**Multimodal Learning Analytics Data Resources:**

**Description of Math Data Corpus and Coded Documents**

Sharon Oviatt, Adrienne Cohen¹ Nadir Weibel², Kevin Hang³ & Kathryn Thompson⁴

Inca Designs

1140 Wing Point Drive, N.E., Bainbridge Island, Wa. 98110

oviatt@incaadesigns.org

¹Univ. of Washington; ²UC San Diego; ³Univ. of New South Wales; ⁴Univ. of Sydney

**Abstract**

This paper provides documentation on critical dataset resources for establishing a new research area called multimodal learning analytics (MMLA). Research on this topic has the potential to transform the future of educational practice and technology, as well as computational techniques for advancing data analytics. The Math Data Corpus includes high-fidelity time-synchronized multimodal data recordings (speech, digital pen, images) on collaborating groups of students as they work together to solve mathematics problems that vary in difficulty level. The Math Data Corpus resources include coding of problem segmentation, problem-solving correctness, representational content or students’ writing, and spoken content of students’ speech. These resources are made available to participants in the data-driven grand challenge for the Third International Workshop on Multimodal Learning Analytics. The primary goal of this event is to analyze coherent signal, activity, and lexical patterns that can identify domain expertise and change in domain expertise early, reliably, and objectively, as well as learning-oriented precursors. An additional aim is to build an international research community in the emerging area of multimodal learning analytics by organizing a series of workshops that bring together multidisciplinary scientists to work on MMLA topics.

**Introduction**

Multimodal learning analytics (MMLA), learning analytics, and educational data mining are emerging disciplines concerned with developing techniques to more deeply explore unique data in educational settings [2, 4]. They also use the results based on these analyses to understand how students learn. Among other things, this includes how they communicate, collaborate, and use digital and non-digital tools during learning activities, and what the impact of them is on acquiring new skills and knowledge. Advances in learning analytics are expected to contribute new empirical findings, theories, methods, and metrics for understanding how students learn. They also can contribute to improving pedagogical support for students’ learning through assessment of new digital tools, teaching strategies, and curricula. In addition, MMLA is expected to advance new computational and engineering techniques related to machine learning and data analytics.

The most recent direction within learning analytics is multimodal learning analytics, which emphasizes the analysis of natural rich modalities of communication during situated interpersonal and computer-mediated learning activities. This includes students’ speech, writing, and nonverbal interaction (e.g., gestures, facial expressions, gaze). A primary objective of multimodal learning analytics is to analyze coherent signal, activity, and lexical patterns in order to uncover entirely new learning-oriented phenomena. It aims to identify domain expertise and changes in domain expertise early, reliably, and objectively, and also to examine learning-oriented precursors.

There are many potential advantages of developing multimodal learning analytics capabilities. First, whereas learning analytics examines students’ keystroke input during technology-mediated learning, MMLA analyzes natural rich modalities of communication (e.g., speech, writing, gestures, gaze) during both human-human and human-computer learning exchanges. As such, multimodal learning analytics techniques can be used to evaluate both interpersonal and technology-mediated learning. This is important, because learning occurs in both contexts. In addition, a human-human baseline of learning is required to evaluate the impact of new educational interfaces. Secondly, the dominant educational technology platform worldwide is mobile cell phones, which are not keyboard-dominant devices. Instead, their input capabilities include speech, touch, virtual keyboard, and in some cases sensors and pen input. In short, mobile cell phone interfaces incorporate natural rich communication modalities, and they are becoming increasingly multimodal [4]. As a result, learning analytics techniques that rely on click-stream analysis are limited in their utility and will become increasingly obsolete as cell phone use accelerates in developing
regions. Thirdly, MMLA techniques are capable of tracking multiple natural communication modalities, which provide a particularly informative window on human thought and learning. They also 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. MMLA 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. One long-term objective of this research is to develop automated prediction of expertise and change in expertise that is more accurate and less affected by biases in social perception than human decision makers. For example, this includes accurately distinguishing domain experts from individuals who are simply socially dominant. Another is to apply these new capabilities to improving assessment of the impact of new digital tools, teaching strategies, and curricula on students’ learning outcomes.

The First International Conference on Multimodal Learning Analytics (http://tltl.stanford.edu/mla2012/) represented an initial intellectual gathering of multidisciplinary scientists interested in this new topic. The Second International Workshop on Multimodal Learning Analytics brought together an international collection of researchers to advance research on multimodal learning analytics with a data-driven grand challenge event (https://tltl.stanford.edu/mla2013). In support of this event, the Math Data Corpus and related coding resources were made publicly available by Incaa Designs (http://http://www.incaadesigns.org). In addition, the ChronoViz multimodal data analysis tool is supported by UCSD for workshop grand challenge participants who wish to use it (http://chronoviz.com/quickstart.html).

The purpose of this paper is to provide documentation on the expanded data resources associated with the Math Data Corpus, which were available for participants in the Third International Data-Driven Grand Challenge Workshop on Multimodal Learning Analytics in 2014. These resources include high-fidelity time-synchronized multimodal data recordings on collaborating groups of students as they work together to solve mathematics problems varying in difficulty. Data were collected on students’ natural multimodal communication and activity patterns, including their speech, digital pen input, facial expressions, gestures, and physical movements. The dataset includes 12 sessions, with six three-student groups who each met twice. In total, approximately 29 student-hours of recorded multimodal data is available during these collaborative problem solving sessions. This data resource includes coding of problem segmentation, problem-solving correctness, representational content on students’ writing, and speech transcriptions of students’ lexical content. Further specifics on this dataset are described below, including detailed appendices. This dataset and its related resources are available to participants in the Third International Data-Driven Grand Challenge Workshop on Multimodal Learning Analytics, after signing a collaborator agreement, at the following website: http://mla.ucsd.edu/agreement.pdf.

**Student Participants**

Participants in this study 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 and represented a range of geometry skills from average to high performers. Performance rankings were based on teacher evaluation of students’ average performance over the one-year course and pilot results, and then confirmed by performance during the study. Participants were paid volunteers and all were native English speakers.

During the data collection, small groups of three students who were gender matched jointly solved problems and mentored one another. Each student group was matched on geometry skill level (low, moderate/high, and high) to facilitate more reciprocal and collaborative interactions. Appendix C contains the unique student ID codes associated with the anonymous dataset.
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. These dimensions included the number of math terms, the number of equations needed to solve the problem, whether or not the equation could be applied directly, whether units required translating (i.e. inches to feet), whether values required translating (i.e. volume to gallons) and the number of steps required. Figure 1(a) shows an example of a low-difficulty problem that included one math term (e.g. square feet) that could be solved by directly applying the equation for area of a rectangle. Figure 1(b) shows an example of a very high-difficulty problem, which contained four math terms (e.g. pyramid, square, volume, cubic feet) and required equations for the area of a rectangle and the volume of a pyramid. The difficulty levels of the problems were validated using: 1) teacher records of percentage correct on similar problems for high school students in introductory geometry, 2) pre-experimental piloting, and 3) students’ percentage of correct solutions in the current study. Appendix D contains all problems used during data collection for both sessions, including problem sets 1 and 2.

Data Collection Procedure
Each of the six student groups met for two sessions, during which students could view the math problems displayed one at a time on a tabletop computer screen. The sessions were facilitated by a tutoring 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. 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. Since a different leader was selected from each group for their second session, a total of 12 separate sessions involving 12 different leaders was conducted.

Each group was instructed to exchange information 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 by a research assistant that during the session each participant would be randomly asked to explain one or more of the group’s math solutions:

“Today you're going to have an informal group session to study and learn about Algebra and Geometry together. As you work together on different math problems to solve them, be sure to exchange information and expertise so everyone understands how you arrived at the solutions and could explain them to someone else. If anyone doesn’t understand how you got a particular solution, then be sure to ask questions and continue discussing it until you feel comfortable that you could explain the steps to someone else. During the session, you will each randomly be asked to explain one or more of the problems.”
Figure 2. Synchronized views from all five video-cameras at the same moment in time during data collection. Videos A, B and C show close-up views of the three individual students, video D a wide-angle view of all students, and video E a top-down view of students’ writing and artifacts on the tabletop.

Within each group, one student was designated as the leader and was told that he or she was responsible for both coordinating the group’s activity and interacting with the computer system that facilitated the group’s problem solving:

“As you work together, one person will be the designated group leader for coordinating things and getting information from the computer. The leader will be the person who coordinates the group and conveys what the group wants to the computer. For today’s session, you’ll be the leader [select student]. You can use a calculator, paper and pens, and the computer while you’re studying. The most important thing you do today will be just studying informally together, and working on understanding math better. Of course, no computer can do that for you! The idea is to collaborate on coming up with solutions and making sure you really understand them. And it’s fine if your group solutions aren’t always correct—you can rework a problem together, but if you get stuck just go on to the next one. So, we’d like you to just focus on the math problems, and have fun working together!”

Regarding the computer display, students were further instructed that:

“The computer really only has a few simple things it can do:

• provide the math problem
provide definitions of math terms, if you want them (for terms highlighted with a box, and others)
provide formulas, if you want them
turn off (to avoid distracting you while you work together on the problem, the important part!)
collect the answer you submit and then provide its answer (only after you’re done & submit your answer…)
provide a sample solution as an example of one approach to solving the problem

When you’re ready to start a new problem, inform the computer that you want it displayed. Then you should read over the problem and make sure you all understand it, including discussing it and making any notes you may need to prepare yourselves to actually begin work. You should let the computer know when you’re ready to begin work, and it will turn off the screen to avoid distraction until you finish. When you all agree you’re done solving a problem, the leader should submit the group solution. Be sure it has been received before requesting the system’s answer. Once you’ve compared it against yours, and discussed it if necessary, you can get the worked example showing the solution. We hope these things help you learn and make things go smoothly while you’re studying together.”

Regarding the digital pens students were given, they were told:

“You’ll also each have your own pen, and please don’t exchange it with the others. It writes just like a normal pen on paper, and you can erase things by crossing them out.”

Following orientation and instructions, each group was given three sample math problems to familiarize themselves with the problems, computer, and materials:

“Now we’ll do some practice together so you can try out some math problems and see exactly how the computer works. When you've finished the first problem set, you’ll see a “new phase” screen. Come and get me and I'll get you set up for the next set.”

After completing the practice problems, the research assistant left the room and the group began the main session that lasted about an hour. After each problem set involving four math problems, the research assistant entered the room and there was a brief pause in group activities to set up for the next set before resuming again.

While working on each problem, the leader would begin by asking to see the next problem. Students then typically discussed it with one another, and the leader could ask for related mathematical terms and equations on behalf of the group. All students could use pen and paper and a calculator as tools to draw diagrams, make calculations, etc. while working on each problem. One of the students usually proposed a solution to the others, which then was discussed among them. Once it was agreed upon as correct and all students understood how it had been solved, the leader then 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 was randomly called upon by the computer to explain how they 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 they then discussed before the leader asked to see the next problem.

**Computer Facilitation of Problem Solving Sessions**

To summarize, the computer assistant provided basic functions that included: (1) displaying each math problem, (2) turning off the display to avoid distraction during problem solving, (3) providing definitions of math terms or equations if requested, (4) accepting answers submitted by the group, (5) displaying or confirming the correct answer, and (6) displaying worked examples of problem solutions. This information was provided on the computer display on the table in front of students, and was controlled by a simulation environment detailed in Arthur and colleagues (2006). The computer was capable of delivering feedback multimodally, such that the graphical display on the monitor was occasionally accompanied by text-to-speech output such as “problem one” when showing the next problem, or “here’s one solution” when displaying a worked solution.

