APPENDIX E:
INDEPENDENT PSYCHOMETRIC QUALITY ASSURANCE REVIEW

AES –Verification Scaling & Equating

Assessment and Evaluation Services (AES) served as subcontractor for the Independent Psychometric Quality Assurance Review. AES reviewed and replicated all psychometric procedures connected to the scaling and equating of the assessments. The prime contractor, Measurement Incorporated, provided to AES all the same data which is provided to their psychometric unit. The prime contractor also provided AES with all necessary software settings, documentation, and the results of its own psychometric analyses for verification by AES. AES performed its analysis independent of the prime contractors work.

AES has experience in performing quality control services in testing programs in Ohio, New Jersey, New York, Virginia and Washington. In those states AES verifies similar psychometric analyses as described in the Michigan program. AES has verified analyses involving Item Response Theory (IRT) equating, scaling, and item analysis. AES staff has expertise in IRT models including but not limited to one-parameter IRT model, Partial Credit Model, and three-parameter IRT model.

AES also has experience in coordinating this work with various contractors. In the past we have found that working with the psychometric staff of another company requires extensive planning and coordinated scheduling. It is essential that the quality control work be extensive and accurate, but it is equally important that it be completed in a timely fashion so that overall project schedules can be met. This requires that both AES and the psychometric staff of the prime contractor work closely in planning for the transfer of data and analysis results to AES. Likewise, it is important for AES to complete the checking and transfer our results to the MEAP office for verification.

AES met with MEAP and Measurement Incorporated to discuss the plans and schedules regarding the implementation of this contract. The intent was co-ordinate the activities between the contractor and AES to ensure that the verification procedures were implemented in a smooth and accurate manner.

Assessment and Evaluation Services provided the verification of scaling and equating activities for the MEAP 2010 analyses. Verifications were done for pre-equating, post-equating, and field test item analysis. The primary scaling of the Reading, Mathematics, Writing, Science, and Social Science was done with the Rasch model using the partial credit model for open-ended items.

This Appendix provides a description of the steps AES undertook to provide replication of the MEAP 2010 analyses. The workflow was organized so that Measurement Incorporated technical staff and Assessment and Evaluation staff worked independently on each step. Once major portions of the analysis were completed, Assessment and Evaluation Services compiled the two sets of results into a comparison spreadsheet. These spreadsheets were then examined by Measurement Incorporated, Assessment and Evaluation Services, and the MEAP office to determine if the replication was successful. When discrepancies between MI and AES results

occurred during the steps they were often resolved before the comparison spreadsheets were completed.

The MEAP Grade 3-9 contains assessments in five subject areas: Reading, Writing, Mathematics, Science, and Social Science. Reading and Writing while scaled separately also provide a combined English Language Arts score. Reading, Writing, and Mathematics were assessed in Fall 2010 at Grades 3 through 8. Science was assessed at Grades 5 and 8 and Social Studies was assessed at Grades 6 and 9. Each of the 22 subject/grade tests had multiple test forms. In most cases the test forms were structured so that forms contained the same census test items and different field tests

The four major analyses for the project are detailed below in steps from the Assessment and Evaluation Services perspective.

The project has been partitioned into four Analysis Sets for description. These did occur sequentially through the Fall and Winter of 2010-2011.

Analysis Set 1- Pre Equating Scaling for Reading, Mathematics, Science, and Social Science

Analysis Set 2- Post Equating Scaling for Reading, Mathematics, Writing, Science, and Social Science

Analysis Set 3-Field Test Item Analysis

Analysis Set 1- Pre Equating Scaling for Reading, Mathematics, Science, and Social Science

Two key components of the test scaling are provided by the MEAP program prior to test administration. The test forms are detailed in a Test Map spreadsheet which provides information on item scoring, form composition, and organization. Within the test map the item parameter analysis information from previous field tests is presented since test items have been previously calibrated to the reporting scale through item response theory based on the data collected in previous years a pre-equating analysis can be done to produce preliminary raw score to scale score tables.

