Problem solving, Domain Expertise and Learning: Ground-truth Performance Results for Math Data Corpus
Sharon Oviatt
Incaa Designs
11140 Wing Point Drive, N.E., Bainbridge Island, Wa. 98110
oviatt@incaadesigns.org; adrienne.o.cohen@gmail.com

Abstract
Problem solving, domain expertise, and learning are analyzed for the Math Data Corpus, which involves multimodal data on collaborating student groups as they solve math problems together across sessions. Compared with non-expert students, domain experts contributed more group solutions, solved more problems correctly and took less time. These differences between experts and non-experts were accentuated on harder problems. A cumulative expertise metric validated that expert and non-expert students represented distinct non-overlapping populations, a finding that replicated across sessions. Group performance also improved 9.4% across sessions, due mainly to learning by expert students. These findings satisfy ground-truth conditions for developing prediction techniques that aim to identify expertise based on multimodal communication and behavior patterns. Regarding prediction, the dominant expert in a group was identifiable 100% of the time by the seventh problem during a session simply by knowing who contributed the most group solutions. This finding provides an initial benchmark result, and also a model for evaluating the speed and reliability of prediction success. Together with the Math Data Corpus, these results contribute valuable resources for supporting data-driven grand challenges on multimodal learning analytics, which aim to develop new techniques for predicting expertise early, reliably, and objectively.

Introduction
Multimodal learning analytics (MMLA) is an emerging field within learning analytics [2] that analyzes natural communication modalities [4, 5, 6]. It tracks rich multimodal information sources (e.g., speech, video, digital pen) to examine how learning occurs and how to best support it with new curricula, digital tools, teacher training, and other resources. Since it involves multiple natural communication modalities and is not limited to click-stream analysis of keyboard-based computer interactions, it can be applied to both interpersonal and computer-mediated learning activities [5, 6]. In addition, the dominant educational technology platform worldwide is mobile cell phones, with adoption continuing to accelerate in developing regions. Their input capabilities include speech, touch, virtual keyboard, and in some cases sensors and pen input, with interfaces becoming increasingly multimodal [4]. To evaluate learning on these devices, new multimodal learning analytics techniques will be required. In comparison, click-stream analyses will have limited utility [5].

MMLA techniques can support a systems-level view that is grounded in converging data on how students communicate, collaborate, and solve problems during learning. In addition, they can examine behaviors supportive of learning at different stages, including indices of social/emotional and motivational state. They also can guide the design of promising next-generation educational technologies (e.g., pen, multimodal, tangible) that are not limited by keyboards, and deepen our understanding of any tradeoffs associated with different learning interventions or digital learning tools.

Virtually all assessments of student learning rely critically on the availability of accurate metrics as forcing functions. Multimodal learning analytics is expected to yield a large array of more sensitive, reliable, objective, and richly contextualized metrics of learning-oriented behaviors and expertise. This includes ones that could be analyzed unobtrusively, continuously, automatically, in natural field settings, and more affordably than traditional educational assessments. Progress in MMLA will transform our ability to accurately identify and stimulate effective learning, support more rapid feedback and responsive interventions, and facilitate learning in a more diverse range of students and contexts. Long-term advances in multimodal learning analytics are expected to contribute new empirical findings, theories, methods, and metrics for understanding how students learn.
One aim of multimodal learning analytics is to identify domain expertise and consolidation of expertise rapidly, reliably, and objectively, and also to identify critical learning-oriented precursors. To support the development of relevant new techniques, multimodal data resources such as the Math Data Corpus have become publicly available to participants in data-driven grand challenge events [for details, see 5]. These resources include coding of communication and behavioral patterns relevant to learning. This paper provides a performance analysis of students’ collaborative math problem solving, domain expertise, and learning in the Math Data Corpus at both the individual and group levels, and also across longitudinal sessions. This includes students’ initiative, accuracy, and time required to solve math problems that vary in difficulty from easy to very hard. It also includes assessment of change in performance indicative of learning across sessions. Additional objectives of this research are to develop a sensitive metric of domain expertise that can successfully distinguish expert from non-expert performance in order to validate the ground-truth conditions required for developing new predictors of expertise. With respect to advances in prediction, another objective of this work is to provide a rapid and accurate method for predicting domain expertise based on students’ performance-related behaviors, which also can provide an initial benchmark result for evaluating the reliability and speed of other new prediction techniques.

