2. FIELD-TESTS AND STATISTICAL ANALYSIS RESULTS

2.1. Assessment development. We recently completed the first phase of assessment development, which involved 150 teachers over 4 years. We have gone through several rounds of the iterative process of design, pilot test, data analysis, and revision of the instrument. We also conducted problem-solving think-aloud interviews with teachers, educators, and mathematicians, to see if their interpretations of the items match our intended objectives.

Our focus throughout has been on creating assessment items that meet our goal of measuring how teachers choose to use their MHoM when doing mathematics in familiar contexts. Thus, items whose content was beyond the grasp of most teachers—and hence did not permit them to demonstrate the use of their mathematical habits—have been excluded. Likewise, we have discarded items that did not yield much variety in how teachers approached them, possibly due to the lack of multiple entry points. We have also learned that situating items in classroom settings (e.g., "If you are the teacher, what feedback would you provide to this student?") can be a distraction, since teachers tended to focus on pedagogical issues in their responses, as opposed to providing a mathematical solution. And we have considered assessment logistics, such as its format, length, and guidelines. For instance, multiple-choice items have been discarded, since we deemed them inadequate for revealing teachers' mathematical thinking.

The current paper and pencil assessment contains 10 items and is designed to be completed in one hour. Each item is in free response format, such as the *Maximum Value* problem described in the previous section. The teachers taking the assessment are instructed to write down their thoughts and approaches, even if they are unable to solve a problem completely.

Our assessment development is not yet complete, and we will continue to update the instrument as we obtain more data. In section 2.3, we will explain how our statistical analysis results have led us to replace or revise three of the P&P items.

2.2. Research methods. Below, we will describe our teacher sample and data collection methods, rubric development and coding process, and statistical analysis conducted. The results of our statistical analysis will be discussed in section 2.3.

2.2.1. Teacher sample and data collection. We administered the P&P assessment to 59 secondary teachers, involved in three different *mathematics content-based* professional development programs. Amongst them, 43 teach high school, 13 teach middle school, and 3 play other roles in secondary education (e.g., math coach, district leader, etc.). Here are brief descriptions of the programs:

- **Program A** provides opportunities for teachers to effectively implement the Common Core Standards for Mathematical Practice. The teachers focus on mathematical topics where these practices can bring insights and coherence to school mathematics.
- **Program B** is intended to increase secondary mathematics or science teachers' knowledge and command of their subject matter and curriculum, as well as prepare them for roles as lead teachers, mentors, and coaches.
- **Program C** is a summer immersion in mathematics that engages secondary teachers in "experiencing mathematics as mathematicians do." The teachers also attend workshops designed to help them unpack the pedagogical approaches used in the immersion experience.

Our sample of 59 teachers includes 20 from Program A, 23 from Program B, and 16 from Program C. These individuals vary greatly in their teaching experiences and in the settings in which they teach.

For each group, the P&P assessment was administered at one of their professional development meetings, in accordance with our guidelines. Furthermore, the teachers completed a background survey that collected information such as their educational background, their teaching experience, and the grade levels and type of courses they teach.

Also, teachers in Program A took a slightly different version of the assessment—their version

contained 12 items, out of which only 8 overlapped with the version taken by teachers in Programs B and C. Thus for Program A teachers, only the overlapping 8 items were included our dataset; consequently, we deemed their data incomplete and excluded them from all statistical analyses. But all 59 teachers' data were used for rubric development.

2.2.2. Rubric development and coding process. For each item on the P&P assessment, we saw variations in the responses by our teacher population. We sought and found recurring regularities and themes that revealed themselves in the data (Guba 1978). The responses to each item were categorized according to the approaches that teachers took, but *not* according to whether they obtained the correct answer. These approaches formed a basis for the codes in our rubric. Moreover, since our main focus is the way in which teachers approach each item, we decided that a response would be credited with a code even if it were incomplete or contained minor calculation errors, as long as the intent of the teacher to use a particular approach was clear.

