

Person-fit of English Language Learners (ELL) in K-12 High-Stakes Assessments

Lei Wan, Brad Wu
Pearson

The No Child Left Behind Act holds states using federal funds accountable for student academic achievement. Particularly, it mandates the inclusion of English language learners in the accountability system, as the ELL population continues to increase rapidly in recent years. According to the most recent statistics (National Center for Education Statistics, 2006), English language learner services were provided to 3.8 million students across the country (11% of all students) in the 2003-04 school year. California and Texas had the largest reported number of students receiving ELL services. In California, there were 1.6 million ELL students (26% of all students), and in Texas, there were 0.7 million ELL students (16% of all students). Due to the significant proportion of ELL students in public education, the NCLB Title I and III have established provisions guiding test development so that the special needs of ELL students can be addressed. Specifically, it is required that assessments should be given “in language and form most likely to yield accurate data on what such (ELL) students know and can do in academic content areas, until such students have achieved English language proficiency”(No Child Left Behind Act, 2002).

Despite the efforts to ensure appropriate test development, the validity and fairness of K-12 state assessments for ELLs has been a serious concern to many researchers (Abedi, 2004; Abedi & Lord, 2001; Butler & Stevens, 2001; Solano-Flores & Trumbull, 2003). These researchers argued that in the current K-12 settings, ELL students’ performance in content assessments may be negatively impacted by the language complexity of test items, and thus the validity and reliability of these tests are suspect for ELL students. Most of the past research on ELL assessment issues was based on mean performance comparison between ELL and language-majority students; some employed psychometric approaches such as generalizability theory, factor analysis, or differential item functioning.

In this study, we used a “person-fit” method to examine the validity of two statewide K-12 content tests for ELL students. Methods evaluating the fit of an individual’s response pattern to a psychometric model have been referred to as “person-fit” methods. Last decade witnessed a fast growth of interest in developing these methods, mainly because people realized that a total score could be an inadequate measure of an individual’s trait level. For example, a total score can hardly capture unusual testing behavior such as cheating, random guessing, faking, and omitting on purpose etc. By contrast, a scrutiny of individual response patterns may make it more likely to detect the unusual testing behavior. Also it may expose other test development and administration issues like low motivation, response alignment errors, mismatched curriculum, and effect of different language and culture etc. (Birenbaum, Kelly, & Tatsuoka, 1993; Lamprianou & Boyle, 2004; Levine & Drasgow, 1982; Meijer & Sijtsma, 2001; Stricker & Emmerich, 1999).

Studies related to person-fit fall into two categories. One category of studies focuses on the development of person-fit statistics. In most cases, person-fit statistics are developed based on how “unexpected” a response pattern is according to a psychometric model (often an IRT model). A response pattern is “unexpected” in the sense that it gives surprisingly correct responses to difficult questions or surprisingly incorrect responses to easy questions. More than a dozen of statistics have been developed, and the most widely used person-fit statistics include U and W (Wright & Masters, 1982; Wright & Stone, 1979), M (Molenaar & Hoijtink, 1990), and l_z and l_{zn} (Drasgow, Levine, & McLaughlin, 1991; Drasgow, Levine, & William, 1985). The other category of studies examines empirical test data using the statistics. They either explore factors that may lead to person-misfit or examine the impact of person-misfit on validity, reliability, and other psychometric properties of tests.

Not enough person-fit research has been conducted using empirical test data. Phillips (1986) investigated the effect of deleting misfitting response patterns on vertical equating results. It was found that removing misfitting patterns had little effect on item parameter estimation or equating. Schmitt, Cortina and Whitney (1993) found that removing aberrant response patterns from a battery of job-related tests would possibly increase concurrent validity. Rudner, Bracey and Skaggs (1996) examined the distribution of person-fit using NAEP TSA data. They did not find many students with aberrant response patterns, or significant relationship between person-fit and proficiency, item order, subscales or other key demographic variables. More recently, Lamprianou and Boyle (2004) studied the accuracy of measurement in the context of Mathematics National Curriculum test data (England) for ethnic minority students and students who speak English as a second language. It was found that these students are significantly more likely to generate aberrant response patterns.

Research Objectives

Person-fit of ELLs in the U.S. K-12 assessments has not been sufficiently investigated. Given that person-fit examination can provide evidence for validating tests for ELL students under NCLB, the aim of the present study is to explore if ELL students are more likely to generate aberrant response patterns than their English-speaking counterparts. Several sources that may potentially contribute to aberrancy were also probed. Specifically this study addresses the following research questions:

1. In general, are students likely to produce aberrant response patterns in high-stakes assessments?
2. Are ELL students more likely to produce aberrant response patterns than language-majority students? Controlled for ability, are ELL students more likely to produce aberrant response patterns?
3. Do content, item difficulty, and item language load affect the distribution of person-fit?