**Multimodal Data Collection**

For each session, a high-resolution digital video close-up of each participant from the waist up was captured using Point Grey Scorpion digital firewire cameras, with resolutions of 1024x768 at a frame-rate of 15 Hz. Unobtrusiveness was a key characteristic for selecting the recording hardware to use in our infrastructure. The Pt. Grey Scorpion cameras have a 50mm x
50mm x 40mm form factor, which can be easily hidden on bookshelves or from the ceiling. In addition, a wide-angle room view and a view of the tabletop with paper, pens, and other artifacts was recorded. Figure 2 shows five video frames taken at the same moment in time during group problem solving. Figure 3 illustrates multimodal data capture, and transmission of recordings to a screen in a separate room where display of problem content was controlled. Transmission of students’ speech and pen input was wireless.

Digital audio recordings were collected of each participant’s speech using Countryman Isomax hyper-cardioid microphones, which were connected to Shure wireless transmitter/receivers. These miniature microphones were used because they are unobtrusive and flesh-toned in color, permit natural movement, and do not occlude facial regions during videotaping. They frequently are worn by stage performers. In addition, a fourth digital audio recording of the room was collected using a studio-quality omni-directional microphone hung above the leader’s head.

Finally, each participant’s writing was collected using Nokia digital pens and large sheets of Anoto digital paper. These pens provided the unique identification of each student, even when they wrote on one another’s paper. In addition to being a “reflective” modality associated with the read-write cycle, pen-based writing is important because it supports mobile use in natural field environments like classrooms. Unlike video, it is easy to record high-quality writing in classrooms. And unlike speech, students are more likely to communicate extensively using pen input in classroom settings. Finally, ink is important for our current research because we aim to analyze students’ work in mathematics, which includes frequent diagramming and written symbols and digits.

**Data Files and Data Labeling**

The MMLA data files available for download on the web ([http://mla.ucsd.edu/data](http://mla.ucsd.edu/data)) are organized into the twelve sessions--“G1D1” is student group 1 session 1, through “G6D2” for student group 6 session 2. Within each session folder is a set of audio, digital pen, video, and additional coding files and explanations (e.g., problem time boundaries, performance coding, coding of written representations) for the three students who participated in that group session.

For every session (every G#D# folder), the dataset is composed of five different data:

1. **“Audio G#D#”** contains four .wav files – one close-talking recording for each student (labeled with their unique ID), and one recording (labeled ALL) of all student voices together.

2. **“DigitalPen G#D#”** contains three sets of digital pen recording of student writing, one for each student (labeled with their unique ID). It also contains a .csv file labeled “G#D# ProblemSet# Linkages_WrittenCoding.csv” that specifies the temporal offset of the first recorded pen stroke with the start of the recorded group session (e.g. the videos and audio files). Within each student folder is a set of: (1) .jpegs showing images of what he or she wrote on each page of digital paper, and (2) .xml files, including ones containing a given student’s strokes for each page of digital paper they used and also a “session.xml” file for each student that encodes the relationship between the single stroke files. The pen stroke files are encoded using a hierarchical
structure based on (a) a note per page, (b) a set of trace per note, which represent the single strokes, and (c) a set of point per trace, representing the inking trajectory within the given stroke. For every point with an (x,z) coordinate, the timestamp and the pressure on the pen nib is recorded. These data result in a continuous record of the writing from the digital pen of every student during the given session. For the digital pens used, pen stroke data was sampled at an average rate of once every 26.7 milliseconds, or about 37 times per second.

All of the audio and digital pen files include continuous recordings of the student for whom the unique student ID is provided on each file for a given session.

3. “Video G#D#” contains two complete sets of video records, high and low resolution. In the “High resolution” folder, there are five high-resolution .avi files taken from different cameras, one wide-angle view of all three students while working (labeled WIDE), one top-down view of their writing and tabletop artifacts (labeled TOP), and three close-up images of each of the three students (labeled with their unique ID). The “Low resolution” folder contains the same data stored in lower resolution.

IMPORTANT NOTE: For the video recordings of individual students’ close-up views, the file labeled with one unique student ID is a continuous video recording of the student leader for that entire session. The other two files are labeled with two student IDs. For these recordings, the first ID identifies the student shown in the video during problems 1-4 and 13-16, and the second ID is the student shown during problems 5-12. That is, an individual video camera sometimes recorded two different students during a session. These always were the non-leader students, who switched seats during the session. In these cases, video analysis of a given student must combine segments from the two different camera recordings.

4. “Coding”

A. “Time boundaries and Performance” includes both Word documents and .csv files that contain information for each session on the problems, time boundaries, and group performance during problem solving. These .csv files can be processed directly, or used in conjunction with corresponding signal files (audio, visual, digital pen) to segment them correctly for analysis purposes (e.g., by problem, by problem difficulty level, by correctly or incorrectly solved problems, etc.). The order of information contained in the .csv files, as shown from left to right with each information type divided by a comma, is:

- **Problem** (e.g., “1A”), with 1-4 indicating problem difficulty level (1-easy; 2-moderate; 3-hard; 4-very hard) and A-D the problem set
- **Problem start** (e.g., “1:59”), or time in minutes and seconds from the beginning of the group session when a given problem statement was first displayed for students to view
- **Problem end** (e.g., “3:08”), or time in minutes and seconds from the beginning of the group session when students discontinued working on a given problem and asked to see the next one
- **MOI start (MOI=moment of insight)**, or time in minutes and seconds from the beginning of the group session when the student who contributed the group’s problem solution first proposed it for a given problem
- **MOI end**, or time in minutes and seconds from the beginning of the group session when the group submitted their solution for a given problem
- **TTS (TTS=time to solution)**, time in seconds to complete a given problem solution, or the number of seconds between problem start time and when students submitted their group’s solution
- **Correct**, or whether students’ problem solution was correct (Y) or not (N) for a given problem
- **Person initiating answer** (e.g., “S12”), or student who first suggested the problem solution that the group later submitted
- **Person explaining** (e.g., “S12”), or student who explained the group’s solution after it was submitted, if prompted by the system to do so; if the system did not ask for an explanation on a given problem, it was coded as “No”

The Word document on “Time boundaries and PerfScores” also contains the information provided in the processable .csv file for a given session. In addition, it includes performance summaries for the group, such as the (1) total percentage of problems solved correctly for the group, (2) average time to solution for problems solved by the group, (3) total number of problem
solutions contributed by each student, (4) number of hard problem solutions contributed by each student (i.e., classified as difficulty level hard or very hard), (5) percentage of correct solutions of those contributed by a given student, (6) expertise score for each student during a given session, (7) cumulative expertise score for each student across their two sessions, (8) qualitative notes about any missing data or time regions to consider excising during data analysis (e.g., audio analysis excise due to extensive laughter), and (9) qualitative notes about student group dynamics during problem solving. The notes also contain information such as how expertise was calculated for each student. Note that group two was the only student group that did not have a clearly dominant domain expert.

B. “Written Representations” is a set of .csv files (converted from excel files), one for each student during each session, that coded student writing for the type of written representation, its semantic content, and start and end of inking. These files can be processed and analyzed directly. They also are required to time-align digital ink signal information with its type and meaning content in order to produce more meaningful analyses of student writing. Note that by doing so, spurious ink marks are filtered out. The order of information contained in the .csv files, as shown from left to right with each information type divided by a comma, is:

- **Type Representation**, which was coded as either a diagram (D), linguistic (L), numeric (N), symbolic (S), or a mark (M) for task relevant content. Any task irrelevant content was coded as “TI.”
  - **Diagramming** was further coded into the number of discretely different diagrams (e.g., “D1”, “D2”), whether a diagram was low, moderate, or high in complexity (e.g., “D1-LC”, “D2-HC”). If a diagram showed content in 3D, this also was coded (e.g., “3D”)
  - **Disfluent** writing during a given written representation was further coded as “DIS-CC” if it involved a content correction (e.g., deleting or changing the meaning of past written content), as “DIS-R” if a student retraced part or all of past writing, as “DIS-A” if a student returned to a diagram to add to it incrementally, and as “DIS-S” if a word was misspelled or a symbol misformed. For example, a written representation would be coded as “S DIS CC” if a student wrote a multiplication symbol over a previous addition symbol during a given entry. The “Notes” section also would identify the original content that was corrected.

In the .csv files, the “Type Representation” field always lists the basic type of representation first (D, L, N, S, M, TI). If the type is “D,” then subtype information follows about diagram number and complexity (e.g., “D1-HC”). Any information identifying a representation as disfluent is listed last in the “Type Representation” string. It identifies that a disfluency occurred and what type (e.g., “DIS-CC”). Since these acronyms and their location in the string are distinct, they can be automatically located and processed. For example, all “TI” representations could be filtered out, and then “D” representations could be processed separately from others. Likewise, all disfluent representations (“DIS”) could be processed separately from those that were not disfluent.

- **Meaning Content**, summarized the semantic content of the written representation (e.g., “18”, “answer”, “pyramid”)
- **Start time** (e.g., “00:31.870”), or time in hours, minutes, and seconds from the beginning of a given session when the student started to write the representation
- **Stop time**, or time in hours, minutes, and seconds from the beginning of a given session when the student finished writing the representation
- **Notes** explains information about the writing, including whether linguistic or numeric content labeled a diagram and was part of a complex diagram structure (e.g., “Labels D1”), or the meaning of sketched content that was added to a diagram incrementally (e.g., “diameter line added”), or what previous written content was modified when there was a disfluency involving a content correction (e.g., “corrects row 7”), or the apparent meaning of any nonstandard abbreviations (e.g., “hd” for “hot dogs”)

See Appendix B for a detailed explanation of how written representations were coded. If a student did no writing during a particular problem, their .csv file is empty. The inter-rater reliabilities for coding different types of written representation, including whether they were disfluent or not, averaged 85-98%.
Note that if digital pen signal data is filtered using the aligned coding of written representations, then spurious ink marks not associated with meaningful writing (e.g., inking when resting one’s hand on paper) are filtered out automatically. In addition, task-irrelevant content (e.g., doodling) can be filtered out entirely, as can disfluent constructions.

C. “Speech Transcriptions” is a set of Excel files (along with corresponding .csv files), one for each of the twelve math problem-solving sessions. The speech transcriptions for each session were created by a three-pass process by three separate professionals, which ended with a final-pass transcription and reliability checks on the accuracy of all transcribed content. Transcriptions include all verbatim spoken content by each of the three students in a given session that fell within the start and end time boundaries of the 16 math problems they solved together. Additional spoken utterances sometimes were contributed by the computer and research assistant, which have been left in the transcripts for context. Each spoken utterance or contribution includes information about start time, end time, student speaker, problem number, and the semantic content spoken. All speech was transcribed verbatim, including confirmations, disfluencies, partial or broken off utterances, and so forth. Unintelligible words are marked “xx.” Comments are marked in square brackets, including episodes of non-speech audio such as laughter, humming, or throat clearing. Double slashes frequently have been used to mark the end of a broken off word or utterance. A question mark at the end of an utterance indicates rising intonation associated with uncertainty and question asking.

The .csv files can be processed and analyzed directly. For example, they support analysis of semantic content, and speech signal features in conjunction with certain types of lexical content or dialogue contributions. They also support analysis of spoken dynamics within the group, such as lexical accommodation between speakers, overlapped speech, and so forth. The order of information contained in the .csv files, as shown from left to right with each type of information separated by a comma, is:

Start time, end time, problem number, speaker, utterance content.

The inter-rater reliability for coding the lexical content of verbatim speech on the final-pass transcripts averaged 95%.

Note that if the speech signal data is filtered using the aligned lexical transcriptions, then only speech within the start and end of the 16 problem boundaries will be analyzed. Some acoustic signals associated with non-speech audio (e.g., laughing, coughing) also will be filtered out. Based on the verbatim transcripts, it also is possible to search for disfluent lexical constructions and either (1) analyze them separately, or (2) exclude them from analyses.