To complete the rubric development, we asked two research mathematicians to assign a numerical value to each code.[1] Specifically, they rated the codes on a scale of 1 to 10 based on alignment with the mathematical habit that each item is intended to measure, with 1 being low alignment and 10 being high. In the *Maximum Value* item, for example, the two mathematicians were asked to study the codes SQUR, TRNS, SYMM, CALC, and PNTS, and score each on a scale of 1 to 10 based on the degree to which the approach exhibited the use of the habit **SUDS2.** *Making use of structure to solve problems*. With the input from these mathematicians, we derived a numerical value for each code.

Using the rubric developed for each item, two of our researchers independently coded all responses.[2] Shown in Table 1 is the Kappa statistic for each item, measuring the level of agreement of the two raters. For all except the `MXB` item, the Kappa value was greater than 0.80, indicating very strong agreement (Hallgren 2012).

Table 1 Kappa value for each assessment item

Item	`MAX`	`SOS`	`TNP`	`PWB`	`THP`	`FFX`	`ELZ`	`SOL`	`EQU`	`MXB`
Kappa	0.832	0.858	0.853	0.874	0.920	0.834	0.820	0.810	0.845	0.363

Then the two researchers shared how they coded each response to sort out any discrepancies they had. And after this discussion, they were in full agreement. The rubrics were also revised based on this conversation—some new codes were added, and clarifications were made on existing codes where needed.

2.2.3. Statistical analysis conducted. Using data from the 39 teachers who completed the full 10-item version of the P&P assessment (i.e., those in Programs B and C), we analyzed the psychometric properties of the instrument. Specifically, we checked for construct validity of the assessment, tested its internal consistency reliability, and performed exploratory analysis of item discrimination. We also examined the relationship between teachers' professional background and their performance on the assessment.

We tested the construct validity of the P&P assessment in three steps, although these are iterative according to the formative needs of creating a validated instrument. The first step was operationalization and initial ordination of target constructs, i.e., the mathematical habits that the assessment items intend to measure. The second step employed principal components analysis (PCA) to determine convergent and discriminant validity of the assessment items. This process identified how well clusters of items converge around the intended MHoM construct and remain distinct from other constructs. In the third step, we

[1] This step is necessary in order to conduct statistical analysis on our data. In our large scale field-testing (described in section REF), we will employ 200 mathematicians for this process.

[2] When a response employed multiple approaches, the one with the highest numerical value was counted towards the teacher's assessment score.

compared the components that were found in the data with the MHoM constructs as intended in the assessment design. We acknowledge that these are preliminary steps—for a more complete validation of the instrument, we would use a larger sample and follow up with structural equation modeling for confirmatory factor analysis.

Our goal for reliability is that the instrument will produce acceptably consistent results with different respondents by seeing how closely the data resemble a hypothesized normal distribution. The standard measure for this is Cronbach's Alpha analysis, conducted with groups of varying composition (DeVellis 2003). We recognize that Cronbach's Alpha may be underestimating results, and thus we also used Guttman Lambda 6 to ensure a maximally accurate calculation of instrument reliability.

We also studied item discrimination by analyzing the relationship between how teachers did on each item and their overall assessment performance. We divided our sample into "high" and "low" groups, according to whether their total assessment score was above or below the median. Then for each item, we computed the median score on that item for both groups. (Ideally, we would see higher median score for the "high" group than for the "low" group.) Also for each item, we conducted the Wilcoxon rank-sum test to see whether the scores on that item for the two groups are statistically significantly different. We remark that the above process is exploratory, and that to more accurately measure item discrimination, we would have to employ logistic regression analysis with a larger sample.