Methodology

Instruments

This study is based on the examination of two statewide tests in two states. The first instrument is a mathematics test in a Midwestern state for students in grades 3-8 and 11. For all grades except grade 11, this math test consists of four content strands: Number Sense, Pattern, Function, & Algebra, Data, Statistics, & Probability, and Geometry & Measurement. For grade 11, the Number Sense strand is not explicitly measured, so there are only three content strands. In this study, we selected samples from grades 5, 8, and 11 so that the grades span from elementary school to high school. There are three types of item in the test. Multiple-choice items are used extensively, and a small number of items are either short-answer items or constructed-response items. Both multiple-choice and short-answer items have a point value of 1, and constructed-response items are scored 0-4.

The other instrument is a science test in a Northwestern state for students in grades 5, 8 and 10. The science test consists of three content strands: System of Science, Inquiry of Science, and Application of Science. Multiple-choice and constructed-response items are used in the test. Constructed-response items are scored 0-2 or 0-4.

Since it is unlikely to solve constructed-response items by guessing, cheating or other test-taking strategies, and the number of these items in the tests is rather small, we dropped constructed-response items from the present study. The number of items used in this study is presented in Table 1.

Insert Table 1

Samples

Random samples of ELL and language-majority students were taken from a recent year's administration of the two tests. For the math test, we took 1000 ELL students and 2000 language-majority students in each grade. For the science test, different number of students was selected in each grade so that the ratio of ELL to language-majority students was about 1:2. The sample size for each test is also presented in Table 1. It should be made clear that in the two states, the actual ratio of ELL students is much lower than one third: in the Midwestern state, the percent of ELLs is 4.5-9.1% across grades, and in the Northwestern state, the percent is about 7%. By selecting many more ELL students, we hoped that it would be easier to detect person-fit issues, if there is any, with ELL students. As it will be described below, the unusually high percent of ELL students will not affect the computation or interpretation of results in this study.

Person-fit Statistic

Literature offers a large number of person-fit statistics based on various psychometric models. Meijer and Sijtsma (2001) did a comprehensive review of the person-fit methods. The review by and large indicates that there is no one "best" method to study person fit. The usefulness of the various statistics depends on the purpose of study, test characteristics (e.g. test length), as well as examinee characteristics (e.g. ability level). Lamprinou and Boyle (2004) also admitted that the person-fit statistics that they used might not be the "sole" candidate. Among all the available person-fit statistics, we

selected l_z (Drasgow, Levine, & Williams, 1985) for this study. This statistic was preferred over other person fit statistics for several reasons.

First, the statistic l_z has been used extensively and evaluated by many researchers. It has been recognized that l_z is one of the statistics with the most accurate detection for aberrant response patterns ((Drasgow & Guertler, 1987; Li & Olejnik, 1997; Nering & Meijer, 1998; Reise & Flannery, 1996; Schmitt, Cortina, & Whitney, 1993). Given that our tests are of moderate length, and test items have a large spread of item difficulty and good discrimination parameters, the l_z statistic is expected to perform rather well (Raise & Due, 1991; Raise & Flannery, 1996).

Second, given the right conditions (e.g. sufficient test length, good item discrimination, broad difficulty dispersion), the distribution of l_z is close to a standard normal distribution at all proficiency levels (Drasgow et al., 1985). Thus, l_z has an expected value of zero and a variance of one. This characteristic makes it simple to determine a cutoff point for judging person misfit. Large negative l_z values indicate response patterns that are lower in likelihood than the model predicts. In this study, a value less than -2 ($P < 0.0228$) was taken as an indication that the response pattern was aberrant in some way. Some researchers (Molenaar & Hoijtink, 1996; Reise, 1995) argued that l_z is standard normally distributed only when true θ values are used. When an estimate of θ is used, the variance of l_z is smaller than expected under the standard normal distribution using the true θ . In this study, we used MLE $\hat{\theta}$ produced by MULTILOG. Since likely our l_z has a smaller variance, a value of -2 is even more extreme than it appears to be in a normal distribution. In another word, what we captured in this study is rather extreme aberrant response patterns.

Lastly, the statistic l_z serves the purpose of this study appropriately in the sense that it does not focus on any specific type of aberrance but identify general misfit. This study is “explanatory” in nature, so indices which test against a specific alternative are not suitable.

The person-fit statistic l_z is developed from the log-likelihood function l_0 . Levin and Rubin (1979) first used the log-likelihood function to assess person fit, and because problems arose using l_0 , later Drasgow et al. (1985) proposed l_z . The log-likelihood function can be written as

$$l_0 = \sum_{g=1}^k \{ X_g \ln P_g(\theta) + (1 - X_g) \ln[1 - P_g(\theta)] \},$$

where subscript g indicates a specific item, k is the total number of items, X_g refers to item score, and $P(\theta)$ is the IRT probability for a person with ability θ to answer an item correctly. For the function l_0 , a distribution under the null hypothesis of fitting response pattern is unknown, implying that the classification of misfit or fit depends on ability.

Therefore, Drasgow et al. later proposed l_z as a standardized version of l_0 , which is less confounded with ability

$$l_z = \frac{l_0 - E(l_0)}{[Var(l_0)]^{1/2}},$$

where $E(l_0)$ and $Var(l_0)$ are the expectation and variance of l_0 , respectively:

$$E(l_0) = \sum_{g=1}^k \{P_g(\theta) \ln P_g(\theta) + [1 - P_g(\theta)] \ln [1 - P_g(\theta)]\},$$

and

$$Var(l_0) = \sum_{g=1}^k P_g(\theta) [1 - P_g(\theta)] \left[\ln \frac{P_g(\theta)}{1 - P_g(\theta)} \right]^2.$$

More importantly, the distribution of l_z is asymptotically standard normal. The characteristic greatly facilitates the interpretation of l_z .

