Mathematics Applications Inventory
Results of Preliminary Pilot Study

November 17, 2009

Howard T. Everson
Center for Advanced Study
City University of New York

The statistical summaries reported here are intended to provide guidance to the developers of the Mathematics Applications Inventory (MAI) scale. The central purpose of these analyses is to improve the MAI by revising, modifying, or removing MAI scale items that may not be functioning optimally, and to increase our understanding of how this early, Alpha version of the MAI is working (i.e., is there preliminary evidence that the MAI is a valid and reliable measurement scale). Another important aspect of the MAI item analyses is related specifically to whether the test items can provide important diagnostic information on what the examinees may have learned or not with respect to the applications of advanced mathematical methods and procedures. To achieve this goal a variety of group statistics were computed for each of the MAI items, as well as for the MAI summary scores. Before we begin, however, there is one major caveat, the sample size, fifty-one students, is much too small to assure the stability of the sample and test item statistics. Much larger samples will be needed in the future to fully refine and develop the MAI.

With this in mind, we begin by reviewing a number of summary statistics for the eleven (11) items that comprise the MAI scale, all of which were computed using item-level response data from the entire pilot sample (51) of students who completed the MAI scale. (Note: Although the analysis file contains additional information on seventy-five (75) Cornell University students who participated in this early phase of the study, only a subset of them, roughly two-thirds, completed the MAI scale.) To start, we look at the mean scores by item because they provide key indices of the MAI items' difficulty for this particular sample of students. Item difficulty is most commonly measured by calculating the percentage of test-takers who answer the item correctly [(*p*-value for an item = (# of people responding correctly) / (# of people taking the test)]. In the

case of the MAI, however, there are a number of items that were awarded partial credit, so the MAI has item scores range from 0 (incorrect) to 0.5 (partially correct) to 1.0 (fully correct).

In general, items with mid-range *p*-values (between 0.40 and 0.60) produce total test scores with the most variation. Thus, most test developers seek to develop tests where the average item difficulty score is about 0.5. Again, speaking generally, test developers often discard items with difficulty levels (*p*-values) between 0 - 0.2 and 0.8 - 1.0 because they are either too difficult or too easy, and do not effectively differentiate among the examinees. Table 1, below, presents the results of the descriptive analyses of the students' performance on each of the eleven MAI items. In this sample, MAI item 1 appears to be the easiest ($p = .85$) , while MAI item 3c looks to be the most difficult ($p = .18$). The remaining MAI items exhibit item difficulty values ranging from 0.36 to 0.81.

Table 1. Item Level Statistics

	N	Minimum	Maximum	*P* values	Std. Deviation
MAI 1	51	.00	1.00	.8529	.33607
MAI 2a	51	.00	1.00	.6863	.46862
MAI 2b	51	.00	1.00	.6275	.48829
MAI 2c	51	.00	1.00	.4412	.49646
MAI 3a	51	.00	1.00	.6275	.46737
MAI 3b	51	.00	1.00	.4216	.47300
MAI 3c	51	.00	1.00	.1765	.34385
MAI 4a	51	.00	1.00	.6471	.48264
MAI 4b	51	.00	1.00	.8137	.38679
MAI 5a	51	.00	1.00	.4314	.45847
MAI 5b	51	.00	1.00	.3627	.45911
Valid N (listwise)	51				

Tables 2a and 2b present data on the item difficulties for students enrolled in courses #1910 and #1920. On average, and as expected, the 24 students in Course #1920 performed at a higher level on the MAI than the 27 students in the #1910 course.

Table 2a. Item-Level Statistics: Course 1910

	N	Minimum	Maximum	*P* Values	Std. Deviation
MAI 1	27	.00	1.00	.8148	.37076
MAI 2a	27	.00	1.00	.6667	.48038
MAI 2b	27	.00	1.00	.6667	.48038
MAI 2c	27	.00	1.00	.3519	.47666
MAI 3a	27	.00	1.00	.5926	.48113
MAI 3b	27	.00	1.00	.4630	.45838
MAI 3c	27	.00	1.00	.0926	.24167
MAI 4a	27	.00	1.00	.5556	.50637
MAI 4b	27	.00	1.00	.7222	.44578
MAI 5a	27	.00	1.00	.3148	.41944
MAI 5b	27	.00	1.00	.2593	.40121
Valid N (listwise)	27				

Table 2b. Item-Level Statistics: Course 1920

	N	Minimum	Maximum	*P* Values	Std. Deviation
MAI 1	24	.00	1.00	.8958	.29411
MAI 2a	24	.00	1.00	.7083	.46431
MAI 2b	24	.00	1.00	.5833	.50361
MAI 2c	24	.00	1.00	.5417	.50898
MAI 3a	24	.00	1.00	.6667	.45842
MAI 3b	24	.00	1.00	.3750	.49454
MAI 3c	24	.00	1.00	.2708	.41649
MAI 4a	24	.00	1.00	.7500	.44233
MAI 4b	24	.00	1.00	.9167	.28233
MAI 5a	24	.00	1.00	.5625	.47348
MAI 5b	24	.00	1.00	.4792	.49955
Valid N (listwise)	24				

In both subsets of examinees MAI item 3 was the most difficult (p = .09 and .27, respectively). For the students in the 1920 course, both MAI item 1 and item 4b were relatively easy. The remaining MAI items exhibit item difficulty values ranging from 0.36 to 0.81. Again, reviewing these p-values for each item allows us to determine if items are being answered in the same direction by all the examinees regardless of background knowledge. However, although p-values provide some information about the difficulty of a test item, they tell us very little about any one item's usefulness in measuring the test's construct, in the case of the MAI we are interested in students' ability to apply mathematics to solve problems.