The independent variables for teachers' background were collected using the survey that accompanied the P&P assessment. They included the following:

- Grade level taught (middle school or high school)
- Current teaching assignment
- Level of involvement in professional development
- Type of undergraduate degree
- Undergraduate major
- Type of graduate degree
- Graduate area of concentration
- Number of years of teaching
- Number of years at the current school

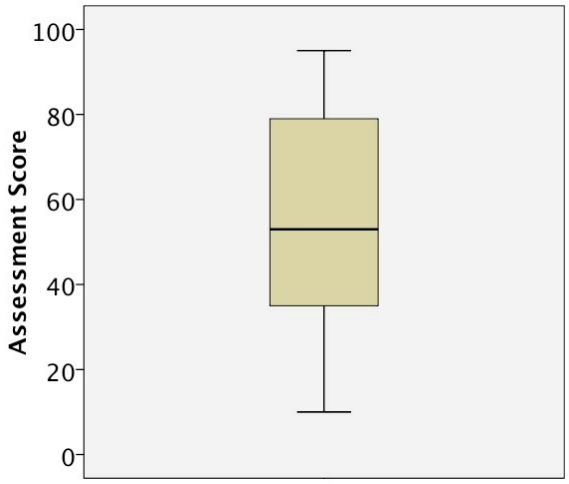
For each variable, we considered the groups in question (e.g., high, medium, and low involvement in professional development programs) and performed the Kruskal-Wallis test to compare their assessment scores and determine whether the difference between the groups are statistically significant.[3]

2.3. Statistical analysis results. Each item on the 10-item assessment is scored out of 10 points, and so the maximum possible score on the assessment is 100 points. The box plot in Fig. 1 summarizes the distribution of assessment scores in our dataset $(n = 39)$.[4] Descriptively, the plot shows a normal distribution with nearly full range of scores, and without any outliers. This suggests that the instrument can be reliably used across a wide range of teachers.

Fig. 1 Box plot of assessment scores

[3] The Wilcoxon rank-sum test was used when there were only two groups.

[4] The mean score was 54.85, with a standard deviation of 26.52. The median score was 53, and the minimum and maximum scores were 10 and 95.



2.3.1. Construct validity results. We conducted principal component analysis to verify the construct validity of the P&P assessment.[5] After examining the scree plot and considering the number of non-trivial components with a cut-off value of 0.45, we decided to include three components in the analysis. Moreover, our assessment was intended to primarily measure three of our habits—SUDS2, LANG1, and LANG2—providing an *a priori criterion* for our 3-component model.

Then we performed PCA to see what kinds of correlational clusters appear in the data.[6] Table 2 shows the rotated solution with a cut-off value of 0.45. We sorted the items based on the habits they intended to measure. Note that the `THP` item appears under SUDS1 and LANG1, since it intended to measure both habits. We also named the components `S2/L1`, `L2`, and `Diff`—rationales for these names, particularly for Component `Diff`, are discussed below.

Table 2 PCA results

[5] To determine whether or not our dataset is amenable to conducting PCA, we ran the KMO and Bartlett tests. The former test generated a value of 0.836 and the latter was significant. These findings supported the conclusion that our dataset is, indeed, amenable to PCA.

[6] Rotation method: Varimax with Kaiser normalization. Rotation converged in 5 iterations.

Habits/Items		Component		
		S2/L1	L2	Diff
SUDS1:	THP	.756		
SUDS2:	MAX	.828		
	SOS	.784		
	TNP	.671		
	PWB	.499	.488	.550
LANG1:	THP	.756		
	FFX	.842		
	ELZ			.896
LANG2:	SOL		.722	
	EQU	.488	.524	
	MXB		.838	

Next, we compared the observed clusters (i.e., the PCA results) with the intended MHoM constructs to see how well they resemble each other. The SUDS2 items group together in Component S2/L1, indicating strong convergent validity. But LANG1 items (THP and FFX) also appear in the same component, suggesting a possible discriminant validity issue. The three LANG2 items group together solely in Component L2, showing both convergent and discriminant validity for these items.

We conjecture that PWB and ELZ clustered together in Component Diff (i.e., "Difficult") due to many teachers' unfamiliarity with the content of each of these items, and those who performed poorly on one item tended to do poorly on the other. Not surprisingly, these were the two items with the lowest median scores across all teachers—with 3 points (out of 10) for PWB and 4 points for ELZ. Let us consider the PWB item, which asks:

> Imagine you have a number *b*, not equal to 1, such that $b^7 = 1$. Find the smallest whole number *n* (i.e., $n \geq 0$) such that $b^{374} = b^n$.