Procedures

In this study, we first estimated item and θ parameters using statewide population so that the high percent of ELL students in the samples would not distort the calibration results. The population size is about 60,000 for the math tests and 20,000 for the science tests. For both short-answer and multiple-choice questions, the three-parameter logistic IRT model was used. Multilog version 7.3 was employed to do the calibration. Experience with similar data sets from previous testing cycles suggested that the statistical fit of the 3PL model should be satisfactory.

Next, we randomly selected the specified numbers (as in Table 1) of ELL and language-majority students, and computed the l_z statistic for each student or response pattern. A response pattern with a l_z value smaller than -2 was classified as “misfit”. A chi-square test was then conducted to check whether the proportion of misfitting response patterns was significantly different between the ELL and language-majority samples. As we suspected that language background may be confounded with achievement levels, we further classified students into different achievement levels based on the estimated θ values. Specifically, θ values smaller than -0.5 indicate low achievers, values in (-0.5, 0.5) indicate medium achievers, and values larger than 0.5 indicate high achievers. The proportion of misfit for ELL and language-majority students was compared respectively within each achievement level. By doing this, we hoped to control for the effect of achievement levels on comparing the likelihood of aberrant response patterns between ELL and language-majority students.

In order to explore potential sources of aberrancy, we further examined person-fit regarding content, item difficulty, and item language load. These examinations were conducted in a similar way as what described above, and the difference is that rather than computing l_z using all available items, l_z was computed separately using items associated with each of the content strands, item difficulty levels, and language load

levels. As it is known, there are four or three content strands for the math tests depending on grade, and three content strands for the science tests. After l_z was computed for each strand, the chi-square significance test was performed to evaluate the difference in misfit rate across ELL and language-majority students. Item difficulty levels were determined based on the b parameter of an item. The most difficult 1/3 of items were labeled “Hard”, the easiest 1/3 of items were labeled “Easy”, and the rest were labeled “Medium”. The chi-square significance test was performed separately within each level of item difficulty. To determine the language load of an item was more or less a judgment call, because there were no established criteria of readability for science and math tests. Finally, we were able to group items into two categories: items with higher language demand were labeled as “Heavy”, and items with less language demand were labeled as “Reduced”. The distinction was particularly blurred for the math tests, since in general the language demand for math was minimal.

Results

Descriptive statistics of the two interest groups (language majority and ELL) across three grades and two subjects are presented in Table 2. It includes mean and standard deviation of theta and l_z statistics for each group. On average, ELL students have smaller thetas than language-majority students, indicating ELLs have lower achievement level. This finding is not surprising, given that the achievement gap between ELL and language-majority students has been well recognized in the field. In terms of the person-fit statistic, ELL students have slightly lower l_z values on all math tests and the grade 5 science test with larger variations.

Insert Table 2

Number of students flagged as person misfit in both language-majority and ELL groups are presented from Table 3 to 10. Chi-square analysis is also presented in these tables.

Insert Table 3 to 10

As shown in Tables 3 and 4, at the test level, ELL status seems to be associated with person misfit as ELL students are more likely to generate aberrant response patterns. Examining responses to all items based on the entire sample, five (the three mathematics tests, grade 5 and grade 8 science tests) out of six tests show significant chi-square at the probability level of 0.05. The difference between the language-majority and ELL groups, however, tends to diminish after controlling for student achievement level. When numbers of misfit are compared within each achievement subgroup, chi-squares are not significant across ELL status except within the medium achievement group for grade 5 science. However, this second finding is not conclusive. As shown in the tables, the percent of flagged “misfitting” students tends to be so small in the ability breakdown analyses that the chi-square tests may not provide valid results. More research needs to be

done to verify if the ELL status truly has little effect on the likelihood of producing aberrant response patterns after partialling out achievement.

Comparisons are also made within each content strand (Tables 5 and 6). Chi-square statistics are not significant in most content strands. In mathematics, Geometry and Measurement is the only strand that consistently demonstrates difference of person misfit between language-majority and ELL students. For the strand of Data, Statistics, & Probability, misfit proportions do not differ significantly between language-majority and ELL groups on any grade. In science, chi-squares are significant only for two strands on grade 5 (System of Science and Inquiry in Science).

No specific pattern can be detected when analyses are performed within each item difficulty category. Based on Tables 7 and 8, observations of person misfit between the two interest groups are not consistent across the three item difficulty levels, grades, and subjects. Chi-squares are not significant in most comparisons.

Lastly, language load for all items is categorized and analysis is done within each level. Tables 9 and 10 show that difference in person misfit rates is significant for grade 5 science regardless of language load level. In mathematics, chi-squares are significant for reduced language load items on grades 8 and 11. The results of mathematics may be somewhat related to the previous finding that the strand Geometry and Measurement consistently demonstrates difference in person misfit, because Geometry and Measurement items rely on images and graphics more than language, so they are almost always classified as reduced language load items.

