with this, the act of cheating has changed over the past years; for most US youth, cheating is not viewed as an aberrant act. Therefore if teachers do not recognize cheating as a reality and students accept this dishonesty, cheating on tests will be a strategic option to maintain good grades.

Some common cheating techniques in classrooms are: *pre-knowledge,* having access to a copy of the test or some part of the test ahead of time. Electronic devices such as cameras in eyeglasses, lapel pins, or pens can be used to gain pre-knowledge of test content. *Copying*, by an individual or in collusion with others, is one of the most widespread strategies to cheat. Students might sit strategically in the classroom to copy from others. *Unauthorized aids* refer to access to information during the test, like information printed on a small piece of paper, a tissue or a stick of gum. Information might be posted on the walls, posters, or the board. *Communication codes* allow students to communicate using signals that only they understand, like fingers to refer to the question number. Finally, *Proxy testing* occurs when students change their exams, the “smart” student answers the test of the “less smart” student. This strategy is also called impersonation and it is more likely to happen in a college setting than in a K-12 school setting.

In order to prevent cheating using these techniques, the author recommends strategies regarding test design. For instance, using scrambled forms of the test is a good way to prevent cheating in multiple choice of true/false items. If it is not possible to create different forms, some instructors use the “phantom form” where tests look different but they are the same, so tests can have different titles or colored paper to highlight differences. Another strategy to reduce communication codes is to assign students to seats instead of allowing them to choose their own seat, this could be done randomly or as they enter the room. Designing tests with items such as fill-in-the-blank, short answers or essay questions is also an option.

Other strategies mentioned by the author are: Prior to the test, organize the desks in strictly aligned columns, extending the width of a seating arrangement, check IDs, and ask students to sign a commitment to not cheat or share information. During the test, students cannot have access to electronic devices, must hand in the test individually avoiding passing test copies down a row or walk around the room during the test. Finally, after the test, collect the answer sheets individually, check that no pages have been removed, and remove all other tests and copies from the classroom.

Lane, S. (2013). Security issues in writing assessment. In J.A. Wollack, & J.J. Fremer (Eds.), *Handbook of test security* (pp. 101–123). New York, NY: Routledge.

The author notes the widely recognized prevalence of academic dishonesty. Students perceive cheating as an acceptable practice, as they do not know what constitutes cheating or what is ethical and moral academic behavior. As a result, students continue cheating because it is a way to obtain good grades. The author address security issues to prevent cheating in large-scale writing assessment as well as classroom writing assessment. First, a discussion about cheating and plagiarism on writing assessment is presented. Second, the strategies to detect plagiarism in classrooms will be addressed. The last part is about strategies for preventing cheating in writing assessment.

The author explains that writing assessment could be stand-alone writing prompts where students produce a response to a prompt or a brief topic, or text-based writing prompts where students have to read a passage, analyze it and write the answer. Also, this type of assessment can be large-scale where students answer one or two small prompts or in classrooms where tests can be longer and include both stand-alone and text-based writing prompts. Since the chapter is focused on cheating in writing assessment, plagiarism is an important topic to discuss because this kind of cheating is unique to writing assessment, and given the access to Internet materials plagiarism is widespread.

Regarding some common cheating techniques in writing assessment the author mentions memorizing or borrowing essays from someone else and submitting them as if they were their own, purchasing a paper from an online paper mill, or copy parts of a paper without an appropriate citation. The author mentions that cheating in writing assessments using a computer can be divided into *plagiarism* or *collusion.* Plagiarism occurs when materials from the Internet are copied by students and used it in their assignments without proper citation. In contrast, collusion occurs when students get together as a group and share materials when they should be working independently.

About detecting plagiarism, the *substring matching method* can be used. This method compares two documents and if they have a proportion of matching strings above a prespecified threshold then plagiarism might have occurred. A second method is similar to the previous one but comparing similarities of keywords. The last method mentioned is *fingerprinting* where similarity is based on lexical sequences of the text. In addition, there are electronic plagiarism detection systems used such as Ferret system, Stanford Mechanism, RECAP, and one of the most popular system is Turnitin.

Lastly, about preventing cheating in writing assessment, the author recommends teachers to provide a comprehensive discussion about plagiarism with students, require students to submit earlier version as well as the final version of their assignments, avoid using the same prompts in assignments, expose prompts to a small number of students, enforce security regulations, address ethical nature of the problem, engage students in practices that improve writing and proper methods of citation, and encourage students to report cases of cheating.