Step 1-Determine Test Form Item/Point Counts and Identify Item Parameters

Test Maps were used to provide the number of test items, test points, and Rasch equated difficulty parameters.

Step 2- Pre Equate based on Field Test Values

Raw Score to Scale Score tables were developed for each test form using the WINSTEPS program.

Step 3- Develop Comparison Spreadsheets

Raw to Scale Score tables from MI and AES were compared to determine if the scale scores, error term, and performance level were equivalent. This analysis was data free and no differences were expected. The actual differences were very small. A few Rasch thetas were different by .0001, while all scale scores, scaled error, and performance levels were identical.

Analysis Set 2- Post Equating Scaling for Reading, Mathematics, Science, and Social Science

Preliminary raw scale to scale score tables were developed from the Pre Equating process. The Post Equating analysis took place after student work had been scored and was based on a data set which included almost all students. The Post Equating analysis checked for the stability of the item difficulty parameters. Differences from the Pre Equated values from the bank and the Post Equated values from the assessment were examined.

Step 1-Check Data File for Unreasonable Values

In large data files often item response values which are implausible are found. This is particularly true in scored data files. AES examined the item score fields for values which were not plausible given the form designation and the item key. When values were found, AES notified MI so these instances could be investigated.

Step 2-Initial WINSTEPS Run

The data was analyzed by the WINSTEPS program to develop initial unanchored Rasch difficulties.

Step 3-Develop an Initial Comparison Spreadsheet

The initial Rasch Difficulties, n-counts, P-values, and Point-Biserials were compared. Again very few and very small differences were found.

Step 4- Examine Item Stability

An item stability criteria of .50 was determined. If the Rasch Difficulty value from the Pre Equating differed more than the equated Post Equating Rasch Difficulty by more than .50 that item was dropped from the equating of Post values to item bank values. MI and AES identified the same items to be dropped from the equating.

Step 5- Post Equating WINSTEPS Run

An equating constant between bank values and initial WINSTEPS values was developed using common items that were not dropped in Step 4. The equating constant was then applied to the initial WINSTEPS run to put the item difficulties on the original scale. New raw score to scale score tables were generated.

Step 6- Develop Post Equating Comparison Spreadsheet

Raw to Scale Score tables from MI and AES were compared to determine if the scale scores, error term, and performance level were equivalent. The actual differences were very small. A few Rasch thetas were different by .0001, one scale score rounded differently due to a .0001 difference, while all other scale scores, scaled error, and performance levels were identical.

Analysis Set 3-Field Test Item Analysis

Field test items were analyzed to provide item data for committee review and parameter values for future form scaling and equating. All subject forms contained field test items. The analysis consisted of information about the performance of the item for the total population and item performance for ethnic and gender groups which yielded differential item performance statistics.

Step 1-Check Data File for Unreasonable Values

In large data files often item response values which are implausible are found. This is particularly true in scored data files. AES examined the item score fields for values which were not plausible given the form designation and the item key. When values were found, AES notified MI so these instances could be investigated.

Step 2-Item Statistics for the Total Group

Analysis was run by grade/subject on all forms to develop Rasch item parameters, p-values, and point-biserials. The census items were used as base values and their difficulty parameters fixed so that field test item parameters were placed on the same scale. This method was also used for the writing field test items.

Step 3-Item Statistics for Ethnic and Gender Group and calculation of Differential Performance Indicators

The data was analyzed by ethnic/gender groupings: female/male and African-American/White. N-counts, p-values, and point-biserials were calculated. Differential Item Performance statistics developed were the Mantel-Haenszel statistics and Standardized Mean Differences.

Step 4- Develop Field Test Comparison Spreadsheet

The field test comparison spreadsheet was item based. For each item the total and group n-counts p-values, point-biserials, and item parameters were compared. Differences were small and were due to data matrix designs. Differential Item Performance statistics were also compared. Differences between MI and AES values were due to performance groupings and specific software differences.