While ELL students do seem more likely to produce aberrant response patterns, no overt pattern is observed in the breakdown analyses. The results seem to suggest after taking account of achievement level, ELL status will not necessarily lead to difference in misfit rate. In addition, person misfit rates are consistently higher for ELL students when examining only the “Geometry and Measurement” strand items in mathematics. Tendency of higher misfit rates for ELL students is not observed in association with other content strands, item difficulty levels, or item language load levels.

Discussion

In searching for possible sources for aberrant response patterns, this study investigated the likelihood of person misfit by examinees’ ELL status, achievement level, content strands, item difficulty, and item reading load. As in the research by Lamprianou and Boyle (2004), results from our study also support that ELL students are more likely to generate aberrant response patterns in multiple choice tests. Two subjects and three grades were examined and the results were fairly consistent. However, the reason why ELL students are more likely to produce aberrant response patterns is not clear. One possible explanation to the difference is that ELL students usually are at a lower achievement level than the language-majority students. In other words, the difference in person misfit rate is probably attributed to students’ achievement level rather than the ELL status. A follow-up chi-square analysis was conducted using the data to check if the

achievement level makes difference in misfit rate. The results are presented in Table 11. It is clear from the table that misfit rates do vary significantly across the performance levels: high achievers rarely produce aberrant response patterns, and low achievers are more likely to generate misfitting response patterns.

Of course, it is very difficult to completely disentangle ELL status and academic performance. The achievement gap between the two interest groups is longstanding in K-12 settings, and the situation can hardly be improved, partly due to the current policy of classifying ELL status (Abedi, 2004). In many states, when an ELL student makes significant progress in content subjects, he or she will be reclassified as English fluent and will no longer be part of the ELL subgroup. The consequence is that the ELL subgroup will always have only low performing students.

It is also important to note that the number of examinees flagged with person misfit is fairly small and it rarely exceeds 5% of the samples in this study. This observation is consistent what Rudner, Bracey and Skaggs (1996) found with the NAEP data: person fit was exceptionally good. To some extent, this provides justification for the psychometric quality of statewide testing programs. However, though the limited cases of person misfit observed may be less critical in a norm-reference test setting, they are of more concern under the context of education accountability. When student test performance is under much higher stakes, such as calculating AYP or determining graduation, overlooking the aberrant response patterns may result in serious misleading decisions. Moreover, as already mentioned, by using a cutoff value of -2, this study captures only extremely aberrant response patterns. If we use a less tolerating criterion, for example, a cutoff point of -1.65 ($P = 0.05$), then the percent of misfitting responses is expected to noticeably increase.

Another interesting finding in our study is that ELL students tend to generate aberrant response patterns more than the language majority students on some (e.g. geometry and measurement) but not other content areas. Such observation is understandable since geometry and measurement strand usually involves less language complexity than other content strands, thus ELL students have higher probability of answering these items correctly. In this case, in comparison to ELL student performance in other content strands, their better performance on geometry and measurement would be considered “unexpected.” Therefore, this finding indicates that content area could be a source of person misfit. Moreover, it suggests that language proficiency demand does unduly affect ELL students in content assessments. ELL students are likely to receive lower ability estimates based on all types of items than what they would receive from performing items with less language constraints.

One limitation of this study is that we had relatively small sample sizes for the science test. This caused very small cell counts in some of the chi-square analyses, hence unreliable chi-square results. With larger sample size for the science test, we might even be able to identify different trends in person misfit in science than mathematics, as math is usually considered to be more “language free” than science. Another limitation of the

study is that we used only l_z to study person fit. Though one could contend that there is no one “best” statistic, one might still incorporate more than one well-researched person-fit statistics for cross-validation purpose.

To delve further into the aberrant responses patterns and investigate possible nature and cause of the misfit patterns, we recommend closer examination of the misfitting response patterns. An in-depth qualitative analysis can be helpful in understanding why students generate such response patterns, which includes test settings, motivation, language barrier, etc. When larger number of aberrant responses is observed, multivariate approach such as cluster and correspondence analysis can be applied to further investigate the relationship between the response patterns and student demographic or test characteristics.

References

- Albedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher*, 33, 4-14.
- Albedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, 14, 219-234.
- Birenbaum, M., Kelly, A. E., & Tatsuoka, K. K. (1993). Diagnosing knowledge states in algebra using the rule-space model. *Journal of Research in Mathematics Education*, 24, 442-459.
- Butler, F. A., & Stevens, R. (2001). Standardized assessment of the content knowledge of English language learners K-12: Current trends and old dilemmas. *Language Testing*, 18, 409-427.
- Drasgow, F., & Guertler, E. (1987). A decision-theoretic approach to the use of appropriateness measurement for detecting invalid test and scale scores. *Journal of Applied Psychology*, 72, 10-18.
- Drasgow, F., Levine, M.V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement*, 15, 171-191.
- Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67-86.
- Lamprianou, I., & Boyle, B. (2004). Accuracy of measurement in the context of Mathematics National Curriculum Tests in England for ethnic minority pupils and pupils who speak English as an additional language. *Journal of Educational Measurement*, 41, 239-259.
- Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology*, 35, 42-56.
- Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, 4, 269-290.
- Li, M. F., & Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement*, 21, 215-231.
- Meijer, R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25, 107-135.
- Molenaar, I. W., & Hoijsink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, 55, 75-106.
- National Center for Education Statistics (2006). Retrieved February 29, 2008, from <http://nces.ed.gov/fastfacts/display.asp?id=96>.

- Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the I_z person-fit statistic. *Applied Psychological Measurement*, 22, 53-69.
- No Child Left Behind Act (2002). P. L. 107-110. Retrieved February 29, 2008 from <http://www.ed.gov/policy/elsec/leg/esea02/107-110.pdf>.
- Phillips, S. E. (1986). The effects of deletion of misfitting persons on vertical equating via the Rasch model. *Journal of Educational Measurement*, 23, 107-118.
- Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement*, 19, 213-229.
- Reise, S. P., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement*, 15, 217-226.
- Reise, S. P. & Flannery, W. (1996). Assessing person-fit on measures of typical performance. *Applied Measurement in Education*, 9, 9-26.
- Ruder, L., Bracey, G., & Skaggs, G. (1996). The use of a person-fit statistic with one high-quality achievement test. *Applied Measurement in Education*, 9, 91-109.
- Schmitt, N., Cortina, J., & Whitney, D. (1993). Appropriateness fit and criterion-related validity. *Applied Psychological Measurement*, 17, 143-150.
- Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, 32, 3-13.
- Stricker, L. J., & Emmerich, W. (1999). Possible determinants of differential item functioning: familiarity, interest, and emotional reaction. *Journal of Educational Measurement*, 36, 347-366.
- Wright, B. D., & Masters, G.N. (1982). *Rating scale analysis*. Chicago: Mesa Press.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press.

Table 1: Number of Items and Sample Sizes for Each Test

| | Number of Items | Number of Language-Majority Students | Number of ELL Students |
|------------------|-----------------|--------------------------------------|------------------------|
| Grade 5 Math | 42 | 2000 | 1000 |
| Grade 8 Math | 44 | 2000 | 1000 |
| Grade 11 Math | 45 | 2000 | 1000 |
| Grade 5 Science | 33 | 800 | 424 |
| Grade 8 Science | 42 | 250 | 128 |
| Grade 10 Science | 42 | 300 | 166 |

Table 2. Average Theta and Lz Statistics in Each of the Interest Group

| | Theta | | | | Lz | | | |
|------------------|-------------------|------|--------|------|-------------------|------|------|-------|
| | Language Majority | | ELL | | Language Majority | | ELL | |
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| Grade 5 Math | .043 | .889 | -.715 | .880 | .203 | .728 | .171 | .857 |
| Grade 8 Math | .066 | .925 | -.727 | .877 | .234 | .858 | .100 | .980 |
| Grade 11 Math | .131 | .903 | -.761 | .757 | .240 | .877 | .142 | .990 |
| Grade 5 Science | .054 | .875 | -1.029 | .729 | .229 | .915 | .088 | 1.152 |
| Grade 8 Science | .043 | .891 | -1.382 | .679 | .172 | .890 | .197 | 1.089 |
| Grade 10 Science | .019 | .890 | -1.386 | .700 | .203 | .886 | .301 | 1.000 |

Table 3. Comparison of Number of Students with/without Person Misfit and the Chi-Square Statistics Stratified by Achievement Level—Mathematics

| | Number of Items | Total Sample Size | Language Majority Flagged as Person Misfit | Language Majority with Normal Person Fit | ELL Students Flagged as Person Misfit | ELL Students with Normal Person Fit | Chi-Square | P |
|-----------------------------|-----------------|-------------------|--|--|---------------------------------------|-------------------------------------|------------|---------|
| Grade 5 Overall | 42 | 3000 | 14 | 1986 | 20 | 980 | 10.0551 | .0015 |
| Grade 5 Low Achievement | 42 | 1080 | 11 | 481 | 16 | 572 | .2588 | .6109 |
| Grade 5 Medium Achievement | 42 | 1203 | 3 | 878 | 4 | 318 | 3.3144 | .0687* |
| Grade 5 High Achievement | 42 | 717 | 0 | 627 | 0 | 90 | N/A | N/A |
| Grade 8 Overall | 44 | 3000 | 23 | 1977 | 28 | 972 | 10.8611 | .0010 |
| Grade 8 Low Achievement | 44 | 1122 | 12 | 508 | 23 | 579 | 2.1131 | .1460 |
| Grade 8 Medium Achievement | 44 | 1111 | 10 | 793 | 4 | 304 | .0051 | .9431* |
| Grade 8 High Achievement | 44 | 767 | 1 | 766 | 1 | 89 | 2.835 | .0922* |
| Grade 11 Overall | 45 | 3000 | 20 | 1980 | 20 | 980 | 5.0676 | .0244 |
| Grade 11 Low Achievement | 45 | 1206 | 1 | 525 | 7 | 673 | 3.17 | .0750* |
| Grade 11 Medium Achievement | 45 | 1001 | 19 | 736 | 11 | 246 | 2.4393 | .1183 |
| Grade 11 High Achievement | 45 | 793 | 0 | 793 | 2 | 74 | 19.4816 | <.0001* |

* 25% or 50% of the cells have expected counts less than 5. Chi-square may not be a valid test.