Bush, M.E. (2006). Quality assurance of multiple-choice tests. *Quality Assurance in Education*, *14*(4), 398-404.

The premise of this article is that designing good multiple-choice test items is a difficult task, and that there are several quality assurance issues to attend to when developing a bank of test items and building test forms. The author splits the discussion into sections about pre-test quality assurance and post-test quality assurance.

For pre-test quality assurance, the author advocates for using as many non-overlapping questions as possible on each test. This allows for the greatest breadth of coverage of topics from the domain being tested, and can minimize the luck factor of students receiving (or not receiving) specific test content that they may or may not know very well. That is, a large number of non-overlapping items would level the playing field as much as possible for all students. Because of the large amount of questions needed to do this – particularly with multiple test forms – and the difficulty in creating non-overlapping items, the author advocates for using multiple versions of items within the item bank (e.g., positively and negatively worded items). This can reduce the amount of time needed to develop items, but clearly, overlapping items would not appear on the same test forms.

The author also describes alternate formats of multiple-choice items that could be used to help avoid scoring problems due to guessing (or educated guessing). They are described as effective, but potentially confusing for test-takers who may try to think too much about test-taking strategies. Two options described are:

* **Negative Marking:** for a multiple-choice item with *n* options, awarding *n* – 1 points for a correct answer, and deducting 1 point for each incorrect answer. Over a large number of items, a student would stand to lose as many points as they would gain from guessing.
* **Liberal tests:** an extension of Negative Marking, which allows students to choose multiple answers, in the hopes that one of them is correct.

For post-test quality assurance, the author suggests reviewing item statistics; by reviewing the items’ difficulty and discrimination indices, items can be evaluated to see if they are too easy or too difficult, if any distractors are being chosen a disproportionate percentage of the time, and if the items are positively contributing to students’ scores on the test.

Alternatively, the author encourages soliciting feedback from the examinees about the test items. That is, by obtaining verbal or written comments from students who took the test, the test developers can potentially learn important information about how the students interacted with and understood the questions they just answered. This can be used to inform future revisions to these items, as well as future development of new items.

Malau-Aduli, B.S., & Zimitat, C. (2012). Peer review improves the quality of MCQ examinations. *Assessment & Evaluation in Higher Education*, *37*(8), 919-931.

The authors examine the effects of implementing an item-review process for multiple-choice questions administered over three years, for an assessment given to medical school students in Australia. This was done by examining item statistics from 2008 (prior to implementing the item review process) and comparing them to item statistics from 2009 and 2010 (after the implementation of the item review process).

The item-review process included peer-reviews of the items by colleagues of the medical school staff and assessment committees. These reviews consisted of two sessions each – in the first session, the reviewers discussed the structure of the test items, wording, and formatting. In the second session, the reviewers focused on the content of the items, particularly to make sure the items tested what was intended, the items were relevant, and the items tested core knowledge.

The item analysis focused on comparing the difficulty, item discrimination, and point-biserial statistics, as well as the KR-20 reliability statistic, for the three years that the test was administered. The results showed that the quality of the test items in 2009 and 2010 greatly improved from 2008. This was highlighted by three main findings:

* The number of items with appropriate difficulty levels increased
* The number of discriminating items on the tests increased
* The number of functional distractors per item increased, and the number of non-functional distractors per item decreased.

Additionally, it is interesting to note the authors’ statement that “with the introduction of the peer review process, item writers were encouraged to develop items that assessed high cognitive-level processes rather than recall of factual knowledge” (p. 923). This suggests that the peer review process not only identified and resolved flaws in the items themselves, but also started to improve the quality of the item writers.

Van der Merwe, H. (2015). Quality assuring multiple-choice question assessment in higher education. *South African Journal of Higher Education*, *29*(2), 279-297.

This article is based on a study of how multiple-choice assessments were developed and administered at a College of Education in South Africa. There were two primary objectives; to examine how teachers view multiple-choice assessments as useful to assess higher-level thinking and multiple intelligences; and to determine a systematic quality assurance process that can be used to improve item quality.