Method

Student Participants
Participants included 18 high school students, 9 female and 9 male, who ranged in age from 15 to 17 years old. All had recently completed Introductory Geometry at a local high school. They represented a range of geometry skills from low to high performers. For the basis of performance ratings, see [5]. Participants were paid volunteers and native English speakers.

During the data collection, small groups of three students jointly solved problems and mentored one another. Each student group was matched on gender and geometry skill level (low to high) to facilitate more reciprocal and collaborative interactions.

Math Tasks
During each session, students engaged in authentic problem solving and peer tutoring as they worked on 16 geometry and algebra problems, four apiece representing easy, moderate, hard, and very hard difficulty levels. These math problems were presented as word problems, and they varied along dimensions known to make them more challenging (for details, see [5]).

Data Collection Procedure
Each of the six student groups met for two sessions, during which students could view the math problems one at a time on a tabletop computer screen. The sessions were facilitated by a system that could present the problems with accompanying visuals, terms and equations related to solving the problems, worked
examples of how to solve problems, and the problem solutions. This system was controlled by a simulation environment, detailed in [1], which could deliver feedback multimodally. One student in the group was designated as the leader for a given session, and this designated leader switched on the group’s second session to a different student. This resulted in a total of 12 sessions involving 12 different leaders.

Each group was instructed to exchange information and expertise as they worked on solving the problems, so everyone understood the solution and could explain it if asked. To ensure that all students participated fully, they were told that during the session each participant would be randomly asked to explain one or more of the group’s math solutions. The group leader was told that he or she was responsible for coordinating the group’s activity and also interacting with the computer system. The leader was instructed that he or she could request the system to: (1) display the problem, (2) provide definitions of math terms, if needed, (3) provide formulas, if needed, (4) turn off, to avoid distracting students while working, (5) collect the group’s answer, (6) show the correct solution after the group’s answer was submitted, and (7) provide a sample solution for how to solve the problem. Usually during each problem, one of the students proposed a solution to the others, which then was discussed among them. Once it was agreed upon as correct and all students understood how to solve it, the leader submitted the group’s answer to the computer. Afterwards, the system displayed the correct answer, so students could verify their solution. If correct, one of the students sometimes was randomly called upon by the computer to explain how he or she had arrived at the solution. If the answer was not correct, the leader could ask to see a worked example of how the solution had been calculated, which the group discussed before the leader moved on to the next problem.

Regarding physical tools, all students could use a calculator, digital pen, and large sheets of digital paper to draw diagrams, make calculations, etc. while working on each problem. They were told that they could write with the digital pens just like a normal pen on paper, and that they could erase things by crossing out. None of the students had used digital pens previously. Following orientation and instructions, each group was given three sample math problems to solve in order to orient them with the problems, computer, and materials. Each group session lasted about 60 to 90 minutes.

**Multimodal Data Collection**

For each group session, twelve data streams were collected involving multimodal input from video cameras, digital pen, and close-talking microphones. Figure 2 illustrates high-resolution data collected from five camera views. Digital audio recordings were collected of each participant’s speech using close-talking microphones. A fourth combined audio recording also was collected using an omni-directional microphone hung above the table. Each participant’s writing was collected using Nokia digital pens and large sheets of Anoto digital paper. The pens provided the unique identification of each student, even when they wrote on one another student’s paper. Figure 3 illustrates multimodal data capture, and transmission of recordings to a screen in a separate room where display of problem content was controlled. Technical details of multimodal data collection are described elsewhere [1, 3].