MAI Scale Reliability Estimates

When building a scale such as the MAI a principal concern is whether the measurement method is reliable, since without reliability results from the MAI would not be replicable, and replicability is central to measuring change in the underlying construct in a pre-post instructional framework. In an abstract sense, reliability refers to the correlation of an item, a total scale score with a hypothetical measure (sometimes referred to as a "true score") of the underlying construct. One can imagine a thought experiment in which the MAI was administered to the same student many, many times with the resulting mean of those scores as representing the "true score". Because we cannot base our estimates on such an approach, we have to estimate reliability using another method. In this instance, particularly given the small sample of examinees, we will use *Cronbach's alpha* (a measure of internal consistency of measurement based on the inter-correlations of the eleven items comprising the MAI scale). *Cronbach's alpha* equals zero when the true score is not measured at all and there is only an error component captured by the measurement. Conversely, α =1.0 when all the items measure the underlying true score without error.

Our estimate of reliability based on this method yielded an α = .77, which is a bit low, but not alarmingly so given the brevity of an 11 item scale like the MAI. In general, we look for scale reliability estimates in the range of .85 or above, and one way to improve a scale's reliability is to add more items measuring the targeted underlying (latent) construct. Thus, the somewhat suppressed αlevels may have occurred for a number of reasons, including too few items (only 11

in this case), and also because a number of the MAI items are not discrete or unique and are actually nested components or multipart items.

Inter-Item Correlations.

The inter-item correlation matrix is another important aspect of test analysis. The correlations in Table 4, below, provide important information about the MAI's internalconsistency, and what could be done to improve it. In a reliable scale, all items should correlate with the total. A low item-correlation provides empirical evidence that the item is not measuring the same construct measured by the other scale items. For example, a correlation less than 0.3 indicates that the corresponding item does not correlate very well with the scale overall and, thus, it may be dropped. Ideally each MAI item should be correlated highly with the other items measuring the same construct (mathematical application ability). MAI items not correlating strongly with the other items measuring the same construct can often be dropped without reducing the test's reliability. The correlations in Table 4 provide asense of the pattern of correlations among the eleven MAI items.

Table 4. Inter-Item Correlation Matrix

	MAI 1	MAI 2a	MAI 2b	MAI 2c	MAI 3a	MAI 3b	MAI 3c	MAI 4a
MAI 1	1.000	.082	.147	.217	.090	.083	.056	.228
MAI 2a	.082	1.000	.615	.220	.232	.248	.350	.120
MAI 2b	.147	.615	1.000	.155	.168	.261	.280	.110
MAI 2c	.217	.220	.155	1.000	.378	.214	.296	.037
MAI 3a	.090	.232	.168	.378	1.000	-.044	.293	.159
MAI 3b	.083	.248	.261	.214	-.044	1.000	.241	.008
MAI 3c	.056	.350	.280	.296	.293	.241	1.000	.202
MAI 4a	.228	.120	.110	.037	.159	.008	.202	1.000
MAI 4b	.324	.113	.472	.228	.162	.001	.252	.176
MAI 5a	.193	.363	.241	.465	.252	.390	.459	.250
MAI 5b	.126	.214	.169	.249	.270	.364	.410	.228

Table 4. Inter-Item Correlation Matrix

(Continued)

	MAI 4b	MAI 5a	MAI 5b
MAI 1	.324	.193	.126
MAI 2a	.113	.363	.214
MAI 2b	.472	.241	.169
MAI 2c	.228	.465	.249
MAI 3a	.162	.252	.270
MAI 3b	.001	.390	.364
MAI 3c	.252	.459	.410
MAI 4a	.176	.250	.228
MAI 4b	1.000	.293	.219
MAI 5a	.293	1.000	.691
MAI 5b	.219	.691	1.000

This overallpattern of correlations suggests that performance on the unique items, i.e., item MAI 1, MAI 2, MAI 3, MAI 4 and MAI 5 are not highly correlated—indicating that students who got, say item 1 correct (or incorrect) did not perform similarly on the other items on the MAI. The within item responses, e.g., item MAI 5a and 5b were more strongly correlated (r = .69) or MAI 2a and 2b (r = .61).

Other Item Characteristics

Table 5 presents a number of "item-total"statistics. The first column indicates the changes in the overall MAI scale mean scores, if an were deleted, giving us a sense of how much any particular item is contributing to the distribution of the scale scores. Similarly, column two shows what would happen to the variance in the MAI scores, if an item were dropped from the scale. The item-total correlation, column 3, is the correlation between an item score and the total MAI score. We expect, for example, that if an examinee gets a question correct she should, in general, have higher overall MAI score than an examinee who got that same question wrong. A low item-total correlation indicates that performance on that particular item is not related to a students'

overall performance on the MAI scale (for example., < .3 for large samples or not significant for small samples) and the researchers may want to consider dropping it.