Table 4. Comparison of Number of Students with/without Person Misfit and the Chi-Square Statistics Stratified by Achievement Level—Science

| | Number of Items | Total Sample Size | Language Majority Flagged as Person Misfit | Language Majority with Normal Person Fit | ELL Students Flagged as Person Misfit | ELL Students with Normal Person Fit | Chi-Square | P |
|-----------------------------|-----------------|-------------------|--|--|---------------------------------------|-------------------------------------|------------|---------|
| Grade 5 Overall | 33 | 1224 | 14 | 786 | 28 | 396 | 19.7029 | <.0001 |
| Grade 5 Low Achievement | 33 | 411 | 2 | 131 | 7 | 271 | .4321 | .5110* |
| Grade 5 Medium Achievement | 33 | 433 | 11 | 302 | 19 | 101 | 20.4143 | <.0001 |
| Grade 5 High Achievement | 33 | 380 | 1 | 353 | 2 | 24 | 16.9790 | <.0001* |
| Grade 8 Overall | 42 | 378 | 3 | 247 | 6 | 122 | 4.4300 | .0353* |
| Grade 8 Low Achievement | 42 | 125 | 0 | 34 | 3 | 88 | 1.1484 | .2839* |
| Grade 8 Medium Achievement | 42 | 131 | 3 | 92 | 3 | 33 | 1.6000 | .2059* |
| Grade 8 High Achievement | 42 | 122 | 0 | 121 | 0 | 1 | N/A | N/A |
| Grade 10 Overall | 42 | 466 | 6 | 294 | 5 | 161 | .4749 | .4907 |
| Grade 10 Low Achievement | 42 | 158 | 0 | 35 | 2 | 121 | .5764 | .4477* |
| Grade 10 Medium Achievement | 42 | 170 | 5 | 126 | 3 | 36 | 1.0066 | .3157* |
| Grade 10 High Achievement | 42 | 138 | 1 | 133 | 0 | 4 | .0301 | .8623* |

* 25% or 50% of the cells have expected counts less than 5. Chi-square may not be a valid test.

Table 5. Comparison of Number of Students with/without Person Misfit and the Chi-Square Statistics Stratified by Content Strand—
Mathematics

| | Number of Items | Total Sample Size | Language Majority Flagged as Person Misfit | Language Majority with Normal Person Fit | ELL Students Flagged as Person Misfit | ELL Students with Normal Person Fit | Chi-Square | P |
|--|-----------------|-------------------|--|--|---------------------------------------|-------------------------------------|------------|-------|
| Grade 5 Overall | 42 | 3000 | 14 | 1986 | 20 | 980 | 10.0551 | .0015 |
| Grade 5 Number Sense | 16 | 3000 | 24 | 1976 | 13 | 987 | 0.0547 | .8150 |
| Grade 5 Pattern, Function, & Algebra | 7 | 3000 | 54 | 1946 | 31 | 969 | 0.3874 | .5336 |
| Grade 5 Data, Statistics, & Probability | 8 | 3000 | 66 | 1934 | 30 | 970 | 0.1937 | .6599 |
| Grade 5 Geometry & Measurement | 11 | 3000 | 69 | 1931 | 57 | 943 | 8.3880 | .0038 |
| Grade 8 Overall | 44 | 3000 | 23 | 1977 | 28 | 972 | 10.8611 | .0010 |
| Grade 8 Number Sense | 11 | 3000 | 59 | 1941 | 49 | 951 | 7.3046 | .0069 |
| Grade 8 Pattern, Function, & Algebra | 13 | 3000 | 60 | 1940 | 39 | 961 | 1.6922 | .1933 |
| Grade 8 Data, Statistics, & Probability | 8 | 3000 | 67 | 1933 | 35 | 965 | 0.0457 | .8308 |
| Grade 8 Geometry & Measurement | 12 | 3000 | 48 | 1952 | 38 | 962 | 4.6927 | .0303 |
| Grade 11 Overall | 45 | 3000 | 20 | 1980 | 20 | 980 | 5.0676 | .0244 |
| Grade 11 Pattern, Function, & Algebra | 18 | 3000 | 46 | 1954 | 41 | 959 | 7.6707 | .0056 |
| Grade 11 Data, Statistics, & Probability | 14 | 3000 | 66 | 1934 | 22 | 978 | 2.8331 | .0923 |
| Grade 11 Geometry & Measurement | 13 | 3000 | 37 | 1963 | 33 | 967 | 6.1507 | .0131 |

