4.0 Reliability and Validity

4.1 Reliability

Reliability is quantification of the consistency of results from a measurement. The ability to measure consistently is a necessary prerequisite to making appropriate score interpretations (i.e., showing evidence of valid use of the results). For the portfolio-based Alternate Maryland School Assessment (ALT-MSA), reliability relates primarily to the consistency with which the specified scoring process can be employed by scorers.

Pearson Educational Measurement (PEM) uses several procedures to help ensure that all ALT-MSA portfolios are scored reliably.

- Training procedures and materials are standardized for all participating scorers. This is true not only within an administration year, but to the extent possible, across administrations.

- The scoring process and scoring rules are clearly documented so there is no ambiguity as to how scoring issues should be handled.

- Validity and reliability reports are reviewed on a regular basis to identify scorer drift, outliers, and general scoring misconceptions (as defined by the portfolios in the validity set). In 2003-2004, the reports were used to inform scorers of their validity and reliability scores. The scoring director analyzed the reports, informed the supervisor of any concerns and the scoring supervisor in turn reviewed pertinent reports with the scorer. Supervisors monitored these scorers by backreading more frequently and checking their reliability and validity rates.

Reader Agreement

Because every portfolio is read at least twice by different readers, agreement between the readers is a common measure of reliability. These data are monitored on a daily basis by PEM during the scoring process. Daily reader agreement reports show the percent perfect agreement of each reader against all other readers.

Tables 16-18 in Appendix A summarize reader agreement for each subject area by content standard and overall for the current test administration. Reader agreement rate is expressed in terms of perfect agreement (i.e., the percentage of cases in which the first reader's score equals the second reader's score).

High inter-reader agreement implies that the scoring process and scoring rules are being applied consistently across readers.

In 2003-2004, the backreading procedure was performed before monitors were scanned into the system. This process included a review of the scored monitor and the portfolio in order to determine scorer accuracy. This process was used to backread single monitors. The procedure also included comparing first and second score monitors in order to increase the number of monitors backread by supervisors each day and provide

immediate attention to any scoring inconsistencies. When an inconsistency was discovered, the scoring supervisor used it as a teachable moment to inform the scorer of the mistake. This allowed the scorer with the opportunity to change their score in order to provide the student with an accurate score. In 2003 – 2004 this process may have resulted in artificially inflated reliability rates. In future years, the backreading and resolution process will be conducted independently to provide more accurate and actionable reader statistics.

4.2 Validity

As previously stated, assessment results must show evidence of reliability for the purpose for which they were intended before they can show evidence of validity. Validity relates to the appropriateness or strength of the assessment results for making specific interpretations about what students know and can do. As documented in Standard 1.1 of the Standards for Educational and Psychological Measurement (1999), validity evidence should be collected for every intended interpretation and use of the scores resulting from a measurement instrument.

The purpose of the ALT-MSA is multifold, as outlined in the first chapter of this document. First and foremost, the assessment is intended to comply with federal mandates, to inform ongoing instruction and to help teachers plan instruction for the following year. A student's ALT-MSA results and portfolio should help teachers determine his/her level of functioning at the time of the assessment, indicate specific skills acquired and those requiring continued instruction, and identify supports and assistive technologies previously employed. This information can be used to inform the review and revision of a student's IEP and support the construction of a well-structured plan for instruction and assessment in the upcoming year. In addition, by reviewing previously submitted portfolios in conjunction with historical data, teachers can get an indication of a student's rate of progress relative to certain subject and content standard areas.

Second, the ALT-MSA is intended to hold teachers/schools/districts accountable for implementing standards-based curriculum and using assessment results to improve student learning. The annual ALT-MSA development and administration process helps to ensure that teachers/schools/districts are focused on the development, instruction, and assessment of challenging performance goals that are aligned with the state content standards.

Finally, ALT-MSA results should inform and support program evaluation at the classroom, school, and district level. This includes identification of both resources that may further support instruction, and topics for professional development of staff.

Intrinsic Rational Validity Evidence

Intrinsic rational validity is evidence that exists as an artifact of the test development process. The evidence is intrinsic, because it is built into the test. It is rational because it is derived from rational inferences about the kind of tasks that will best meet the measurement goals of the assessment (Ebel, 1983).

To a large extent, the process that was implemented by the MSDE to develop and design the ALT-MSA is, in and of itself, evidence for the use of ALT-MSA test results in supporting the goals defined above. The MSDE took great care to ensure the right people were involved in all aspects of developing and implementing the ALT-MSA program. Advisory specialists in alternate assessment met at length on many occasions to determine what the assessment should look like given the assessment mandates and intent. In addition, the state implemented a structured process to support the identification of desired assessment components and designs. This process included an Advisory Committee review of the alternate standards and assessments for many states across the nation. Such a comprehensive review helped to ensure ALT-MSA results would be viewed as useful and important to teachers and parents alike.