Table 5. Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
MAI 1	5.2353	6.704	.270	.174	.768
MAI 2a	5.4020	5.990	.470	.519	.747
MAI 2b	5.4608	5.938	.467	.584	.748
MAI 2c	5.6471	5.983	.435	.355	.752
MAI 3a	5.4608	6.258	.346	.265	.763
MAI 3b	5.6667	6.317	.314	.306	.767
MAI 3c	5.9118	6.267	.524	.317	.745
MAI 4a	5.4412	6.436	.252	.147	.775
MAI 4b	5.2745	6.373	.391	.442	.757
MAI 5a	5.6569	5.615	.673	.635	.721
MAI 5b	5.7255	5.883	.536	.534	.739

The squared multiple correlation, R^2 , is the amount of variance in the item predicted from the performances on all other items in the scale. The larger the R^2, the more the item is contributing to internal consistency. The lower the R^2, the more the test developers ought to consider dropping the item. We should note, too, that the R^2 of some items may be low even on a scale which has an acceptable level of reliability. Finally, in column 5 we have the "Cronbach's Alpha if Item Deleted" estimates, which tell us the value of *alpha* if the given item were removed from the scale. That is, the 'scale if deleted' option lets the researcher assess the reliability of each item. Often test developers may want to drop items where the alpha if deleted is higher than the overall alpha as another way to improve thereliability of the scale. However, it may be the case that even when an item has high random error, it is, in fact, measuring the construct of interest. (We can take up the issue of how best to determine if an item is contributing to the overall measurement of the targeted construct at a later time, when we have larger sample sizes.)

Item Discrimination Indices

To measure how well a test item separates those test takers who show a high degree of mathematical application ability from those who do not, we often calculate a discrimination index (D). In general, this index compares, for each test item, the performance of those who scored the high on the test (U – upper group, those above the mean score of 6) with those who scored the worst (L – lower group, those with scores below the group mean of 6). The logic of the D statistic is simple. Tests are more difficult for those who score poorly (lower group). If an item is measuring the same thing as a test, then the item should be more difficult for the lower group. The D statistic provides a measure of each item's discriminating power with respect to the upper and lower groups. Tables 6a and 6b provide descriptive statistics for both the low and high scoring students.

Table 6a. Descriptive Statisticsfor Lower Scoring Students

	N	Minimum	Maximum	Mean	Std. Deviation
MAI Total	29	.00	6.00	4.1379	1.50533
Valid N (listwise)	29				

Table 6b. Descriptive Statisticsfor Higher Scoring Students

	N	Minimum	Maximum	Mean	Std. Deviation
MAI Total	22	6.50	11.00	8.6591	1.46699
Valid N (listwise)	22				

Tables 7a and 7b provide the item-level *P* values for both the low and high scoring students.

Table 7a *P* Values for Lower Scoring Students

	N	Minimum	Maximum	*P* Values	Std. Deviation
MAI 1	29	.00	1.00	.8103	.38762
MAI 2a	29	.00	1.00	.4483	.50612
MAI 2b	29	.00	1.00	.4138	.50123
MAI 2c	29	.00	1.00	.2241	.41374
MAI 3a	29	.00	1.00	.4483	.48816
MAI 3b	29	.00	1.00	.2414	.41449
MAI 3c	29	.00	.50	.0345	.12894
MAI 4a	29	.00	1.00	.5172	.50855
MAI 4b	29	.00	1.00	.7069	.45350
MAI 5a	29	.00	1.00	.1379	.26378
MAI 5b	29	.00	1.00	.1552	.33014
Valid N (listwise)	29				

Table 7b. *P* Values for Higher Scoring Students

	N	Minimum	Maximum	*P* Values	Std. Deviation
MAI 1	22	.00	1.00	.9091	.25054
MAI 2a	22	1.00	1.00	1.0000	.00000
MAI 2b	22	.00	1.00	.9091	.29424
MAI 2c	22	.00	1.00	.7273	.45584
MAI 3a	22	.00	1.00	.8636	.31554
MAI 3b	22	.00	1.00	.6591	.44685
MAI 3c	22	.00	1.00	.3636	.44137
MAI 4a	22	.00	1.00	.8182	.39477
MAI 4b	22	.00	1.00	.9545	.21320
MAI 5a	22	.00	1.00	.8182	.36337
MAI 5b	22	.00	1.00	.6364	.46756
Valid N (listwise)	22				

Next Steps

In subsequent discussions we ought to consider how best to enlarge the sample, and the possibility of conducting a "distractor analysis" which would look at the pattern of right and wrong answers on those items that were structured as multiple-choice type items. We also ought to discuss how best to make use of the other demographic and achievement data in the students' files, e.g., in larger samples should we consider conditioning our analyses on the scores on the attitudinal scales).