Table 6. Comparison of Number of Students with/without Person Misfit and the Chi-Square Statistics Stratified by Content Strand—
Science

| | Number of Items | Total Sample Size | Language Majority Flagged as Person Misfit | Language Majority with Normal Person Fit | ELL Students Flagged as Person Misfit | ELL Students with Normal Person Fit | Chi-Square | P |
|---------------------------------|-----------------|-------------------|--|--|---------------------------------------|-------------------------------------|------------|--------|
| Grade 5 Overall | 33 | 1224 | 14 | 786 | 28 | 396 | 19.7029 | <.0001 |
| Grade 5 System of Science | 16 | 1224 | 26 | 774 | 25 | 399 | 4.8599 | .0275 |
| Grade 5 Inquiry in Science | 12 | 1224 | 17 | 783 | 29 | 395 | 17.0305 | <.0001 |
| Grade 5 Application of Science | 5 | 1224 | 13 | 787 | 8 | 416 | .1126 | .7372 |
| Grade 8 Overall | 42 | 378 | 3 | 247 | 6 | 122 | 4.4300 | .0353 |
| Grade 8 System of Science | 21 | 378 | 11 | 239 | 7 | 121 | .2132 | .6443 |
| Grade 8 Inquiry in Science | 13 | 378 | 12 | 238 | 7 | 121 | .0793 | .7782 |
| Grade 8 Application of Science | 8 | 378 | 11 | 239 | 7 | 121 | .2132 | .6443 |
| Grade 10 Overall | 42 | 466 | 6 | 294 | 5 | 161 | .4749 | .4907 |
| Grade 10 System of Science | 21 | 466 | 10 | 290 | 10 | 156 | 1.8837 | .1699 |
| Grade 10 Inquiry in Science | 15 | 466 | 8 | 292 | 2 | 164 | 1.0876 | .2970 |
| Grade 10 Application of Science | 6 | 466 | 16 | 284 | 5 | 161 | 1.3381 | .2474 |

Table 7. Comparison of Number of Students with/without Person Misfit and the Chi-Square Statistics Stratified by Item Difficulty—
Mathematics

| | Number of Items | Total Sample Size | Language Majority Flagged as Person Misfit | Language Majority with Normal Person Fit | ELL Students Flagged as Person Misfit | ELL Students with Normal Person Fit | Chi-Square | P |
|----------------------------------|-----------------|-------------------|--|--|---------------------------------------|-------------------------------------|------------|-------|
| Grade 5 Overall | 42 | 3000 | 14 | 1986 | 20 | 980 | 10.0551 | .0015 |
| Grade 5 Low Difficulty Items | 14 | 3000 | 56 | 1944 | 27 | 973 | 0.0248 | .8749 |
| Grade 5 Medium Difficulty Items | 14 | 3000 | 18 | 1982 | 7 | 993 | 0.3227 | .5700 |
| Grade 5 High Difficulty Items | 14 | 3000 | 40 | 1960 | 27 | 973 | 1.4961 | .2213 |
| Grade 8 Overall | 44 | 3000 | 23 | 1977 | 28 | 972 | 10.8611 | .0010 |
| Grade 8 Low Difficulty Items | 14 | 3000 | 54 | 1946 | 52 | 948 | 12.2244 | .0005 |
| Grade 8 Medium Difficulty Items | 14 | 3000 | 26 | 1974 | 14 | 986 | 0.0507 | .8219 |
| Grade 8 High Difficulty Items | 16 | 3000 | 45 | 1955 | 25 | 975 | 0.1828 | .6689 |
| Grade 11 Overall | 45 | 3000 | 20 | 1980 | 20 | 980 | 5.0676 | .0244 |
| Grade 11 Low Difficulty Items | 15 | 3000 | 42 | 1958 | 29 | 971 | 1.8465 | .1742 |
| Grade 11 Medium Difficulty Items | 15 | 3000 | 14 | 1986 | 15 | 985 | 4.4569 | .0348 |
| Grade 11 High Difficulty Items | 15 | 3000 | 55 | 1945 | 34 | 966 | 0.9785 | .3226 |

Table 8. Comparison of Number of Students with/without Person Misfit and the Chi-Square Statistics Stratified by Item Difficulty—
Science

| | Number of Items | Total Sample Size | Language Majority Flagged as Person Misfit | Language Majority with Normal Person Fit | ELL Students Flagged as Person Misfit | ELL Students with Normal Person Fit | Chi-Square | P |
|----------------------------------|-----------------|-------------------|--|--|---------------------------------------|-------------------------------------|------------|--------|
| Grade 5 Overall | 33 | 1224 | 14 | 786 | 28 | 396 | 19.7029 | <.0001 |
| Grade 5 Low Difficulty Items | 11 | 1224 | 35 | 765 | 31 | 393 | 4.6837 | .0304 |
| Grade 5 Medium Difficulty Items | 11 | 1224 | 10 | 790 | 6 | 418 | .0585 | .8088 |
| Grade 5 High Difficulty Items | 11 | 1224 | 15 | 785 | 15 | 409 | 3.2045 | .0734 |
| Grade 8 Overall | 42 | 378 | 3 | 247 | 6 | 122 | 4.4300 | .0353 |
| Grade 8 Low Difficulty Items | 14 | 378 | 9 | 241 | 8 | 120 | 1.3841 | .2394 |
| Grade 8 Medium Difficulty Items | 14 | 378 | 8 | 242 | 9 | 119 | 2.8931 | .0890 |
| Grade 8 High Difficulty Items | 14 | 378 | 8 | 242 | 5 | 123 | .1272 | .7214 |
| Grade 10 Overall | 42 | 466 | 6 | 294 | 5 | 161 | .4749 | .4907 |
| Grade 10 Low Difficulty Items | 14 | 466 | 7 | 293 | 2 | 164 | .7186 | .3966 |
| Grade 10 Medium Difficulty Items | 14 | 466 | 6 | 294 | 2 | 164 | .4005 | .5268 |
| Grade 10 High Difficulty Items | 14 | 466 | 9 | 291 | 4 | 162 | .1373 | .7109 |