Content- and Curricular-Related Validity Evidence

Content-related validity evidence addresses the extent to which the assessment tasks adequately align to the material or standards intended as the focus of assessment. Several features of the annual ALT-MSA development process provide evidence that the results measure the intended content standard or access skills objectives. For one, it is clearly specified in teacher training and the Test Administration and Coordination Manual that mastery objectives must be aligned to state content standards or access skills objectives. The goal of the assessment to measure skills aligned to the state standards is highlighted as often as possible.

In addition, content experts from the MSDE review every mastery objective to ensure alignment to, and appropriate representation of, the underlying objective identified by the test examiner. These experts provide feedback to test examiners regarding how the mastery objective can be improved and whether alignment is an issue.

Face Validity

Face validity addresses the question of whether or not the assessment appears to measure what it supposed to measure. This is an extremely important component of any assessment program. If parents, teachers, or community members do not perceive a test as relevant or do not understand its purpose, they are less likely to give it their attention and support. The extent to which a test possesses face validity is typically gauged by the response of stakeholders to using test results to inform instruction and monitor accountability. One way to obtain this information is through a well-crafted survey administered to parents, teachers, and other stakeholder groups of interest.

The MSDE asks teachers, test coordinators and school administrators to complete a survey about the ALT-MSA development and administration process. The survey includes Likert-type statements (i.e., agree, strongly agree, etc. . .) and open-ended questions intended to (in part) provide some insight to how the ALT-MSA is perceived. This information is used by the MSDE to gauge test acceptance and determine what can be done to improve it in the future.

Consequential Validity Evidence

When establishing evidence to support the appropriateness of a test relative to a set of assessment goals, it is important to evaluate both the intended and unintended

consequences of the assessment process and results (Messick, 1993). This is especially the case for a portfolio-based assessment such as the ALT-MSA where the assessment development and administration process can be relatively complex and labor-intensive.

In addition to providing information about how the ALT-MSA is perceived by stakeholders, survey results may assist the MSDE in making inferences about the consequences of the ALT-MSA (both positive and negative). For example, one of the open-ended questions posed to teachers and test coordinators in the survey is: "Next year as test coordinator/teacher I plan to have . . ." If, in reviewing responses to this question, we find a significant number teachers state that they "plan to develop assessment tasks that better reflect their student's IEP," the MSDE has some evidence that the assessment process is influencing instruction. In this case the process is working as intended by increasing the alignment between the assessment tasks and the student's IEP. In a similar manner, survey responses may shed light on some unintended, negative consequences of the ALT-MSA that can be addressed before the next administration.

Criterion-Related Validity Evidence

Although the primary evidence for the validity of the ALT-MSA lies in the process used by the MSDE to develop and design the assessment, it is also informative to collect criterion-related validity evidence. The term criterion-related validity refers to the degree to which a test correlates with one or more outcome criteria. The key is the degree of relationship between the assessment items or tasks and the outcome criteria. To help ensure a good relationship between the assessment and the criterion, the criterion should be relevant to the assessment and it should also be reliable.

For each student portfolio submitted for scoring, readers review the contents for evidence of "important components of an instructional program" or positive practices (see Section 3.3). Students receive a score of 1 on a positive practice if there is evidence of that practice in their portfolio and a score of 0 if there is not. It is suggested that scores on the ALT-MSA will be strengthened if these components are present in the student's instructional program. Consequently, a positive relationship between student mastery percentage and positive practice scores is expected. Table 19 in Appendix A provides the correlation between student mastery percentage scores and positive practice scores for the current administration. A correlation reflects agreement between relative standing on one variable and relative standing on the other. A significant correlation means that the correlation coefficient is statistically different from zero. For the ALT-MSA, positive correlations suggest that the presence of a positive practice indicator is related to higher mastery percentage scores.

When reviewing this data it is important to note that *only one of the at least two* scorers scoring any given portfolio assigns positive practice scores. Reader agreement scores for positive practice indicators are not available. Therefore, the reliability or consistency with which scorers can assign these scores is unknown. Similarly it is important to remember that positive practice scores indicate the extent to which indicators of positive practice are *observable* in the submitted portfolio. It is quite possible that best practice was followed, but the *indicators are not outwardly apparent or identifiable in the*

submitted portfolio materials. These factors suggest the correlations be interpreted with caution until the reliability and validity of the positive practice scores can be verified.

The extent to which the issues described above are influencing the resulting correlations is currently unknown. However, if they are having an effect it is likely that the values reported in Table 19 are attenuated. This should be taken into account when comparing the degree of the correlations relative to expectations