5. “Archives” contain a set of high and low resolution zipped video files. While the single data can be downloaded separately, in order to facilitate the download of the complete datasets, we created a zipped archive for every session. These zip files are named G#D#_Data_LowResVideo.zip and G#D#_HighRes_Videos. They are stored in the “G#D#/Archives/” folder for every session. The G#D#_Data_LowResVideo.zip files contain all the data described above in a single archive, excluding the high resolution videos (which are very big files). The G#D#_HighRes_Videos.zip contain only the high-resolution video files. Both archive are particularly big (i.e., approximately 1GB for the low resolution archives and 10GB for the high resolution ones). Plan for long download times (2-3 hours) when downloading these files.

6. “ChronoViz” contains a set of integrated multimodal data files, which can display all of the above data together along a timeline. These data are being prepared by Nadir Weibel for participants who would like to display them using the ChronoViz multimodal analysis tool. They will be added to the dataset after the main website is open, and will be described in a separate document.

Multimodal Data Synchronization & Processing

Video Processing & Synchronization—To accomplish tight time synchronization among the five video streams, the videos were recorded on three Unix machines with clocks synchronized via the Network Time Protocol (NTP). Each raw Bayer video frame was compressed by a method using Haar wavelet transformations with quad-tree coding similar to the technique in [6], resulting in about 50% compression. Every frame was time-stamped and written to disk in an in-house media file format. A new output file was created every minute to avoid exceeding the maximum file size.

The video collection software suite was composed of several software modules, each responsible for a particular task, e.g. frame-grabbing, color space conversion, video rendering, file writing, etc. See [1] for a flowchart illustrating the video capture
process. The data streams are piped between modules using NIST Smart Flow [5]. Smart Flow is used to decompose the task of video collection into a number of specialized agents, each responsible for a specific task. Smart Flow is the infrastructure that permits agents to communicate with each other. It allows software modules to publish or subscribe to data flows of interest, such as RGB video or PCM audio. It simplifies inter-process communication, especially for applications running on different machines.

Audio Processing & Synchronization—The audio signals from close-talking microphones worn by each meeting participant were transmitted wirelessly with a Shure UHF transmitter. The Shure receiver was connected to a Mobile Pre M-Audio USB external analog/digital converter. Using an external analog/digital converter, as opposed to an internal soundcard, eliminated electrical interference noise introduced by the machine. All audio streams were recorded with a single Windows XP machine using DirectSound into the same in-house file format as video. We recorded audio on Windows because of better soundcard support and the benefits of DirectSound, such as the automatic recognition of a range of sound cards and USB analog/digital converters, and the automatic buffering of incoming audio data. USB analog/digital converters supported recording multiple audio streams on a single machine without having to install additional hardware.

Ink Processing & Synchronization—During data collection, ink was captured using Nokia digital pens, and was recorded and time-stamped locally on the pen. When the pen was physically docked in a USB enclosure, the data was sent by Anoto components via COM to our software on the host machine, which saved it to disk in an XML file. To effectively synchronize the digital ink with audio-visual media streams for each student in each session, we used ChronoViz to hand synchronize their first pen down using two reference videos, a close-up video of the student writer and a top-down video of all students’ digital paper and pens. These videos were viewed while watching the start of the dynamic ink trace in the digital ink recording. These playbacks were viewed at ¼ speed. To find the exact moment when the ink trace began, the scorer quickly tapped forward on the pause button to advance the recording a few milliseconds at a time. We then re-checked the digital ink synchronization at the end of the file to ensure that the most accurate initial synchronization time had been selected. The accuracy of synchronization of the inked data with the audio-visual recordings averages 0.5 to 1 second.

Data Stream Synchronization—Upon completion of a session, the saved audio and video files were encoded to universally viewable MPEG-4 media streams. A video stream was paired with an audio stream to form a single media stream (e.g., for a close-up of a given student), and the different audio-visual streams all were synchronized for the entire study session. The network of computers running the data collection software suite was comprised of machines running both Windows and Unix. The fundamental feature enabling this heterogeneous network of machines to work together as a single data-collecting unit was that all machines periodically synchronized their clocks with NTP. By recording data in our in-house file format, with a header containing rich meta-data for every frame, the streams could be synchronized with any input source that has a clock synchronized via NTP. This also means that our infrastructure is extensible, and in the future could incorporate data from touch-sensitive white-boards used by teachers in classrooms, slideshow presentations during instructional interactions, supplementary sensors such as accelerometers, and other sources.

To estimate video-to-video synchronization fidelity, we examined manual contact with physical artifacts during videotaped data recordings, including different pair-wise combinations of cameras and also videos recorded on separate machines. We also took estimates both at the beginning and end of every recorded session, in order to estimate possible drift in synchronization. Videos were shown to be synchronized with a mean departure of 0.025 seconds (0.0005 seconds per minute), and to within one frame or 0.067 seconds for 100% of the cases.

To estimate audio-to-video synchronization fidelity, we examined speakers producing utterance-initial consonants (e.g., ‘p’). The audio and video signal were shown to be synchronized with a mean departure of 0.077 seconds (0.0016 seconds per minute), and to within two frames or 0.12 seconds for 80% of cases.

Using the ChronoViz multimodal analysis tool, all data streams are shown as linked files with a start entry time for each file. In the Math Data Corpus, each video file already was synchronized with the corresponding audio file (see above description). A given student’s digital ink file was hand synchronized using ChronoViz with the corresponding audio-visual file of the student. To estimate the accuracy of digital ink synchronization over the session with the audio-visual recordings (i.e., any pen synchronization drift for the three Nokia digital pens), the average departure in correspondence between pen local timestamps and audio-visual timestamps was compared from start to end of sessions. Based on this estimate, the mean change was 0.004
seconds per minute.

The above summary of the accuracy of multimodal data stream synchronization provides guidance to those analyzing the present dataset on the degree of temporal fidelity that can be achieved in examining any temporal phenomena.

**Data Summarization, Ground-truth Conditions for MMLA, and Detailed Appendices**

**Segmentation of Data**—As discussed above, the data were segmented from each session into time phases representing the start and end of each of the 16 problem-solving episodes. The moment of insight (MOI) was recorded, or time when one student first proposed the problem’s solution, which the group later submitted as their answer. Time of answer submission also was recorded. Based on answer submission time, the total time required by the group to solve the problem (TTS) was summarized (i.e., time from start of each problem display until students submitted their answer). After submitting an answer, remaining time spent on the problem involved examining and discussing the answer and worked example. On some problems, one student also was asked by the computer to explain how the group arrived at their problem solution.

**Ground-Truth Scoring of Problem Difficulty, Performance and Expertise**—Each problem was coded: (1) by difficulty level (1-4 representing low to very high difficulty, respectively), (2) as correctly or incorrectly solved, (3) by which student initiated the answer, and by (4) time to solution (TTS).

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

See Appendix A for detailed information on problem segmentation, problem difficulty level, problem-solving performance, summaries of student expertise level, and related information on this dataset. See “Data Files and Data Labeling,” section on “Coding” for related information on dataset organization. In addition, a companion paper to this one is available on the website and in the ACM Digital Library, which documents that (1) students classified as “domain experts” in this dataset were a reliably distinct and non-overlapping group from those identified as non-experts, and (2) significant learning occurred during the longitudinal sequence represented in this dataset (i.e., from session 1 to session 2). This and related ground-truth analyses are available in:


Together, the above resources establish ground-truth conditions for summarizing domain expertise at both problem-centric and student-centric levels.

**Written Representations**—The written content for each student during the sessions was coded for type of representation, type of diagrams, presence of written disfluencies, and task relevant versus irrelevant content. See Appendix B for detailed written representations coding scheme.

**Speech Transcriptions**—The verbatim content of speech for each student during the group sessions was coded for lexical content, start time, end time, and math problem.

**Student Identification Coding**—Each of the 18 participating students was coded with a unique numeric ID (1-18), gender status (M, F), leadership status for a given session (L, NL), and expert status (E, NE). Their coded data files also are labeled with their group (1-6), session (1-2), and problem set (1-2) numbers. The master list of students’ unique identification coding is provided in Appendix C.
Problem Sets and Content—The problem statements, answers, example solutions, and any related visuals are provided in full in Appendix D for the 16 math problems that students worked on in problem sets one and also problem set two.

Multimodal Analysis Tools
Although not required to support participants’ analyses during the competition, ChronoViz [7] is available for analyzing time-synchronized multimodal data streams. ChronoViz is a free analysis and visualization tool that can be downloaded from: http://chronoviz.com/. A custom built version of ChronoViz that better supports the Math data corpus is available at http://mla.ucsd.edu/data/ChronoViz.zip. It supports visualizing the data, including inspecting the relation between different data streams (e.g., speech and writing activity) at any given point in time. It also supports adjusting the granularity of the time scale to view micro-events, coarser level events, or an overview, and integrating researcher annotations indexed along the timeline. ChronoViz has been used to analyze various types of data, including speech, writing, visual images, and physical activity patterns, as will be required for the present research competition. In addition to supporting synchronized multimodal data streams, it has been used previously to analyze multi-person naturalistic interactions and data from situated field settings by Nadir Weibel, Adam Fouse, Jim Hollan and their collaborators from UCSD. A growing community is using this tool to analyze different kind of multimodal data. ChronoViz can be used in conjunction with other more specialized signal analysis tools, such as Praat for speech signal analysis. Workshop participants are free to use other analysis tools of their preference. For questions or help with using ChronoViz, please see: http://chronoviz.com/quickstart.html. If additional help is required, contact Nadir Weibel at UCSD, weibel@ucsd.edu.

Collaboration Agreement
This dataset was collected on a former research project by Sharon Oviatt’s lab. IRB clearance has been obtained for collaborating researchers to analyze them. Workshop participants will be required to sign a collaborating researcher agreement, which involves agreeing to observe a protocol that protects the anonymity of the original research participants. While all names of student participants in this dataset have been converted to anonymous identification numbers, they may still be identified by images or voice in our rich multimodal recordings. The second requirement for those who access and use this dataset is that they openly publish their findings for community benefit and clearly cite the present document as the source of their analyzed data. The correct reference is:


In addition, any further coding, transcription, or other data resource enhancements performed on this dataset by workshop participants must be contributed to the publicly available dataset resources for others to use in the future. All contributors will be fully acknowledged for their work. A copy of the collaborator agreement is listed in Appendix E.

Data-Driven Grand Challenge Competition and Evaluation
The purpose of the data-driven grand challenge event organized as part of the Third International Workshop on Multimodal Learning Analytics is to use the Math Data Corpus to explore and develop new techniques that can successfully:

• Predict which student in a 3-person group is the dominant expert
• Predict which problems will be solved correctly versus incorrectly by the group (i.e., of the 32 available)

These predictive techniques should aim to identify this information (1) with high reliability, (2) as rapidly as possible (i.e., during the first 1 versus 5 mins. of group problem solving) (3) using a technique that generalizes across all problem difficulty levels, and/or (4) in a way that accurately distinguishes the domain expert from group leader.

Participants can use either empirical or engineering techniques, and any source of signal or coded information available in these multimodal data resources. The Math Data Corpus is a well structured dataset for examining these particular prediction
challenges. Participants can focus on developing predictors of any one or more of the aims listed above. For evaluation purposes, each participant will be required in their workshop paper to describe replicable methods for achieving their reported level of predictive success with any one or more of the above aims,

(1) a clear quantitative summary of their method’s precise predictive capabilities, and
(2) a description of the scope of their predictive capabilities, if more than one objective is included.

Reporting is on the honor system. However, participants may be asked to present their data and techniques to the evaluation committee or by others at the workshop. They therefore should be prepared to reveal and defend the specifics of any prediction methods, techniques, or results presented.