Table 9. Comparison of Number of Students with/without Person Misfit and the Chi-Square Statistics Stratified by Reading Load—
Mathematics

| | Number of Items | Total Sample Size | Language Majority Flagged as Person Misfit | Language Majority with Normal Person Fit | ELL Students Flagged as Person Misfit | ELL Students with Normal Person Fit | Chi-Square | P |
|-------------------------------|-----------------|-------------------|--|--|---------------------------------------|-------------------------------------|------------|--------|
| Grade 5 Overall | 42 | 3000 | 14 | 1986 | 20 | 980 | 10.0551 | .0015 |
| Grade 5 Heavy Reading Items | 14 | 3000 | 48 | 1952 | 29 | 971 | 0.6665 | .4143 |
| Grade 5 Reduced Reading Items | 28 | 3000 | 87 | 1913 | 54 | 946 | 1.6410 | .2002 |
| Grade 8 Overall | 44 | 3000 | 23 | 1977 | 28 | 972 | 10.8611 | .0010 |
| Grade 8 Heavy Reading Items | 13 | 3000 | 61 | 1939 | 30 | 970 | 0.0057 | .9400 |
| Grade 8 Reduced Reading Items | 31 | 3000 | 31 | 1969 | 29 | 971 | 6.1990 | .0128 |
| Grade 11 Overall | 45 | 3000 | 20 | 1980 | 20 | 980 | 5.0676 | .0244 |
| Grade 11 Heavy Reading Items | 17 | 3000 | 48 | 1952 | 31 | 969 | 1.2741 | .2590 |
| Grade11 Reduced Reading Items | 28 | 3000 | 21 | 1979 | 31 | 969 | 16.4486 | <.0001 |

Table 10. Comparison of Number of Students with/without Person Misfit and the Chi-Square Statistics Stratified by Reading Load—
Science

| | Number of Items | Total Sample Size | Language Majority Flagged as Person Misfit | Language Majority with Normal Person Fit | ELL Students Flagged as Person Misfit | ELL Students with Normal Person Fit | Chi-Square | P |
|-------------------------------|-----------------|-------------------|--|--|---------------------------------------|-------------------------------------|------------|--------|
| Grade 5 Overall | 33 | 1224 | 14 | 786 | 28 | 396 | 19.7029 | <.0001 |
| Grade 5 Heavy Reading Items | 13 | 1224 | 17 | 783 | 21 | 403 | 7.3667 | .0066 |
| Grade 5 Reduced Reading Items | 20 | 1224 | 24 | 776 | 24 | 400 | 5.2057 | .0225 |
| Grade 8 Overall | 42 | 378 | 3 | 247 | 6 | 122 | 4.4300 | .0353 |
| Grade 8 Heavy Reading Items | 23 | 378 | 5 | 245 | 5 | 123 | 1.1944 | .2744 |
| Grade 8 Reduced Reading Items | 19 | 378 | 6 | 244 | 5 | 123 | .6798 | .4097 |
| Grade 10 Overall | 42 | 466 | 6 | 294 | 5 | 161 | .4749 | .4907 |
| Grade 10 Heavy Reading Items | 27 | 466 | 4 | 296 | 2 | 164 | .0139 | .9062 |
| Grade10 Reduced Reading Items | 15 | 466 | 11 | 289 | 6 | 160 | .0008 | .9770 |

Table 11. Comparison of Number of Students with/without Person Misfit and the Chi-Square Statistics Stratified by Achievement – Mathematics and Science

| | Number of Items | High Achiever Flagged as Misfit | High Achiever Flagged as Person Fit | Medium Achiever Flagged as Misfit | Medium Achiever Flagged as Person Fit | Low Achiever Flagged as Misfit | Low Achiever Flagged as Person Fit | Chi-Square | P |
|------------------|-----------------|---------------------------------|-------------------------------------|-----------------------------------|---------------------------------------|--------------------------------|------------------------------------|------------|--------|
| Grade 5 Math | 42 | 0 | 717 | 7 | 1196 | 27 | 1053 | 29.4870 | <.0001 |
| Grade 8 Math | 44 | 2 | 765 | 14 | 1097 | 35 | 1087 | 24.3213 | <.0001 |
| Grade 11 Math | 45 | 2 | 791 | 30 | 971 | 8 | 1198 | 32.2206 | <.0001 |
| Grade 5 Science | 33 | 3 | 377 | 30 | 403 | 9 | 402 | 25.8964 | <.0001 |
| Grade 8 Science | 42 | 0 | 122 | 6 | 125 | 3 | 122 | 5.7017 | 0.0578 |
| Grade 10 Science | 42 | 1 | 137 | 8 | 162 | 2 | 156 | 6.4811 | 0.0391 |