Based on interviews of experienced educators at this College of Education, most of the participants only used multiple-choice questions to assess lower-order or fact-based questions. However, some of these teachers also recognized that it is possible to develop higher-order multiple-choice items, which could possibly assess multiple intelligences. Overall, the educators viewed higher-order multiple choice items as important, but difficult to write. Particularly with assessing multiple intelligences, many of the educators believed that it could be done with multiple-choice questions, but they did not know how to write those types of items themselves. The author suggests that an inadequate item-writer training is a major reason for this.

To combat these shortfalls and increase the quality of item-writing, the author recommends using a systematic item development process, with these features:

* ensure conceptual understanding of the uses or multiple-choice items
* formal training for item writers
* peer review and discussion of the items
* professional editing by an outsider to the item writing process
* considering student feedback about the test items
* analyze item statistics and student performance

By implementing this process, the level of skill of the item writers should improve; the multiple levels of review should lead to further refinement of the items and content; and the post-test analyses can inform and validate decisions

Rios, J.A., Liu, O.L., & Bridgeman, B. (2014). Identifying low‐effort examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research, 161*, 69-82.

Examinee low motivation might be a threat to the validity of test score interpretation. Thus, it is important to identify low-effort examinees, especially when it is a low-stakes test for test takers but high-stakes for the educational institution or group. In their the literature review, Rios, Liu, and Bridgeman introduce two approaches to obtain a proxy of motivation: (a) self-report effort (SRE) measure and (b) response time effort (RTE) measure.

In SRE, test takers are asked the extent of their test-taking effort. Although measuring SRE is advantageous because it takes fewer resources to be administered, its employment has some limitations. First, test takers might not report their level of effort: some students might exaggerate their effort, and the other students might report reduced effort. Second, SRE is not measuring examinee motivation at the item level, but at the test level. It might be problematic because examinee’s effort could change during different test stages.

In RTE, test takers response time, which is readily accessible in computer-based testing, was used to differentiate engaged and disengaged examinees. This approach assumes that low-effort examinees will not spend ‘minimum time’ to read the item stem and all options. The first step to identifying low-effort examinees is to set ‘the minimum time’, in other words, a predetermined “threshold value that indicates rapid guessing” (p.71). Then, based on the threshold value, an overall RTE index can be calculated. If the RTE index is smaller than the threshold value, the examinee’s response is considered as that of a low-motivated examinee.

To compare SRE and RTE, empirical data of the ETS Proficiency Profile (EPP) which is a low-stakes test for test takers, was used. EPP was administered to 132 college seniors from five campuses in the Midwest state university system. Rios et al. focus on the examinee level in this comparison, because SRE doesn’t provide item level response. In this article, the overall RTE index is referred to as ‘the proportion of items on the test for which the examinee’s response time was less than the established threshold values’ (p.72). The examinee is considered unmotivated when the overall RTE index is greater than 10%.

Both SRE and RTE had “a stronger relationship with test performance than did measures of ability, such as cumulative GPA (rs = .06), SAT-Verbal (rs = .40), and SAT-Quantitative (rs = .32) scores” (p.76). However, the comparison shows that RTE filtering works better than SRE if response time are accessible. Out of 132 test takers, RTE detected 30 unmotivated examinees, SRE 19. RTE found out 11.62% of examinees who exaggerated their effort in SRE, and 2% of examinees who reported reduced effort. Additionally, RTE was found to have a slightly stronger relationship with test performance. When unmotivated examinees were filtered, the EPP score difference between motivated and unmotivated was greater with RTE (*d* range from 0.17~0.37) than SRE(*d* range from 0.13~0.19). Thus, RTE filtering may be the preferred method if response times are not paper-pencil format and available because RTE was found to identify more examinees whose have lower average scores than SRE.

Wise, S.L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163-183.

Throughout the literature review, Wise and Kong introduce two test taker strategies on tests: (a) solution behavior and (b) rapid-guessing behavior. In solution behavior, test takers actively engages in the test items and try to show what he or she knows and can do. However, in rapid-guessing behavior, test takers answers the question so rapidly they cannot read the question stem and options. Thus, most of answers in the rapid-guessing behavior are considered to be random, and the probability of a correct answer will be near the chance-level. The authors develop the concept of rapid-guessing behavior further by including the context of low-stakes test, which is the authors’ main focus. Under the global definition of the rapid-guessing behavior, rapid-guessing behavior might imply low motivation rather than hurrying-to-finish behavior.