Acknowledgments
The data described in this paper were collected as part of a study run by Rebecca Lunsford and Alex Arthur, which is available elsewhere [3]. For further details on the research infrastructure that was used to collect these data, see [1]. Rachel Coulston and Marissa Flecha-Garcia assisted with subject recruitment, preparation of problem sets, and data collection. Xiao Huang assisted with equipment setup and calibration. Funding by DARPA Contract No. NBCHD030010 and NSF Grant No. IIS-0117868 originally supported this prior research. Preparation of this dataset and related resources for the MMLA grand challenge series of events, and also preparation of the present paper, were supported by Inca Designs (http://www.incaadesigns.org/). This paper is an expanded and revised version of the one prepared for the Second Intl. MMLA Workshop.

References
Appendix A: Ground-Truth Scoring of Expertise and Problem Difficulty

1. Group 1, Session 1

Group: G1d1 (Males, High performing)  81.3% correct

Problem set: GEO set 2

<table>
<thead>
<tr>
<th>Prob#</th>
<th>Start-End</th>
<th>Answer initiate-report (TTS)</th>
<th>P Cor/Incor</th>
<th>Person Initiate Answer</th>
<th>Person Explain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A</td>
<td>0:23-1:23</td>
<td>MOI: 0:45-1:06</td>
<td>Correct:</td>
<td>S1</td>
<td>No</td>
</tr>
<tr>
<td>3A</td>
<td>3:07-4:38</td>
<td>MOI: 4:25-4:28</td>
<td>Correct:</td>
<td>S1</td>
<td>No</td>
</tr>
<tr>
<td>4A</td>
<td>4:42-6:10</td>
<td>MOI: 5:10-5:33</td>
<td>Incorrect:</td>
<td>S1</td>
<td>No</td>
</tr>
</tbody>
</table>

(technical break)

<table>
<thead>
<tr>
<th>Prob#</th>
<th>Start-End</th>
<th>Answer initiate-report (TTS)</th>
<th>P Cor/Incor</th>
<th>Person Initiate Answer</th>
<th>Person Explain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1B</td>
<td>7:26-8:10</td>
<td>MOI: 7:40-8:05</td>
<td>Correct:</td>
<td>S1</td>
<td>No</td>
</tr>
<tr>
<td>2B</td>
<td>8:14-9:06</td>
<td>MOI: 8:48-8:58</td>
<td>Correct:</td>
<td>S1</td>
<td>No</td>
</tr>
<tr>
<td>3B</td>
<td>9:10-11:44</td>
<td>MOI: 10:53-10:56</td>
<td>Incorrect:</td>
<td>S3</td>
<td>No</td>
</tr>
<tr>
<td>4B</td>
<td>11:46-15:48</td>
<td>MOI: 13:-49-14:22</td>
<td>Incorrect:</td>
<td>S3</td>
<td>S1</td>
</tr>
</tbody>
</table>

(technical break)

<table>
<thead>
<tr>
<th>Prob#</th>
<th>Start-End</th>
<th>Answer initiate-report (TTS)</th>
<th>P Cor/Incor</th>
<th>Person Initiate Answer</th>
<th>Person Explain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1C</td>
<td>16:55-19:10</td>
<td>MOI: 18:00-18:30</td>
<td>Correct:</td>
<td>S1</td>
<td>S3</td>
</tr>
<tr>
<td>2C</td>
<td>19:12-20:40</td>
<td>MOI: 20:07-20:12</td>
<td>Correct:</td>
<td>S1</td>
<td>S2</td>
</tr>
<tr>
<td>4C</td>
<td>23:11-26:26</td>
<td>MOI: 24:30-26:26</td>
<td>Correct:</td>
<td>S1</td>
<td>No</td>
</tr>
</tbody>
</table>

(technical break)

<table>
<thead>
<tr>
<th>Prob#</th>
<th>Start-End</th>
<th>Answer initiate-report (TTS)</th>
<th>P Cor/Incor</th>
<th>Person Initiate Answer</th>
<th>Person Explain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1D</td>
<td>29:32-30:33</td>
<td>MOI: 30:15-30:20</td>
<td>Correct:</td>
<td>S1</td>
<td>No</td>
</tr>
<tr>
<td>2D</td>
<td>30:37-32:11</td>
<td>MOI: 31:34-31:54</td>
<td>Correct:</td>
<td>S2</td>
<td>No</td>
</tr>
</tbody>
</table>
### Performance Summary: 13/16 correct

S1- 12 answers contributed (5 hard), 92% correct (Expert & Assigned Leader)  
S2- 2 answers contributed (1 hard), 100% correct  
S3- 2 answers contributed- 7 answers contributed (2 hard), 0% correct  

**Expertise** = number of easy problems solved or missed (+1 or -1 pt.); number of moderate problems solved or missed (+2 or -2 pts.); number of hard problems solved or missed (+3 or -3 pt.); number of very hard problems solved or missed (+4 or -4 pt.)

**TTS mean** = 79.9 (all)

#### 2. Group 1, Session 2

**Group: G1d2 (Males, High performing) 87.5% correct**

**Problem set: GEO set 1**

<table>
<thead>
<tr>
<th>Prob#</th>
<th>Start-End</th>
<th>Answer initiate-report (TTS)</th>
<th>Person Initiate Answer Person Explain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A</td>
<td>0:05-0:40</td>
<td>MOI: 0:27-0:32 (27)</td>
<td>Correct: S1</td>
</tr>
<tr>
<td>2A</td>
<td>0:41-1:46</td>
<td>MOI: 1:00-1:24 (43)</td>
<td>Correct: S1</td>
</tr>
<tr>
<td>4A</td>
<td>2:52-4:22</td>
<td>MOI: 3:55-4:10 (78)</td>
<td>Correct: S1</td>
</tr>
</tbody>
</table>

(technical break)

| 1B    | 5:00-7:07 | MOI: 5:40-6:37 (97)          | Incorrect: S3                         | No                                    |
| 2B    | 7:11-8:26 | MOI: 8:09-8:18 (77)          | Correct: S1                           | No                                    |
| 3B    | 8:32-11:08| MOI: 10:30-11:00 (148)       | Correct: S1                           | No                                    |
| 4B    | 11:15-12:21| MOI: 11:56-12:00 (45)        | Correct: S2                           | S1                                    |

(technical break)

| 1C    | 13:29-14:36| MOI: 13:56-14:03 (34)        | Correct: S3                           | S3                                    |
| 2C    | 14:39-16:48| MOI: 15:36-16:02 (83)        | Correct: S1                           | S2                                    |
| 3C    | 16:52-17:45| MOI: 17:37-17:40 (48)        | Correct: S1                           | No                                    |

(technical break)

Performance Summary: 14/16 correct

S1 - 12 answers contributed (6 hard), 92% correct (Expert) 23
S2 - 1 answers contributed (1 hard), 100% correct (Assigned leader) 4
S3 - 3 answers contributed - 3 answers contributed (1 hard), 66.7% correct 3

**Expertise**= number of easy problems solved or missed (+1 or -1 pt.); number of moderate problems solved or missed (+2 or -2 pts.); number of hard problems solved or missed (+3 or -3 pt.); number of very hard problems solved or missed (+4 or -4 pt.)

TTS= Total time to problem solution for all problems, correct and incorrect. Start of problem to end of problem submission.

Cumulative expertise score across both sessions:
S1- 20 + 23 = 43 total (dominant expert; assigned leader session 1)
S2- 5 + 4 = 9 total (assigned leader session 2)
S3- -7 + 3 = -4 total

D1:
1- S3 0, S2 0, S1 4
2- S3 0, S2 2, S1 6
3- S3 -3, S2 3, S1 6
4- S3 -4, S2 0, S1 4

D2:
1- S3 0, S2 0, S1 2
2- S3 0, S2 0, S1 8
3- S3 3, S2 0, S1 9
4- S3 0, S2 4, S1 4

Note on Group: Group solves problems collaboratively and works well together, although dominant student is by far the strongest. S3, the weakest one, is an active contributor and sometimes disagrees.

TTS mean=76.8 (all)

TTS Grand Mean=78.4 (Difficulty Levels- 55.1, 62.6, 83.3, 112.4)
3. Group 2, Session 1

Group: G2d1  (Females, Low performing)  37.5% correct

Problem set: GEO set 2

(GENERAL NOTE: Sometimes there’s no MOI because they just guess)

<table>
<thead>
<tr>
<th>Prob#</th>
<th>Start-End</th>
<th>Answer initiate-report (TTS)</th>
<th>P Cor/Incor</th>
<th>Person Initiate</th>
<th>Answer</th>
<th>Person Explain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A:</td>
<td>3:49-5:27</td>
<td>MOI: 4:54-5:20 (91)</td>
<td>Correct:</td>
<td>S5</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2A:</td>
<td>5:30-7:43</td>
<td>MOI: 6:22-6:33 (63)</td>
<td><strong>Incorrect</strong>:</td>
<td>S6</td>
<td>S6</td>
<td></td>
</tr>
<tr>
<td>3A:</td>
<td>7:45-14:32</td>
<td>MOI: 10:26-10:52 (127)</td>
<td><strong>Incorrect</strong>:</td>
<td>S4</td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

(technical break)

1B:  20:23-26:59  MOI: 25:58 (335)  **Incorrect**:  S5  No

(Note: No MOI range because they just guessed)

2B:  27:02-31:30  MOI: 27:41-28:11 (69)  **Incorrect**:  S6  No

3B:  31:35-35:06  MOI: 33:12-34:08 (153)  **Incorrect**:  S6  No

4B:  35:08-39:13  MOI: 36:57-37:21 (133)  **Incorrect**:  S6  S5

(technical break)


3C:  54:00-1:02:17  MOI: 55:50-1:00 (360)  **Incorrect**:  S5  No

(NOTE: Unplanned technical break during this problem, Then off-task discussion about writing pen angle; **Excise 56:45-58:55**)

4C:  1:02:18-1:06:52  MOI: 1:05:57 (219)  **Incorrect**:  S5 & S4  No

(Note: No MOI range because they just guessed)

(technical break)
1D: 1:07:52- 1:10:33   MOI: 1:09:43-1:10:22 (150) Correct: S6 No
2D: 1:10:35- 1:12:40   MOI: 1:11:34-1:12:32 (117) Correct: S5 No
(NOTE: Unplanned technical break during this problem, Excise: 1:14:34-1:15:34)

Performance Summary: 6/16 correct

S5- 5.5 answers contributed (hard), 54.5% correct (Assigned Leader) -2
S6- 9 answers contributed (hard), 33% correct -9
S4- 1.5 answer contributed (1.5 hard), 0% correct -5

Expertise= number of easy problems solved or missed (+1 or -1 pt.); number of moderate problems solved or missed (+2 or -2 pts.); number of hard problems solved or missed (+3 or -3 pt.); number of very hard problems solved or missed (+4 or -4 pt.)

TTS mean=186.5
4. Group 2, Session 2

Group: G2d2 (Females, Low performing) 50% correct

Problem set: GEO set 2

(GENERAL NOTE: Sometimes there’s no MOI because they just guess)

<table>
<thead>
<tr>
<th>Prob#</th>
<th>Start-End</th>
<th>MOI</th>
<th>Correct/Incor</th>
<th>Person Initiate</th>
<th>Answer</th>
<th>Person Explain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A:</td>
<td>0:23-1:46</td>
<td>1:17-1:32 (69)</td>
<td>Correct: S5</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4A:</td>
<td>12:53-16:00</td>
<td>13:22-14:18 (85)</td>
<td>Incorrect: S6</td>
<td>No</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(technical break)

| 1B:   | 17:17-18:56 | 18:18-18:40 (83) | Correct: S6 | No |
| 3B:   | 20:16-29:16 | 24:00-29:03 (527) | Correct: S6 | No |

(technical break)

| 3C:   | 40:47-46:49 | 45:47-45:49 (302) | Incorrect: S6 | No |
| 4C:   | 46:50-51:48 | 50:16-50:30 (220) | Incorrect: S6 | No |

(technical break)

| 1D:   | 52:19-53:56 | 53:29-53:49 (90) | Correct: S5 | No |
| 3D:   | 55:60-1:00:35 | 58:42-59:40 (220) | Incorrect: S6 | S4 |
Performance Summary: 8/16 correct

S5- 2 answers contributed (0 hard), 100% correct 2
S6- 11 answers contributed (7 hard), 36.4% correct  -17
S4- 3 answer contributed (1 hard), % correct  3 (Assigned Leader session 2)

Expertise= number of easy problems solved or missed (+1 or -1 pt.); number of moderate problems solved or missed (+2 or -2 pts.); number of hard problems solved or missed (+3 or -3 pt.); number of very hard problems solved or missed (+4 or -4 pt.)