This item measures the habit **SUDS2.** *Making use of structure to solve problems*—specifically, the habit of using the structure of periodicity.[7] However, many teachers could not imagine a number *b* with such a property, likely due to their insufficient experience with complex numbers. They claimed that $b^7 = 1$ must imply that $b = 1$, or they did not write down any meaningful mathematical work. Thus, we were not putting these teachers in a situation where they could appropriately demonstrate the use of their mathematical habits. In future versions of the assessment, the PWB item will be replaced by another item that measures teachers' use of periodicity, but without the mathematical content getting in the way of their MHoM use.[8]

The PCA results will guide further revision of the P&P assessment so that we can better zero in on the target MHoM construct(s) for each item. The results will also help us think about the theoretical basis

[7] Here is an approach that shows strong alignment with SUDS2: Write $374 = 7 \cdot 53 + 3$. Then use rules of exponentiation to obtain $b^{374} = (b^7)^{53} b^3 = 1^{53} b^3 = b^3$. Thus, $n = 3$ is the desired value of *n*.

[8] The ELZ item had a similar issue about content. Also, many teachers misunderstood what was being asked, which led them to respond to a completely different question. In the future, this item will be revised to add more clarity to the problem statement.

for the MHoM constructs. For instance, we will ask ourselves, "Are there other possible ways of delineating these mathematical habits, and what are the advantages and disadvantages of each?"

2.3.2. Reliability results. For reliability testing, we found Cronbach's Alpha value of 0.87 and Guttman Lambda 6 value of 0.90. Both are well above the generally accepted threshold of 0.7 for a reasonably reliable instrument (Nunnally 1978).

2.3.3. Item discrimination results. We performed an exploratory analysis of how teachers' item-by-item performance compared with their total assessment score. Shown in Table 3 are the per-item median scores for "high" and "low" groups, as well as the *p*-values of the Wilcoxon rank-sum test comparing the two groups.[9] The items have been sorted in the table according to the habits that they intended to measure. Note again that the `THP` item appears under both SUDS1 and LANG1.

Table 3 Item discrimination results

Habits/Items		**Low**	**High**	***p*-value**
SUDS1:	`THP`	3	10	.000
SUDS2:	`MAX`	0	8	.000
	`SOS`	6	10	.000
	`TNP`	2	10	.000
	`PWB`	0	10	.000
LANG1:	`THP`	3	10	.000
	`FFX`	0	7.5	.000
	`ELZ`	4	7	.028
LANG2:	`SOL`	7	10	.005
	`EQU`	2	10	.000
	`MXB`	2	6	.190

The "high" group outperformed the "low" group on every item, with statistical significance ($p < .05$) for all but the `MXB` item. This suggests that each item (perhaps with the exception of `MXB`) is able to discriminate between groups of teachers of high and low overall performance, as measured by this assessment. There is also a wide range of scores in the items, which allows the instrument to assess teachers with a variety of levels of MHoM.

Given its poor item discrimination, as well as its low Kappa value for inter-rater reliability (see section 2.2.2), we will replace the `MXB` item with another in future versions of the assessment.

2.3.4. Teacher background and assessment performance. Using the Kruskal-Wallis test, we studied the relationship between teachers' professional background and their total score on the P&P assessment. Of our nine independent variables described in section 2.2.3, three of them—grade level taught, current teaching assignment, and type of undergraduate degree—contained groupings with statistically significant

[9] Recall that the teachers were divided into "high" and "low" groups, according to whether their total assessment score was above or below the median.

($p < .05$) difference.[10] For each of these three variables, Table 4 shows (1) the size of each group, (2) the median, minimum, and maximum assessment scores for each group, and (3) the *p*-value of the Kruskal-Wallis test (or the Wilcoxon rank-sum test when there are two groups) comparing the groups.