In addition to being a reflective modality associated with the read-write cycle, pen input supports mobile and field use, such as classrooms. Unlike speech, students are more likely to communicate using pen input in classroom settings. Collection of written work based on digital pens also is similar to existing non-digital classroom work practice, and may be less prone to privacy concerns than recording of students’ speech and images. Finally, writing permitted capture and analysis of students’ work in mathematics, including frequent diagramming, symbols, and digits.

**Data Coding & Performance Metrics**

*Segmentation of Problems*— For each session, the data were segmented into start and end of the 16 problem-solving episodes.

*Initiator of Problem Solution*— The student who first proposed the group’s problem solution was scored for every problem, as well as whether the problem they solved was correct or not. Students’ correct problem solving also was coded by difficulty level.
Prediction of Domain Expertise Over Time—Change in the likelihood of predicting which student in a group was the dominant expert was coded as a function of problem position during a session depending on whether a student (1) contributed the group’s solution, or (2) contributed a correct problem solution.

Group Interaction Dynamics—Each group’s unique style of interacting with one another was described qualitatively, including how active and cooperative they were, the extent and nature of peer tutoring, disagreement during problem solving, social dominance of individuals, and other factors.

Results
Data were available for analysis on 18 individual students, 12 three-person group sessions, and 192 math problem-solving episodes (48 at each of four problem difficulty levels).

Validation of Problem Difficulty Levels
The average percentage of problems solved correctly and time-to-solution changed from 95.8% and 99.8 seconds on easy problems, to 81.3% and 118.4 seconds on moderate ones, to 81.3% and 150.4

Correctness of Problem Solution—Each problem was scored for correctness or incorrectness of the group’s solution, and also by difficulty level.

Time to Solution—The total time required by the group to solve each problem was coded, from the time a problem was displayed until the group submitted their answer. It also was summarized for each problem difficulty level.

Cumulative Expertise Rating—To assess domain expertise, each student’s cumulative problem-solving performance was calculated across their group’s two sessions. When students contributed an answer, the following numeric values were assigned: total number of easy problems solved versus missed (+1 or -1 pt.), moderate problems (+2 or -2 pts.), hard problems (+3 or -3 pt.), and very hard problems (+4 or -4 pt.). Based on totals, the dominant expert in each student group was identified.

Learning Across Sessions—Change in group problem-solving accuracy was summarized between sessions one and two as an index of learning, and summarized for different students (i.e., expert, non-expert, group leader).

Figure 2. Synchronized views from all five video cameras, including close-up of all three students, wide-angle view and a top-down view of students’ writing and artifacts on the table.

Figure 3. Integration of multimodal data capture using cameras, wireless microphones and digital pen input.
seconds on hard, and 64.6% and 158.5 seconds on very hard problems. This decline in percent correct and increase in TTS for easy to very hard provides ground-truth validation of problem difficulty levels.

**Problem-Solving Initiative, Correctness and Time-to-Solution as Indicators of Expertise**

*Problem-solving Initiative*—Overall, the expert students initiated an average of 21 problem solutions out of 32 total across their two sessions, compared with 5.5 for each of the two non-expert students in the group. That is, experts were almost four times more active in contributing group problem solutions.

Figure 4 shows the average number of solutions that expert versus non-expert students contributed for different problem difficulty levels. It reveals a strong divergence in problem solving initiative between groups for the moderate to very hard problems.

*Problem-solving Correctness*—Overall, experts contributed an average of 19.1 correct problem solutions across the two sessions, compared with 4.5 for each of the two non-expert students in the group. That is, experts had over a 4 times higher rate of contributing correct solutions. When examining correct solutions only on hard and very hard problems, experts had over a 5 times higher rate of contributing the correct solution, compared with a non-expert student (i.e., means 9.8 versus 1.9, respectively). Since non-experts were relatively inactive in contributing group solutions, their frequency of making errors was only 1.9 times higher than that of expert students (i.e., means 3.1 versus 1.6, respectively). As a result, the frequency of errors was a relatively poor indicator of domain expertise.