Cumulative expertise score across both sessions:
S5  2 + -2 = 0 total  (Dominant expert; Assigned leader session 1)
S6  -9 + -17 = -26 total
S4  -5 + 3 = -2 total (Assigned leader session 2)

(Note: Extremely weak group; S6 took risks to initiate answers rather than guess, but did not have expertise)

D1:
1- S5 1, S6 1, S4 0
2- S5 2, S6 -6, S4 0
3- S5 -3, S6 0, S4 -3
4- S5 -2, S6 -4, S4 -2

D2:
1- S5 2, S6 2, S4 0
2- S5 0, S6 0, S4 0
3- S5 0, S6 -3, S4 3
4- S5 0, S6 -16, S4 0

Notes on Group: This is the only group without a strong expert to lead. They collaborate, affiliate, and are all engaged. However, they take a very long time to solve problems. They often reach an impasse, at which point S6 guesses at the answer and encourages the group to submit it. This is the only group that engages in guessing behavior, which consistently fails to solve the problem.

TTS mean=190.3

TTS Grand Mean= 188.4  (Difficulty Level- 139.9, 150.5, 297.8, 165.5)
5. Group 3, Session 1

Group: G3d1  (Female, High performing)  92.9% correct

Problem set: GEO set 2

Prob#  Start-End  Answer initiate-report (TTS)  P Cor/Incor  Person Initiate Answer Person Explain

1A:  NOTE: Missing data

2A:  NOTE: Missing data


4A: 2:53-6:18   MOI: 4:17-4:34   (101)  Incorrect: S7  No

(technical break)

1B  7:22-9:16   MOI: 8:37-9:01   (99)  Correct: S8  No


3B: 11:52-14:35  MOI: 13:42-14:23  (158)  Correct: S7  No


(technical break)


(Excise: 23:50-25:17 due to unexpected technical break)


(Excise: 31:44-32:31 for extended out of control laughing)

4C: 34:32-40:15  MOI: 36:57-38:18  (226)  Correct: S7  No

(technical break)

1D: 41:10-43:05  MOI: 42:22-42:52  (102)  Correct: S9  No

Performance Summary: 13/14 correct

S7 - 11 answers contributed (7 hard), 90.9% correct 23 (Assigned Leader, Dominant Expert)
S9 - 1 answers contributed (0 hard), 100% correct 1
S8 - 2 answer contributed (1 hard), 100% correct 5

Expertise= number of easy problems solved or missed (+1 or -1 pt.); number of moderate problems solved or missed (+2 or -2 pts.); number of hard problems solved or missed (+3 or -3 pt.); number of very hard problems solved or missed (+4 or -4 pt.)

TTS mean=137.5

6. Group 3, Session 2

Group: G3d2 (Females, High performing) 100% correct

Problem set: GEO set 1

<table>
<thead>
<tr>
<th>Prob#</th>
<th>Start-End</th>
<th>Answer initiate-report</th>
<th>TTS</th>
<th>P Cor/Incor</th>
<th>Person Initiate</th>
<th>Answer Person Explain</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4:14-5:46</td>
<td>MOI: 5:18-5:32</td>
<td>(78)</td>
<td>Correct: S8</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>9:41-12:22</td>
<td>MOI: 11:04-11:52</td>
<td>(131)</td>
<td>Correct: S7</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>32:12-34:37</td>
<td>MOI: 33:51-34:23</td>
<td>(131)</td>
<td>Correct: S7</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>34:38-36:07</td>
<td>MOI: 36:28-36:57</td>
<td>(139)</td>
<td>Correct: S7</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Performance Summary: 16/16 correct**

S9- 3 answers contributed (1 hard), 100% correct  6
S7- 12 answers contributed (7 hard), 100% correct  33  (Dominant expert)
S8- 1 answer contributed (0 hard), 100% correct  1  (Assigned Leader session 2)

**Expertise**= number of easy problems solved or missed (+1 or -1 pt.); number of moderate problems solved or missed (+2 or -2 pts.); number of hard problems solved or missed (+3 or -3 pt.); number of very hard problems solved or missed (+4 or -4 pt.)

**Cumulative expertise score across both sessions:**
S7- 23 + 33 = 56 total  (Dominant expert, assigned leader session 1)
S9- 1 + 6 = 7 total
S8- 5 + 1 = 6 total (Assigned leader session 2)

**D1:**
1- S7 1, S9 1, S8 1
2- S7 6, S9 0, S8 0
3- S7 12, S9 0, S8 0
4- S7 4, S9 0, S8 4

**D2:**
1- S7 1, S9 2, S8 1
2- S7 8, S9 0, S8 0
3- S7 12, S9 0, S8 0
4- S7 12, S9 4, S8 0

Notes on group: S7 is by far the dominant expert, although she remains very reserved throughout. S8 resents and tests her dominance a few times. S8 also initiates a lot of group laughter. S8 is socially dominant. All are engaged, and collaborating and working actively together.

**TTS mean=137.9**

**TTS Grand Mean=137.7  (Difficulty Level- 90.7, 140.7, 145, 169)**
7. Group 4, Session 1

Group: G4d1 (Male, Moderate performing) 62.5% correct

Problem set: GEO set 1

<table>
<thead>
<tr>
<th>Prob#</th>
<th>Start-End</th>
<th>Answer initiate-report (TTS)</th>
<th>P Cor/Incor</th>
<th>Person</th>
<th>Initiate</th>
<th>Answer</th>
<th>Person</th>
<th>Explain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A:</td>
<td>1:59-3:08</td>
<td>MOI: 2:29-3:00 (61)</td>
<td>Correct:</td>
<td>S10</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2A:</td>
<td>3:09-5:42</td>
<td>MOI: 5:00-5:18 (129)</td>
<td>Incorrect:</td>
<td>S10</td>
<td>S12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4A:</td>
<td>8:34-9:44</td>
<td>MOI: 9:20- 9:38 (64)</td>
<td>Correct:</td>
<td>S12</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(technical break)

| 1B:   | 11:21-13:11 | MOI: 12:14-12:36 (75) | Correct:    | S12    | No       |
| 3B:   | 15:37-18:33 | MOI: 17:42-18:00 (143) | Incorrect:  | S12    | No       |

(technical break)


(technical break)

| 1D:   | 32:09-33:19 | MOI: 32:42-33:04 (55)   | Correct:    | S12    | No       |
| 2D:   | 33:20-35:16 | MOI: 34:38-34:52 (92)   | Correct:    | S12    | No       |
| 4D:   | 37:56- 43:43 | MOI: 42:00-42:54 (238)  | Correct:    | S10    | S11      |
Performance Summary: 10/16 correct

S11- 2 answers contributed (2 hard), 0% correct (Assigned Leader) -7
S10- 4 answers contributed (1 hard), 50% correct 1
S12- 10 answer contributed (5 hard), 80% correct 12  (Dominant Expert)

Expertise= number of easy problems solved or missed (+1 or -1 pt.); number of moderate problems solved or missed (+2 or -2 pts.); number of hard problems solved or missed (+3 or -3 pt.); number of very hard problems solved or missed (+4 or -4 pt.)

TTS mean=103.8

8. Group 4, Session 2

Group: G4d2 (Male, High performing) 87.5% correct

Problem set: GEO set 2

<table>
<thead>
<tr>
<th>Prob#</th>
<th>Start-End</th>
<th>Answer initiate-report (TTS)</th>
<th>P Cor/Incor</th>
<th>Person Initiate Answer Person Explain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A:</td>
<td>0:28-1:49</td>
<td>MOI: 0:56-1:42 (74)</td>
<td>Correct:</td>
<td>S11</td>
</tr>
<tr>
<td></td>
<td>(technical break)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2B:</td>
<td>15:49-17:08</td>
<td>MOI: 16:25-17:00 (71)</td>
<td>Correct:</td>
<td>S10 &amp; S11</td>
</tr>
<tr>
<td></td>
<td>(technical break)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4C:</td>
<td>30:05-32:29</td>
<td>MOI: 31:12-31:54 (109)</td>
<td>Correct:</td>
<td>S12</td>
</tr>
<tr>
<td></td>
<td>(technical break)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1D:</td>
<td>33:19-36:36</td>
<td>MOI: 34:04-36:29 (190)</td>
<td>Correct:</td>
<td>S12</td>
</tr>
</tbody>
</table>
Performance Summary: 14/16 correct

S11- 3 answers contributed (hard), 66.7% correct 0
S10- 4 answers contributed (hard), 100% correct (Assigned Leader) 8
S12- 9 answer contributed (hard), 88.9% correct (Dominant Expert) 20

**Expertise** = number of easy problems solved or missed (+1 or -1 pt.); number of moderate problems solved or missed (+2 or -2 pts.); number of hard problems solved or missed (+3 or -3 pt.); number of very hard problems solved or missed (+4 or -4 pt.)

Cumulative expertise score across both sessions:
S11- -7 + 0 = -7 total (Assigned leader session 1)
S10- 1 + 8 = 9 total (assigned leader session 2)
S12- 12 + 20 = 32 total (Dominant Expert)

D1:
1- S12 3, S10 1, S11 0
2- S12 4, S10 -4, S11 0
3- S12 -3, S10 0, S11 -3
4- S12 8, S10 4, S11 -4

D2:
1- S12 2, S10 1, S11 1
2- S12 0, S10 4, S11 0
3- S12 6, S10 3, S11 3
4- S1212, S10 0, S11 -4

Notes on Group: S12 is dominant expert, and also very loud and socially dominant. He sometimes teases others, calls them derogatory names, and adopts “punk” behavior trying to be cool. All are engaged and collaborating actively. S10 is quiet, sometimes nonverbally gesturing information. S11 usually does not understand the problems, and sometimes impulsively offers a wrong answer as a guess and then encourages the group to adopt it.