Table 4 Relationship between teacher background and assessment score

Grade level taught[a]	**Frequency (%)**	**Median (min, max)**	***p*-value**
Middle school	13 (36%)	36 (11, 90)	.008
High school	23 (64%)	70 (11, 95)	
# of content areas	**Frequency (%)**	**Median (min, max)**	***p*-value**
1	14 (39%)	39.5 (11, 73)	.014
2 or more	22 (61%)	73.5 (11, 95)	
Undergrad degree[b]	**Frequency (%)**	**Median (min, max)**	***p*-value**
BS	28 (74%)	53.5 (10, 95)	.023
BA	5 (13%)	89 (47, 95)	
B.Ed.	5 (13%)	36 (19, 51)	

[a] For this variable, as well as for the number of content areas, 3 individuals in our sample who are not currently classroom teachers were excluded from our analysis.
[b] There was a fourth group ("Other") containing just one teacher, which was excluded from our analysis.

For the undergraduate degree variable, the Kruskal-Wallis test yielded $p = .023$, indicating that there was a statistically significant difference between the three groups. We conducted post-hoc analysis (using pairwise Wilcoxon rank-sum test with Bonferroni adjustment) to determine where the difference lies and found that the BA group outperformed the B.Ed. group with statistical significance. However, given the particularly small size of these two groups (with $n = 5$ each), we must take this result as preliminary at best and revisit the issue in the future with a larger sample.

The background survey item for the teaching assignment variable asked the teachers to list the classes that they are currently teaching. We were concerned with variety in content areas, but not whether they taught different levels/tracks of the same course. Thus, for example, a teacher who listed "Algebra 2 Honors" and "Algebra 2 Non-Honors" was coded as teaching one content area, namely Algebra 2.[11] For each teacher, we found the number of distinct content areas taught. Then we compared the assessment scores of two groups—those teaching one content area and those teaching 2 or more content areas—and found that the "2 or more" group obtained higher scores with statistical significance ($p = .014$). Moreover, analysis of hierarchical clustering using the dendrogram shown in Appendix B indicates that the sample is fairly even in its variance, without any clusters of teachers that stand out in terms of the relationship between variety of teaching assignments and assessment performance. This confirms and extends what we had observed in the reliability values, that the instrument demonstrates strong reliability and works in a consistent manner.

[10] For the undergraduate major variable, the Kruskal-Wallis test yielded $p = .040$. However, post-hoc analysis (using pairwise Wilcoxon rank-sum test with Bonferroni adjustment) did *not* reveal significant differences between any pair of groups.

[11] However, we did consider Algebra 1 to be a *different* content area from Algebra 2.

2.4. Statistical analysis summary. The data and statistical analysis reflect our progress in developing a valid and reliable instrument designed to measure teachers' mathematical habits of mind. Indeed, the validity and reliability values reported above are encouraging. However, a larger sample is required (and is planned, as described in section REF.) before we can make any definitive assertions about our instrument.

The teacher background data revealed that those with more variety in their teaching assignments (i.e., 2 or more content areas) performed better on the assessment. This finding, while intriguing, will need to be revisited with a larger sample before any inferences can be made. It may be that the demands of teachers having to shift between content areas engender a kind of cognitive nimbleness that translates well into performance on the assessment. It could just as well be that these teachers are the most accomplished ones in their schools, and thus they are the "go-to" teacher when there is a need to teach an unanticipated section of other content areas.

A possible confounding factor might be the unbalanced distribution of middle school and high school teachers in the groups for the number of content areas variable, as shown in Table 5. Note how the "2 or more" group contains a much higher percentage of high school teachers, who tended to earn higher assessment scores than the middle school teachers in our sample.

Table 5 Distribution of middle school and high school teachers

# of content areas	# of MS teachers	# of HS teachers
1	9	5
2 or more	4	18

We remark that in the background survey, the middle school teachers who teach 1 content area reported that they teach exclusively Pre-algebra or "Middle school math." To determine the degree to which each variable—grade level taught and teaching assignment—is a predictor of the assessment score, we would need a larger sample and conduct multiple regression analysis.

Moreover, the findings of non-significant relations between the other independent variables and assessment performance will be revisited as the sample size grows.