As shown in figure 5, the percentage of correctly solved problems out of those attempted averaged 100%, 95%, 92% and 86% for experts on low, moderate, hard, and very hard problems. In contrast, non-experts averaged 93%, 60%, 65%, and 41% on comparable problems. Overall, experts averaged 93.5% correct solutions, compared with 64.8% for non-expert students. Once again, a strong divergence was evident between experts and non-experts on the moderate to very hard problems.

*Time to Solution*—The average time for correctly solved problems increased across problem difficulty levels, as would be expected. Experts’ time for low to very hard problems averaged 103.8, 112.6, 121.4, and 160.0 across sessions. In comparison, non-experts averaged 91.7, 104.6, 169.6, and 214.7 across the same difficulty levels. Figure 6 illustrates that non-experts averaged much longer to solve hard and very hard problems than experts (mean 192.2 vs. 140.7), although the groups did not differ on easy or moderate ones (mean 98.2 vs. 108.2).

**Validation of Distinct Expert Performance**

A cumulative expertise score was calculated for each student based on the percentage of problems they solved correctly versus incorrectly at different problem difficulty levels, as described earlier. This measure effectively distinguished expert from non-expert performance, because expert students not only solved more problems correctly overall, they were especially likely to solve the more difficult problems correctly.
Figure 6. Average time in seconds for experts versus non-experts on correctly solved problems varying in difficulty level.

Figure 7 illustrates the distribution of expertise ratings for student experts versus non-experts during sessions one and two. These data are based on 32 math problem solutions across a group’s two sessions. They confirm that these students were distinct non-overlapping populations, which satisfies the ground-truth condition for developing prediction techniques that aim to distinguish these groups. Furthermore, these data confirm that clear separation of student groups replicated across both sessions. In fact, figure 7 shows that the groups separated further and became more distinct in expertise ratings during session two.

**Learning between Sessions**

To evaluate possible learning across sessions, each group’s percentage of correctly solved problems was compared between session 1 to 2, as shown in figure 8. Analyses revealed an average improvement in group performance of 9.4% on the second session, which was a significant change by Wilcoxon Signed Ranks test, $T=15$ (N=5), $p < .031$, one-tailed. In fact, five of six groups improved in their performance, with the sixth group showing no change.

Follow-up analyses were conducted on change in the number of correct solutions contributed by different types of student. Non-expert students and group leaders did not improve significantly across sessions, Wilcoxon Signed Ranks test, $T=0$ (N=7) and $T=4$ (N=13), both N.S. However, domain experts did solve marginally more problems correctly on the second session, $T=6$ (N=5), $p < .063$, one-tailed, in spite of small sample size.

Time to solution averaged 133.9 and 130.1 seconds on the first and second sessions. This lack of reduction probably reflected that students were mentoring one another to ensure that everyone knew how to solve the problem, not trying to solve them as quickly as possible.
Figure 9. Change in likelihood of reliably predicting which student in a group is the dominant expert at different points in a session, depending on whether they (1) simply contributed group solution, or (2) contributed correct problem solution

also was surprisingly effective at identifying the dominant expert by the first problem (50%), most of the first three solutions (83%), most of the first five (92%), and the first seven (100%). That is, knowing who was most active in initiating problem solutions reliably identified the expert 100% of the time after just seven problems, or approximately fourteen minutes. If focused only on hard and very hard problems, the expert can be identified 66% of the time on the very first problem, or in the first two minutes of a session.

Group Performance and Interaction Styles

Figure 8 illustrates that student groups varied considerably in their math problem-solving ability from low (group 2) to moderate (groups 1, 4, 6) to high (groups 3, 5). Most worked very actively and collaboratively together. They discussed the problems and checked that everyone understood and could explain the solutions. The dominant domain experts varied in their personalities, and they were frequently not the most socially dominant student so disagreement was common as students worked out answers.