TTS mean=123.5

TTS Grand Mean=113.7 (Difficulty Level- 92.4, 107.6, 115, 139.5)
9. Group 5, Session 1

Group: G5d1  (Male, High performing)  93.8% correct

Problem set: GEO set 2

<table>
<thead>
<tr>
<th>Prob#</th>
<th>Start-End</th>
<th>Answer initiate-report (TTS) P Cor/Incor</th>
<th>Person Initiate</th>
<th>Answer</th>
<th>Person Explain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A:</td>
<td>1:15-4:54</td>
<td>MOI: 3:24-4:37 (202) Correct:</td>
<td>S15</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>(Excise: 2:54-3:13; Unexpected technical interruption)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4A:</td>
<td>15:22-19:28</td>
<td>MOI: 17:21-17:46 (144) Incorrect:</td>
<td>S14</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>(technical break)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1B:</td>
<td>21:30-23:53</td>
<td>MOI: 23:01-23:46 (136) Correct:</td>
<td>S14</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2B:</td>
<td>23:54-26:07</td>
<td>MOI: 25:03-25:45 (111) Correct:</td>
<td>S14 &amp; S15</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>3B:</td>
<td>26:08-31:01</td>
<td>MOI: 28:48-30:29 (261) Correct:</td>
<td>S14</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>(technical break)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3C:</td>
<td>47:24-51:19</td>
<td>MOI: 50:03-51:10 (226) Correct:</td>
<td>S14</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>(technical break)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2D:</td>
<td>57:36-1:00:07</td>
<td>MOI: 57:58-1:00:02 (156) Correct:</td>
<td>S13</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>3D:</td>
<td>1:00:08-1:02:05</td>
<td>MOI: 1:01:03-1:01:39 (91) Correct:</td>
<td>S14</td>
<td>S15</td>
<td></td>
</tr>
<tr>
<td>4D:</td>
<td>1:02:06-1:05:15</td>
<td>MOI: 1:03:35-1:04:12 (126) Correct:</td>
<td>S13</td>
<td>S13</td>
<td></td>
</tr>
</tbody>
</table>
Performance Summary: 15/16 correct

S13 - 3 answers contributed (1 hard), 100% correct (Assigned Leader) 7
S15 - 2 answers contributed (1 hard), 100% correct 5
S14 - 11 answer contributed (6 hard), 90.9% correct (Dominant Expert) 20

**Expertise**: number of easy problems solved or missed (+1 or -1 pt.); number of moderate problems solved or missed (+2 or -2 pts.); number of hard problems solved or missed (+3 or -3 pt.); number of very hard problems solved or missed (+4 or -4 pt.)

Notes on group: This group did not actively collaborate on solving problems. S14 mentored S13. S14 was the dominant expert. During session #1, S13 was leader, although the weakest in expertise. S14 and S15 held back, and let S13 try to solve the problems, which he struggled with. S14 actively mentored S13 and would give him prompts for next steps to perform. In the process, S14 would not take over and do the problems himself, but rather let S13 do them. S13 consistently tries very hard. S15 mainly observed. In session 2, S14 continues mentoring S13, spending a lot of time with corrections and explanations.

**TTS mean=188.9**
10. Group 5, Session 2

Group: G5d2 (Male, High performing) 100% correct

Problem set: GEO set 1

<table>
<thead>
<tr>
<th>Prob#</th>
<th>Start-End</th>
<th>MOI:</th>
<th>%</th>
<th>Correct</th>
<th>Person</th>
<th>Initiate</th>
<th>Answer</th>
<th>Person Explain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A</td>
<td>0:27-2:14</td>
<td>1:40-2:06</td>
<td>(99)</td>
<td>Correct:</td>
<td>S13</td>
<td>No</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(technical break)

| 1B    | 10:37-16:30 | 14:47-16:19 | (342) | Correct: | S14    | No       |        |                |

(technical break)


(technical break)

| 1D    | 42:06-43:56 | 43:26-43:46 | (100) | Correct: | S15    | No       |        |                |
| 2D    | 43:57-45:27 | 44:46-45:00 | (63)  | Correct: | S14    | No       |        |                |

Performance Summary: 16/16 correct
S14- 12 answers contributed (7 hard), 100% correct (Dominant Expert, Asgd. Leader) 32
S15- 3 answers contributed (1 hard), 100% correct 7
S13- 1 answer contributed (hard), 100% correct 1

**Expertise**= number of easy problems solved or missed (+1 or -1 pt.); number of moderate problems solved or missed (+2 or -2 pts.); number of hard problems solved or missed (+3 or -3 pt.); number of very hard problems solved or missed (+4 or -4 pt.)

**Cumulative expertise score across both sessions:**
S14- 20 + 32 = 52 total (dominant expert; assigned leader session 2)
S15- 5 + 7 = 12 total
S13- 7 + 1 = 8 total (assigned leader session 1)

D1:
1- S14 2, S15 1, S13 1
2- S14 5, S15 0.5, S13 1
3- S14 12, S15 0, S13 0
4- S14 0, S15 4, S13 4

D2:
1- S14 2, S15 1, S13 1
2- S14 6, S15 2, S13 0
3- S14 12, S15 0, S13 0
4- S14 12, S15 4, S13 0

TTS mean=173.7

TTS Grand Mean=181.3 (Difficulty level- 152.9, 181, 167.5, 223.8)
## 11. Group 6, Session 1

**Group:** G6d1 (Females, High performing)  87.5% correct  

**Problem set:** GEO set 2

<table>
<thead>
<tr>
<th>Prob#</th>
<th>Start-End</th>
<th>Answer initiate-report (TTS) P</th>
<th>Cor/Incor</th>
<th>Person Initiate</th>
<th>Answer Person Explain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A:</td>
<td>0:34-1:44</td>
<td>MOI: 1:31-1:41 (67)</td>
<td>Correct:</td>
<td>S16</td>
<td>No</td>
</tr>
<tr>
<td>4A:</td>
<td>4:59-7:00</td>
<td>MOI: 6:18-7:00 (121)</td>
<td>Correct:</td>
<td>S17 (NOTE: S16 <em>tries to usurp leadership</em> role, but S17 corrects her &amp; repeats answer)</td>
<td>No</td>
</tr>
</tbody>
</table>

(technical break)

| 1B:   | 7:36-8:48 | MOI: 7:58-8:40 (64)            | Correct:  | S18             | No                   |
| 2B:   | 8:49-10:58| MOI: 10:43-10:52 (123)         | Correct:  | S17             | No                   |

(technical break)

| 1C:   | 16:04-17:47| MOI: 17:09-17:21 (77)          | Correct:  | S16             | S16                  |
| 3C:   | 20:03-21:29| MOI: 20:50-21:22 (79)          | Correct:  | S18             | No                   |

(technical break)

| 1D:   | 25:18-26:13| MOI: 26:01-26:12 (54)          | Correct:  | S18             | No                   |
| 2D:   | 26:14-27:07| MOI: 26:45-26:52 (38)          | Incorrect:| S16             | No                   |
Performance Summary: 14/16 correct

S17- 5.5 answers contribute (4.5 hard), 100% correct (Assign leader + Expert) 18.5
S16- 3.5 answers contributed (0.5 hard), 71.4% correct 1.5
S18- 7 answers contributed (2 hard), 87.5% correct 8

Expertise= number of easy problems solved or missed (+1 or -1 pt.); number of moderate problems solved or missed (+2 or -2 pts.); number of hard problems solved or missed (+3 or -3 pt.); number of very hard problems solved or missed (+4 or -4 pt.)

Cumulative expertise score across both sessions:
S17- 18.5 + 18 = 36.5 total (dominant expert; assigned leader session 1)
S18- 8 + 1 = 9 total
S16- 1.5 + 5 = 6.5 total (assigned leader session 2)

D1:
1- S18 2, S16 2, S17 0
2- S18 4, S16 -2, S17 2
3- S18 6, S16 1.5, S17 4.5
4- S18 -4, S16 0, S17 12

D2:
1- S17 2, S18 1, S16 1
2- S17 6, S18 2, S16 0
3- S17 6, S18 6, S16 0
4- S17 4, S18 -8, S16 4

Notes on Group: S17 is dominant expert. Group works actively and collaboratively together. S16 is most socially dominant.

TTS mean= 106.8
12. Group 6, Session 2

Group: G6d2  (Females, High performing)  87.5% correct

Problem set: GEO set 2

<table>
<thead>
<tr>
<th>Prob#</th>
<th>Start-End</th>
<th>MOI: Start-End</th>
<th>Problem</th>
<th>Correct/Incorrect</th>
<th>Person Initiate</th>
<th>Answer</th>
<th>Person Explain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A</td>
<td>0:12-0:54</td>
<td>0:45-0:54</td>
<td>(42)</td>
<td>Correct:</td>
<td>S17</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2A</td>
<td>0:55-2:28</td>
<td>1:53-2:02</td>
<td>(67)</td>
<td>Correct:</td>
<td>S17</td>
<td>S18</td>
<td></td>
</tr>
<tr>
<td>3A</td>
<td>2:29-3:41</td>
<td>3:18-3:33</td>
<td>(64)</td>
<td>Correct:</td>
<td>S17</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>4A</td>
<td>3:42-5:32</td>
<td>4:04-4:20</td>
<td>(38)</td>
<td>Incorrect:</td>
<td>S18</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(technical break)</td>
</tr>
<tr>
<td>1B</td>
<td>6:56-7:51</td>
<td>7:31-7:38</td>
<td>(42)</td>
<td>Correct:</td>
<td>S18</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2B</td>
<td>7:52-9:16</td>
<td>8:46-9:08</td>
<td>(76)</td>
<td>Correct:</td>
<td>S18</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>3B</td>
<td>9:17-11:08</td>
<td>10:36-11:03</td>
<td>(106)</td>
<td>Correct:</td>
<td>S18</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(technical break)</td>
</tr>
<tr>
<td>1C</td>
<td>15:16-17:56</td>
<td>16:24-17:05</td>
<td>(109)</td>
<td>Correct:</td>
<td>S17</td>
<td>S16</td>
<td></td>
</tr>
<tr>
<td>3C</td>
<td>19:10-20:46</td>
<td>20:02-20:35</td>
<td>(85)</td>
<td>Correct:</td>
<td>S18</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>4C</td>
<td>20:47-23:42</td>
<td>23:05-23:36</td>
<td>(169)</td>
<td>Correct:</td>
<td>S16</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(technical break)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Note: Problem was changed from 4x10cm to 3x5 feet; correct answer 15 sq. ft.)</td>
</tr>
<tr>
<td>2D</td>
<td>25:59-26:48</td>
<td>26:31-26:42</td>
<td>(43)</td>
<td>Correct:</td>
<td>S17</td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

Performance Summary:  14/16 correct
Appendix B: Written Representation Coding

Kejun Xu, updated April 30, 2008 (revised by S. Oviatt & Adrienne Cohen, Dec. & May 2013)

Linguistic (L)
Definition: Words, abbreviations, acronyms, individual letters all were counted as linguistic content.
For example: “G”, “answer”, “msg”, “she’s”, “in-depth” each counted as one linguistic item.
Note: If the content is a label for a diagram, indicate that in the comments section (i.e. “Label D1”).

Numeric (N)
Definition: It denotes a number or a system of numbers.
For example: 5/12, 78%, 0.2, 9, etc.
Notes: 5/12 is counted as two separate numbers, whereas 78% is one number and one symbol. If the content is a label for a diagram: indicate that in the comments section (i.e. “Label D1”).

Symbolic (S)
Definition: A conventional sign or symbol relating to operations, quantities, relations, etc. For example, computational symbols (+, %, π), gestural symbols (arrows, check mark, cross outs), and other conventional symbols ($, * ) were counted in this category.
For example: “5/12” counted as two numbers and one symbol.
“78%” counted as one number, one symbol.
“$59” counted as one number, one symbol
“a + b = c” counted as 3 linguistic, 2 symbols
Note: Grammatical marks that did not pertain to content meaning (such as dashes within words, periods, colons, semi-colons) were NOT counted in this category. However, marks that did pertain to content meaning (e.g. colon between 6 and 7.5 to represent the ratio “6:7.5”) were coded. Small dots to indicate multiplication were not coded, because they usually fall within a <10 pt. minimum threshold for filtering out spurious digital ink.

Also, simple individual marks for emphasis (underlining, encircling) were counted separately as “Elemental marks” (see below). If the content is a label for a diagram, indicate that in the comments section (i.e. “Labels D1”).

Diagram (D)
Definition: A drawing shows a spatial figure, geometric arrangement, usually with relations among parts. A diagram is composed of a set of marks or symbols arranged in a meaningful way. Often content domains have conventional diagrams associated with them (e.g., Venn, punnett square) For example:

Note: Left example was 1 diagram; right example was 1 diagram, 1 word.
(any labels or marks connecting labels with a diagram are coded separately as L (linguistic), N (numeric), and M (elemental mark).
NOTE: Symbols within a diagram that refine its meaning, such as diameter line or right angle to indicate 90 degree angle were counted as part of the basic diagram, not as separate symbols.