It is completely up to test takers how much effort they want to put into engaging with the test items. Though some examinees don’t respond to test item, test givers don’t much regard those examinees’ test scores as threats to validity of the test score interpretation. In low-stakes tests, however, test givers have a responsibility to care about test taker’s effort or rapid-guessing behavior. Thus, the primary question of the authors is how to measure test taker effort? or How to differentiate solution behavior and rapid-guessing behavior? Wise et al. recommend considering response time information which is advantageous for a few reasons. First, it is not difficult to be acquired in the Computer-Base Test (CBT). Second, examinee’ effort can be measured in “the level of individual items” (p.179). To measure the response time information or examinees’ efforts, the authors coin new term, Response Time Effort (RTE). To calculate the RTE, the specification of item threshold should be clarified. Each item threshold comes from two predictors, item length and if an item used a figure, an illustration, or some other reading material.

For RTE analysis, the authors used empirical data from the Information Literacy Test (ILT), a low-stakes test, which was administered in the fall semester of 2003 to 506 freshman at a medium-size south eastern university. The response time distributions across items are investigated and some spike patterns were founded, long-time spikes and short-time spikes. Most of the ILT items have short-time spikes, which tend to be at the end of the speed test. Thus, it might imply that rapid-guessing behavior was happening during the ILT testing session.

Through data analysis, research hypotheses were tested and supported. Five hypotheses led to these findings: (a) RTE scores were found to be reliable. (b) RTE were correlated with other measures of examinee test taking efforts such as total test performance, self-reported effort, person fit, and total test time. (c) RTE were not correlated with measures of academic ability scores from the high-stakes test (i.e. SAT-Verbal, SAT-Quantitative). (d) “Two different examinee behaviors, with occurrences of rapid-guessing behavior yielding item responses whose accuracy did not exceed chance levels” (p.179). In other words, if RTE is low, the probability of correct response is not greater than when test takers randomly respond. (e) When motivation was filtered based on RTE scores, the results are highly similar to the result filtered by self-report effort. Employing RTE scores is expected to bring positive effects in measurement practice. First, they can be indicators to monitor test taker’s effort on each item and the whole test. Second, they can work for motivation filtering.

Rios, J.A., Guo, H., Mao, L., & Liu, O.L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing, 17*(1), 74-104.

According to Rios, Guo, Mao, and Liu, it is important to study the impact of the careless responding on low-stake test, because this threatens the validity of test-score interpretation and may lead to misleading information to policy make and institutional administrators. Through the literature review, Rios et al. investigate the impact of careless responding on the test results. Researchers have shown careless responding may result in biased individual-level ability estimates, item parameter estimates, reliability estimates, and correlations with external variables. However, “the true degree of bias on aggregated scores” has not been sufficiently investigated and “is still largely unknown” (p.76).

To deal with the threat of random careless responding, Rios et al. introduce two approaches: (a) flag and (b) filter. The authors pinpoint that the flagging random careless responses has a limitation of disregard for mean score differences and convergent, discriminant validity coefficients. Besides, this filtering procedure assumes that random careless responding is “unrelated to true ability” but the authors argue that this assumption is flawed. So, between the two approaches, the authors are more interested in the latter approach (i.e. filter).

Filtering the random careless responding has two approaches: (a) examinee-level filtering and (b) response-level filtering. In examinee-level filtering, an examinee’s total test score is deleted if the number of random careless responses for that examinee exceeds an a priori cut-off. As an example, Wise & Kong suggest 10% as a standard to create a subgroup of unmotivated respondents. However, if we filter the participant based on the standard of random careless responses, we might not use up to 25% of the sample data (Rios, Liu, & Bridgeman, 2014), which cannot be retained for any future study (e.g., longitudinal tracking) Thus, the authors recommend (b) response-level filtering using Item Response Theory (IRT) and provide practical recommendations for practitioners on when they face random careless responding problem.

Rios et al. evaluate the degree of underestimation in the true mean and compare the effectiveness of two filtering procedures, examinee-level filtering and response-level filtering. Two findings are:

(a) True mean was underestimated by 0.20 SDs, when total amount of careless response surpasses 6.25% ( easy test) or 12.5% (moderate or difficult test).

(b) True mean inflates 0.42 SDs, when careless responding is associate with ability and unmotivated test takers’ data was deleted.

Findings from this study can be used by test practitioners to deal with careless responses for group-based low-stakes assessments.