The groups also varied considerably in interaction dynamics between individuals. For example, all students in the lowest performing group had extremely weak math skills, and they were the only group with no dominant domain expert. They averaged far more time when solving problems, and often resorted to guessing. Of the moderately performing groups, Group 4 included a domain expert who was socially dominant and sometimes teased and berated the other students by calling them names. One of the other two students sometimes impulsively contributed incorrect answers, and then argued for adopting them. Of the high-performing groups, Group 3 had a very reserved but capable domain expert, but another weaker student resented and frequently challenged her. Group 5 had a dominant expert who actively mentored the weakest student, including correcting, explaining, and prompting him for next steps. This expert held back and did not directly solve any problems. In summary, the student groups varied considerably in their math abilities and interaction patterns. In this regard, they constitute a diverse high-school population, and an excellent challenge dataset for developing MMLA predictors of domain expertise.

Discussion

The performance of domain experts was strikingly different from non-experts during these collaborative math problem-solving sessions. They were almost four times more active in contributing group problem solutions, and also over 4 times more likely to contribute correct solutions. These differences became more pronounced on problems beyond the easiest difficulty level. Domain experts also took 27% less
time to solve hard and very hard problems correctly, compared with non-experts, although there was no difference on easy to moderate problems. A comparison of group performance also revealed a 9.4% improvement during the second session, with learning evident among the more expert students.

A cumulative expertise score based on both (1) correctness of solutions and (2) problem difficulty level was applied to all student participants. It validated that expert and non-expert students were distinct non-overlapping populations, a finding that replicated across sessions. This performance distinction between expert and non-expert students provides the required ground-truth condition for using the Math Data Corpus as a basis for developing new prediction techniques that aim to detect domain expertise and change in expertise.

With respect to advances in prediction, a rapid and accurate basis was revealed for predicting domain expertise based simply on knowing who contributed most group solution, whether correct or not. That is, identifying the most initiating student in contributing solutions was a surprisingly effective way to determine the dominant expert (50% of the time on the first problem, 83% on the third, 92% on the fifth, and 100% of the time by the seventh problem). This information alone reliably identified the dominant expert 100% of the time after just seven problems, or in approximately fourteen minutes. If a focus is placed only on hard and very hard problems, then 66% of the time the expert can be reliably determined on the very first problem, or first two minutes of a session.

Perhaps counter-intuitively, simply contributing the majority of solutions was more effective at identifying domain expertise rapidly across serial problem positions, compared with contributing the majority of correct solutions—a finding of value for real-time prediction. Student activity in simply contributing solutions does not require any judgment regarding correctness, so it also is easy to score and a good candidate for automation. This metric therefore provides a particularly valuable index of expertise, either individually or combined with other factors. One caveat is that this metric potentially can fail in cases where no domain expert is present at all, because non-experts may contribute many solutions that are incorrect. For this reason, future research should explore combining this metric with others to ensure highly reliable identification in a wider variety of real-world group compositions.

This predictor of domain expertise based solely on problem-solving initiative provides an initial benchmark, and also model for evaluating the prediction speed and accuracy of other newly developed prediction techniques. Together, the Math Data Corpus and present results contribute critical resources for supporting data-driven grand challenges on multimodal learning analytics, which aim to develop new techniques to predict expertise early, reliably, and objectively. The Math Data Corpus includes a range of low to high performing students and different group interaction styles. It also is focused on authentic problem-solving and peer mentoring sessions involving collaborating student groups, with evidence of learning over time. However, future efforts will need to expand the available data resources to include other domains and learning contexts, as well as larger datasets, so that promising new MMLA prediction techniques can be generalized appropriately.

Acknowledgments
This research was supported by Incaa Designs (http://www.incaadesigns.org/).

References