Mark number of whole diagrams made during a problem with D1, D2 etc. separately from diagrammatic elements coded with timings. If user returns to elaborate or correct a diagram, it still is coded by same diagram number.

(D0 indicates diagrams that aren’t analyzed, due to being outside problem sets. All diagrams constructed within the time bounds of problem were coded. Task-irrelevant diagrams were coded “TI,” followed by typical diagrammatic coding.

Diagram Sub-coding:
In addition to its “D” diagram designation, diagramming should be coded for the following 2 characteristics, which may apply and involve adding a coding acronym:

(1) Degree of Complexity—
Low (1 individual geometric shape, such as an individual rectangle) (LC)

Moderate (2 or 3 combined geometric shapes, such as 3 circles in Venn diagram) (MC)

High (4 or more combined geometric shapes) (HC)

NOTE: A single geometric shape can be a circle, square, rectangle, etc. It also can be an individual line (e.g., to indicate diameter). The grid (diagram above, left) is composed of six lines, which would be coded HC. A punnett square is similarly composed of 6 lines (HC)

Note: Count number of geometric shapes as 1 apiece, presence of labeling as 1, presence of 3D as 1 (Note: Right diagram above would count as 3 points for two circles and added labeling (MC); A square would count as 1 (LC); a 3-D square would count as 2, geometric shape with 3D added (MC); a 3-D cone with a separate line for diameter and labeling added would count as 4 (HC)

(2) 3-dimensional or not (i.e., 2D); this is included under content meaning and only if “3-D” (otherwise if left blank, assumed as 2-D) (3-D)
Example coding:
D2-HC-DIS A Cone (3D) – diameter added later
(second diagram drawn, 3-D cone of high complexity, likely with diameter/height lines and labeling; in this case, the initial diagrammatic marks were made, the diagram abandoned for other writings, then the author came back and added the diameter to the cone diagram)
Example coding:
D2-HC-DIS CC Cone (3D) – radius line changed to diameter line later
(second diagram drawn, 3-D cone of high complexity, with radius line and labeling; in this case, the initial diagrammatic marks were made, the diagram abandoned for other writings, then the author came back and changed/corrected the radius to a diameter line)

Elemental Marks (M)
Individual marks, often for emphasis or highlighting items or relations (e.g., circle, underline), or for indicating the relation between two things (line connecting two items).

Task-irrelevant Content (TI)
In addition to its basic representational coding, any content written during a problem interval that is not related to problem content is coded with start and end time and marked as “task irrelevant” or off-task writing (coding starts with “TI” then follows normal coding scheme).

Written Content Outside of Problem Intervals
Any writing that occurs outside of the bounds of one of the 16 defined problem intervals (i.e., between start and end of the problem is not coded or analyzed. A problem starts when the leader says “Display new problem” and ends when the leader asks for the next problem.

**Disfluencies:** In addition to its basic representational coding, writing will be coded if it involves any of the following types of disfluencies: (DIS)

1. **Content correction disfluencies** - The meaning of content was changed/corrected by adding, deleting, or deleting and changing representation (e.g., simply cross-out, or cross out and re-write next to or over something). Both of these considered only one content disfluency. Additional examples included insertion caret with content added, adding content to the end, etc.

Corrections could include misspellings that are corrected, such as changing “sroll” to “scroll,” symbols that were corrected (e.g., radius to diameter mark), diagrams that were corrected (e.g., circle to oval), numbers that were corrected (6 to 60), etc.

Note: Any diagrams that were revisited after doing other work that did not involve the same diagram or its labels were marked as DIS-A if the user simply added or elaborated more details. However, it was coded as DIS CC if the meaning of the content was corrected/changed (eliminating part, changing part, or adding something that corrected the meaning rather than just elaborating existing meaning (e.g., extending radius line into diameter line).

2. **Incremental diagram elaborations:** When a user gradually continued refining a diagram in different bursts of time, with other content written in between (formulas, etc.) that were not the same diagram or its labels, then this was coded as DIS-A

3. **Retracing** - Going back to a previously written ink mark and retracing over it one or more times, or aesthetic touch-ups of previous ink (Note what content was retraced; IF CONTENT IS CHANGED, IT IS A CONTENT DISFLUENCY DIS CC) (DIS-R)

3. **Misspelled words or malformed symbols** - Unresolved misspellings or malformed symbols that are never corrected, such as “w” or pi missing a leg, “fl” (instead of “ft”), etc. (DIS-S) This included unconventional abbreviations, such as “f” instead of “ft” for feet.

Note: Some disfluencies may overlap, particularly with misspellings (i.e. content correction of one misspelling only to make a second error). When this happens, both must be coded as separate events so an accurate count of total disfluencies can be obtained.
Appendix C: Student Unique Identification Coding

“S#”=student’s number  
“G#D#”=group and session number  
“P#” = Problem set number  
“F/M”=gender  
“L/NL”=leader/non-leader  
“E/NE”=expert/non-expert

Group 1:
S1_G1D1P2_MLE/S1_G1D2P1_MNLNE
S2_G1D1P2_MNLNE/S2_G1D2P1_MLNE
S3_G1D1P2_MNLNE/S3_G1D2P1_MNLNE

Group 2:
S4_G2D1P1_FLNE/S4_G2D2P2_FNLNE
S5_G2D1P1_FNLNE/S5_G2D2P2_FNLNE
S6_G2D1P1_FNLNE/S6_G2D2P2_FNLNE

Group 3:
S7_G3D1P2_FNE/S7_G3D2P1_FNLNE
S8_G3D1P2_FNLNE/S8_G3D2P1_FNLNE
S9_G3D1P2_FNLNE/S9_G3D2P1_FNLNE

Group 4:
S10_G4D1P1_MNLNE/S10_G4D2P2_MNLNE
S11_G4D1P1_MNLNE/S11_G4D2P2_MNLNE
S12_G4D1P1_MNLNE/S12_G4D2P2_MNLNE

Group 5:
S13_G5D1P2_MNLNE/S13_G5D2P1_MNLNE
S14_G5D1P2_MNLNE/S14_G5D2P1_MNLNE
S15_G5D1P2_MNLNE/S15_G5D2P1_MNLNE

Group 6:
S16_G6D1P1_FNLNE/S16_G6D2P2_FNLNE
S17_G6D1P1_FNLNE/S17_G6D2P2_FNLNE
S18_G6D1P1_FNLNE/S18_G6D2P2_FNLNE
Appendix D: Problem Content for Problem Sets 1 and 2

Problem Set 1:

**Problem #1**

Randall's skateboard shop made $1350 on Monday. At the end of the day, he subtracted out his employee Aaron's wages and found he had $1320 left. If Aaron makes $10/hour, how many hours did he work?

*Answer:* Aaron worked 3 hours

*Example Solution:*

\[
\begin{align*}
$1350 - \text{wages} &= $1320, \text{wages} = $30 \\
30 / 10 &= 3 \text{ hours}
\end{align*}
\]

**Problem #2**

Ruth's 2-story house is 65 feet wide and 25 feet front to back. How many square feet of floor space are there in her house?

*Answer:* 3250 ft²

*Example Solution:*

\[\text{area of a rectangle} = (\text{length})(\text{width})\]

Area of one floor = 65 ft x 25 ft = 1625 ft²
Both floors = 2 x 1625 ft² = 2 x 3250 ft²

**Problem #3**

Susanna is making a Christmas tablecloth for a round table that is 4 feet in diameter. She wants the tablecloth to hang 1 foot down from the edge. How much material does she need?

*Answer:* 28.27 square feet (9 π)

*Example Solution:*

\[
\begin{align*}
\text{area of a circle} &= \pi r^2 \\
\text{Area} &= \pi (4/2+1)^2 = 9 \pi \text{ sq ft} \\
&= 28.27 \text{ sq ft}
\end{align*}
\]

**Problem #4**

Bruce's kite is stuck in a tree. If there's 50 feet of string from Bruce to the kite, and Bruce is 30 feet away from the tree, how high up is the kite stuck?

*Answer:* 40 feet

*Example Solution:*

\[
\begin{align*}
c^2 &= a^2 + b^2 \\
(\text{Height})^2 &= (50)^2 - (30)^2 = 1600 \\
\text{Height} &= 40 \text{ feet}
\end{align*}
\]
Problem Set 1 (cont.):

Problem # 5

Gabe got his little brother a soccer ball when he made the school team. If the radius of the ball is 5 inches, what size box, in cubic inches, would he need to package the soccer ball?

Answer: 1000 cubic inches

Example Solution:

Volume of a rectangular prism = (length)(width)(height)

volume = (10 in)(10 in)(10 in)
= 1000 cubic inches

Problem # 6

Leila is making paper cut-outs for her scrapbook. She has just cut out a blue regular triangle. If the perimeter is 9 inches, what is the height of the paper triangle?

Answer: 2.6 inches

Example Solution:

\[ c^2 = a^2 + b^2 \]

\[ (Height)^2 = (3)^2 - (1.5)^2 \times 6.75 \]

Height = 2.6 inches

Problem # 7

The diameter of Hannah’s head is 10 inches and she has a bandanna that’s 20 by 20 inches square. If she folds the bandanna in half diagonally, will it fit around her head?

Answer: No, it’s too small.

Example Solution:

Circumference of a circle = \(2\pi\) (radius)

Circumference = \(2\pi\) (5 inches)
= 31.41 inches

\[ c^2 = a^2 + b^2 \]

(Bandanna diagonal) = \((20)^2 + (20)^2 = 800\)

Diagonal = 28.28 inches

Problem # 8

Sally has a clock which is 20 cm in diameter. When the clock strikes 6 o’clock, what is the area to the right of the clock hands?

Answer: 157.08 square centimeters (50\(\pi\))

Example Solution:

area of a circle = \(\pi r^2\)

Area of the clock = \(\pi (10)^2 = 314.16\) sq cm
Right side of the clock = \(\frac{1}{4}\) (314.16 sq cm)
= 157.08 sq cm
Problem Set 1 (cont.):

**Problem #9**

Twenty-five feet of copper tubing weighs 12 pounds. How much will 150 feet of the same tubing weigh?

*Answer: 72 pounds*

*Example Solution:*

\[
\frac{25}{12} = \frac{150}{x}
\]

\[
25 \times (x) = 12 \times 150
\]

\[
x = 72
\]

**Problem #10**

Tim is drawing a trapezoid for his Geometry class. The parallel sides are 11 and 15 cm long and it has an area of 130 cm². What is the height of the trapezoid?

*Answer: 10 centimeters*

*Example Solution:*

area of a trapezoid = \(\frac{1}{2} \times \text{base}_1 + \text{base}_2 \times \text{height}\)

\[\text{height} = \frac{2 \times \text{area}}{\text{base}_1 + \text{base}_2}\]

\[\text{Height} = \frac{2 \times 130}{11 + 15} = 10 \text{ cm}\]

**Problem #11**

Daniel is building a half pipe so he can do bike tricks. If he wants the radius of the pipe to be 16 feet, how long should the plywood be for the curved inner area?

*Answer: 50.27 feet\(\cdot\pi\) (16\pi)*

*Example Solution:*

Circumference of a circle = \(\pi \times \text{radius}\)

Circumference of \(\frac{1}{2}\) circle = \(\pi \times \text{radius}\)

Circumference = \(\pi \times 16\) feet

= 50.27 feet

**Problem #12**

Mary is buying pizza for her friends. The pizza place sells 12-inch pizzas for $6 and 16-inch pizzas for $9. Which option would give the kids the most pizza for their money?

*Answer: The 16-inch pizzas.*

*Example Solution:*

area of a circle = \(\pi \times \text{radius}^2\)

Area of the 12-inch pizza = \(\pi \times (6)^2 = 113.1\) sq in

Cost per square inch = $6.00/113.1 = $0.05

Area of the 16-inch pizza = \(\pi \times (8)^2 = 201.1\) sq in

Cost per square inch = $9.00/201.1 = $0.045
Problem Set 1 (cont.):

Problem #13 3D

A bus traveled 120 miles in 2 1/2 hours. How far will it travel in 8 hours?

Answer: 384 miles

Example Solution:

\[
\frac{120}{2.5} = \frac{x}{8}
\]

2.5 (x) = 120(8)

x = 384

Problem #14 2D

John made a round cake for his party and put a candle in the center. He cut a piece of cake starting from the candle. If the straight sides of the slice are 5 inches each, what is the area of the cake top?

Answer: 78.54 square inches

Example Solution:

Area of a circle = \( \pi r^2 \)

Area of the cake = \( \pi (5)^2 = 78.54 \) sq in

Problem #15 3D

Gary's back yard is 15 feet by 40 yards. If his mower is 3 feet wide, and he can mow 10 yards per minute, how long will it take him to mow the yard?

Answer: 20 minutes

Example Solution:

5 rows at 4 minutes per row = 5 x 4 = 20 minutes

Problem #16 4D

Sarah is making a pyramid lamp to decorate her room. It has a square base that is 12 inches on each side and a height of 18 inches. How much cardboard does she need for each triangular side?

Answer: 113.84 square inches

Example Solution:

Area of a triangle = \( \frac{1}{2} \) (base)(altitude)

area = \( \frac{1}{2} \) (12 in)(18.97 in) = 113.84 square inches
Problem Set 2:

**Problem 1A**

A living room is 18 ft. long and 15 ft. wide, with a 12 ft. arched ceiling. How many square feet of gold carpeting are needed to cover the floor space?

*Answer:* 270 square feet

*Example Solution:*

\[
\text{area of a rectangle} = (\text{length})(\text{width})
\]

\[
\text{Area} = (18 \text{ feet})(15 \text{ feet}) = 270 \text{ square feet}
\]

**Problem 2A**

Thad has a new silver kite shaped like a rhombus. If the diagonals are 7 cm wide and 12 cm long, what is the surface area of his kite?

*Answer:* 42 cm²

*Example Solution:*

\[
\text{area of a rhombus} = \frac{1}{2}(\text{diagonal}_1)(\text{diagonal}_2)
\]

\[
\text{Area} = \frac{1}{2}(12\text{cm})(7\text{cm}) = 42\text{cm}^2
\]

**Problem 3A**

Josh made a party hat shaped like a triangle. If it has a base of 16 inches and a height of 7 inches, what is the surface area of the front of it?

*Answer:* 56 square inches

*Example Solution:*

\[
\text{area of a triangle} = \frac{1}{2}(\text{base})(\text{altitude})
\]

\[
\text{area} = \frac{1}{2}(16\text{in})(7\text{in}) = 56 \text{ square inches}
\]

**Problem 4A**

Laurie and Zak are knitting baby booties, and have figured that a ball of yarn with a 3-inch diameter is required for each bootie. If they have a ball of yarn with a diameter of 1 foot, how many booties will it yield?

*Answer:* 64 booties

*Example Solution:*

\[
\text{volume of sphere} = \frac{4}{3}\pi r^3
\]

\[
\text{Amount for one bootie} = \frac{4}{3}\pi (3\text{in})^3 = 113 \text{ in}^3
\]

\[
\text{Amount in large ball} = \frac{4}{3}\pi (12\text{in})^3 = 7234 \text{ in}^3
\]

\[
\text{Total booties possible} = \left(\frac{1\text{ bootie}}{113}\right) 7234 = 64 \text{ booties}
\]
Problem 1B

After running a Hawaiian ice stand, Sally took the $48 she made that day and subtracted the $10 she made in tips. She then divided the remaining money by the $2 cost of the Hawaiian ices and found the number of ices she had sold. How many did she sell?

Answer: 19 ices

Example Solution:

\[
\frac{48 - 10}{2} = x
\]

x = 19 ices

Problem 2B

Four boys are organizing a high school car wash as a fund raiser, and they need to pass out 360 flyers total. Ed will distribute 60 flyers, and Paul will pass out the same number. Dan and Dave, twin boys who do everything identically, will each distribute the same number. How many flyers will each boy pass out?

Answer: Paul does 60, the twins do 120 each.

Example Solution:

Paul = Ed = 60
Twin A = Twin B = x
60 Ed + 60 Paul + 2x = 360
2x = 360 − 120 = 240
x = 120

Problem 3B

You want to paint your bedroom this summer, and you know one gallon of paint covers about 450 square feet. How many gallons of paint do you need to paint just the walls of your bedroom, if the floor is 12 by 15 feet and the walls are 7 feet high?

Answer: One Gallon (exactly 0.84 gallons)

Example Solution:

7’ × 12’ (two walls like this)
and 7’ × 15’ (two walls like this)
The square footage is:
2(7x12) + 2(7x15) = 378 sq ft,
which can be covered by 1 gallon of paint.

Problem 4B

The largest tepee in the United States belongs to the Crow nation, Native Americans who live on the Great Plains. It is 43 feet high and 42 feet in diameter. What is its volume to the nearest cubic foot?

Answer: ~19,858 cubic feet

Example Solution:

\[
\text{volume of cone } = \frac{1}{3}(\text{area of the base})(\text{height})
\]

\[
\text{area of a circle } = \pi r^2
\]

Area of the base = \(\pi \times 21^2 = 1385.45\) sq ft
Volume of the tepee = \(\frac{1}{3}(1385.45\) sq ft)(43 ft) ≈ 19,858 cubic feet
Problem Set 2 (cont.):

Problem 1C

The area of a circular plant stand is $36\pi$ cm$^2$. How many centimeters is its diameter, and will a potted plant 10 cm in diameter fit on this stand?

**Answer:** 12 cm, and yes, it fits.

**Example Solution:**

\[
\text{area of a circle} = \pi r^2
\]

\[
\text{radius} = \sqrt{\frac{36\pi}{\pi}} = 6 \text{ cm}
\]

\[
\text{diameter} = 2r = 2(6) = 12 \text{ cm}
\]

Problem 2C

Matt is using a recipe that normally feeds his family, but is increasing it to cook enough for the whole lab. If the ratio of hot dogs to cups of chili for his family is 6 to 7.5, then for the lab how many hot dogs should he cook to go with 45 cups of chili?

**Answer:** 36 hot dogs

**Example Solution:**

\[
\frac{6}{7.5} = \frac{x}{45}
\]

\[
45(0.8) = x
\]

\[
x = 36
\]

Problem 3C

A hamster wheel is 10 cm in diameter. If the hamster runs so the wheel turns 100 revolutions, then how far did the hamster run?

**Answer:** 3141.6 cm

**Example Solution:**

\[
\text{Circumference of a circle} = 2\pi \text{ (radius)}
\]

\[
\text{Radius} = \frac{1}{2} \text{ diameter} = \frac{1}{2} \times 10 \text{ cm} = 5 \text{ cm}
\]

\[
\text{Distance run} = 100 \times (2\pi(5\text{cm})) = 3141.6 \text{ cm}
\]

Problem 4C

The Transamerica building in San Francisco is a pyramid 800 feet tall, with a square base 149 feet on each side. What is its volume in cubic feet?

**Answer:** 5,920,266.67 cubic feet

**Example Solution:**

\[
\text{volume of pyramid} = \frac{1}{3}(\text{area of the base})(\text{height})
\]

\[
\text{area of a rectangle} = (\text{length})(\text{width})
\]

\[
\text{Area of the base} = 149^2 = 22201 \text{ sq ft}
\]

\[
\text{Volume of the building} = \frac{1}{3}(22201\text{ sq ft})(800\text{ ft}) = 5,920,266.67 \text{ cubic feet}
\]

Aloha Geometry Text - Page 61, #4

[answer from book]
Problem 1D

Sarah is planting a spring garden with marigolds in a cedar planter box shaped like a parallelogram. If the base of the planter is 4 cm and the height is 10 cm, what area is available for planting?

*Answer:* 40 square centimeters

*Example Solution:*

\[
\text{area of a parallelogram} = (\text{base})(\text{height})
\]

Area = (4 cm)(10 cm)  
= 40 square centimeters

Problem 2D

Find the circumference of a multi-colored beach ball with a radius of 3.6 inches.

*Answer:* 22.62 inches

*Example Solution:*

\[
\text{circumference of a circle} = 2\pi
\]

\[2\pi (3.6) = 22.62 \text{ inches}\]

Problem 3D

Fred wants to set up a salt water aquarium to fully cover one wall of his bedroom that is 9 ft. long and 7 ft. high. He wants the depth to be just 1 ft. How many cubic feet of water will he need to fill the aquarium?

*Answer:* 63 cubic feet

*Example Solution:*

\[
\text{Volume of a rectangular prism} = (\text{length})(\text{width})(\text{height})
\]

\[\text{volume} = (9 \text{ ft})(1 \text{ ft})(7 \text{ ft})\]

Problem 4D

Bart is paving a walkway in front of the library using a large steamroller. The steamroller’s cylindrical wheel is 7 feet tall and 5 feet wide. How many square feet does a single revolution of the wheel cover?

*Answer:* 110 square feet

*Example Solution:*

\[
\text{Circumference of a circle} = 2\pi (\text{radius})
\]

\[\text{lateral area} = (\text{circumference})(\text{height})
\]

\[= (2\pi)(h)
\]

\[= (2\pi*3.5 \text{ ft})(5 \text{ ft})
\]

\[= 109.956 \text{ sq ft}\]
Appendix E: Collaborator Agreement

Math Data Corpus Collaborator Agreement

The Math Data Corpus was collected by Inca Designs with IRB consent, assent, and media release forms, and they permit sharing of data by collaborating partners for research and education purposes only. After signing a collaborator agreement, this dataset is available to participants who wish to use them for the data-driven grand challenge Intl. Multimodal Learning Analytics Workshop to be held in 2014 at the ICMI Conference in Istanbul, Turkey. Once the collaborator agreement is signed and submitted, a password will be provided for downloading the dataset. The data and a detailed description of the dataset is available at: http://mla.ucsd.edu/data. Once a signed copy of the collaborator agreement has been submitted to oviatt@incaadesigns.org, a username and password will be provided to access them.

The purpose of the collaborator agreement is to:

1. protect the original research subjects’ legal rights
2. agree to limit any analyses to those intended for participation in the MMLA workshop, unless further permission to use them for other purposes is obtained in writing
3. agree to report any findings freely to other attendees at the MMLA workshops and the broader scientific community
4. agree to contribute any additional transcription, coding, or enhancements performed on the dataset with others in the scientific community to build a common resource for future community use
5. agree to cite the database and its published description in any workshop paper or other publication as:


With respect to protection of the original research subjects’ legal rights, all collaborators who sign below agree that:

1. they have read, understood, and agree to abide by the original research IRB consent and media consent forms (see attached information); this includes protecting all data files, and the anonymity of the subjects represented in the data files, by completing any work on password-protected computers in a private work environment to prevent unauthorized individuals from observing or copying the files
2. only they or individuals working in their group will be given access to these data files, and all individuals working with these files will have completed human subjects training and certification
3. they will use these data for research and educational purposes only as part of the MMLA grand challenge workshop, and will destroy their copy of the data by Dec. 30th 2014
4. They will seek permission in writing if they wish to use the dataset for any other purpose (e.g., students completing masters or PhD theses on related topics) or beyond the date indicated above.

I understand and agree to abide by the above conditions and restrictions for using this dataset:

Name: __________________________ Date: __________________________

Signature: __________________________

(MMLA Workshop participant and collaborator)