

Lessons from the Design of Formative Feedback Tools for Teachers

Suzanne Rhodes, MA

Department of Educational Psychology

Wisconsin Center for Education Research

University of Wisconsin–Madison, USA

serhodes@wisc.edu

Richard Halverson, PhD

Department of Educational Leadership & Policy Analysis

Wisconsin Center for Education Research

University of Wisconsin–Madison, USA

halverson@education.wisc.edu

Objective: We present findings from the initial phase of our collaborative design process to build handheld, formative assessment tools for teachers. Specifically we report on 1) a survey of the variety of teacher-level data collection processes in the school, and 2) the early tool-prototyping process. **Background:** Prior work demonstrated the important role played by formative feedback systems. The study also revealed that school data systems have at least two levels: a district-sponsored, technologically-complex summative system and a distributed, fragmented teacher-driven formative system with information rarely exchanged across these levels and a lack of formative assessment tools for teachers.

Methods: Via a Design-Based Research process we collaborated with 13 elementary school teachers, leaders, and staff. We examined district-level summative data systems as well as teacher-level data collection practices and tools and iteratively designed handheld formative assessment tools. **Results:** Although we found patterns in how teachers collected and recorded data, we also found that teachers valued the ability to customize their data-collection tools. The early design process highlighted the tension between standardization and customization. **Conclusion:** A design intended to create a common vocabulary about classroom assessment can be perceived as a threat to teacher

autonomy and the fragmentation of the teacher-level of a school data-system might be seen as a feature rather than a bug. **Application:** ...TBD

Education data systems typically pull data from the classroom to meet summative accountability demands rather than to formatively support decision-making and improve teaching and learning. “Summative feedback describes the *results* of processes, while formative feedback is used to *inform* and *adjust* the process as it unfolds” (Halverson, Prichett, & Watson, 2007). According to Black and Wiliam’s (1998) meta-analysis of 43 studies of classroom assessment’s effects on student learning, “there is a firm body of evidence that formative assessment is an essential feature of classroom work and that development of it can raise standards...we know of no other way of raising standards for which such a strong *prima facie* case can be made on the basis of evidence of such large learning gains.” Unfortunately, summative assessment cannot currently inform formative assessment because as Black & Wiliam (1998) claim, “teachers find it too difficult to reconcile the roles of summative and formative assessment and feel forced to provide summative data.” Therefore, districts and states collect data that is often disconnected from the processes that actually occur in classrooms. Summative data, then, does not usually help school teachers and leaders improve teaching practice for the purposes of increasing student learning and achievement.

Our prior NSF-funded work explored how leaders help their schools develop the capacity to act on achievement data also demonstrated the important role played by formative feedback systems (Halverson, Prichett, Grigg, & Thomas, 2005; Halverson, Prichett, & Watson, 2007; Prichett, 2007). The study revealed that school data systems have at least two levels: a district-sponsored, technologically-complex summative system and a distributed (typically fragmented) teacher-driven formative system. We found that information was rarely exchanged formally across these levels largely because of the lack of design attention paid to the teacher level and tools that provide teachers the kinds of formative information necessary for student learning are often left out of data system designs.

This paper describes a collaborative design experiment to build formative feedback data tools networked and integrated with district data systems that teachers can customize the data they record about student learning. We report on two aspects of this initial research that we began in Fall 2007: 1) a survey of the variety of teacher-level data

collection processes in the school, and 2) a report on the early tool-prototyping process. Our paper describes how a design intended to create a common vocabulary about classroom assessment can be perceived as a threat to teacher autonomy, and why the fragmentation of the teacher-level of a school data-system might be seen as a feature rather than a bug.

Methods

Design-based Research approach

We selected a design-based research approach for our study as “design is central in efforts to foster learning, create usable knowledge, and advance theories of learning and teaching in complex settings...and for understanding how, when, and why educational innovations work in practice” (Design-Based Research Collective, 2003). A design-based research approach involves collaborative, iterative analysis, design, and development processes similar to game design and development processes. Using a collaborative design approach to build tools for professional practice encourages a cross exchange of knowledge to refine the resulting design: researchers test theories of action against practitioner experience and tool affordances; practitioners make their theories of action explicit in tool design and learn from reflecting on their practice; and designers understand the constraints of the environment under which professional tasks are undertaken. In additional, this approach can promote intersubjectivity between all stakeholders solidly and rapidly. A design-based model for research *surfaces the real constraints and affordances* that shape what professionals see as possible, *identifies the critical junctures* for which formative feedback tools might be constructed, and *provides an authentic opportunity for testing* the effects of new tools on professional practice.

References needed?

Participants

We collaborated throughout the research process with 13 teachers, leaders, and staff in an intermediate school (grades 3-5) who have taught in the classroom between 3-20 years. We selected the school because of its reputation for effective data use to inform student learning and its established record of improving student test scores. The participants included at least two teachers from each of the three grade levels as well as physical education and art teachers. In addition, our participant sample also included the

school's instructional coach, reading specialist, school principal and the district technology coordinator.

Materials and procedure

We first reviewed the district-level data systems and the two school-wide math and reading curriculum programs (both were designed by third party vendors). We then collected samples of teacher-level student data collection and assessment artifacts. We asked teachers for clarification on symbol/grade meanings and for details on how the documents were used when needed. We also observed classrooms and collected artifacts that were used during the time observed or collected them from individual teachers and curriculum materials out of the context of the classroom. The 40 artifacts surveyed included grade sheets, student-teacher conference tools, weekly quizzes and assessment tools from grade-level subject-matter third-party curricula, observation checklists, rubrics, report cards, and student self-evaluation forms. All teacher-level artifacts were paper-based and included student data was de-identified. Our intent was to develop a detailed typology of assessment practices and tools at the teacher, student, and classroom level in order to identify the critical junctures for which formative feedback tools might be used in practice and to inform the tool interface design in a way that would reveal as well as both support and stretch classroom formative feedback practice.

Parallel to and integrated with the assessment artifacts survey, we engaged with the teachers as researchers and designers during design meetings. At these meetings we discussed and applied assessment practices to the designs, reviewed assessment software tools and devices, examined sketches, discussed feature and functionality needs and priorities. We iteratively developed use cases, narrative walk-throughs, action flowcharts, and an information architecture diagram, which served to merge the perceived requirements data gathered from the interactive design meetings and the document analysis. We also developed and collaboratively critiqued a series of wireframes of the interface design and revised all design products. All design and development processes and artifacts were organized and managed via a private wiki website accessible to all team members and included commenting features where teachers could review, verify, and confirm information and designs.

Analysis

We performed a document analysis, which is the systematic examination of documents in order to describe an activity and identify instructional and assessment needs and challenges, on the teacher-level assessment artifacts surveyed. Questions we asked of the artifacts during the analysis included:

- Who sponsors this artifact? (single teacher, curriculum team, curriculum program)
- Who uses this artifact? (student, teacher, both?)
- Type of artifact (rubric, grade sheet, conference form, self-assessment, etc.)
- What is the type of assessment (formative, summative, both)?
- How frequently is this artifact used?
- For which subject and grade level is the artifact used?
- In what context is this assessment used (whole class instruction, student-teacher conference, small group instruction/assessment)
- What kinds of data are collected via this tool (grades, soft skills, numbers, text comments)?
- Does the teacher customize or augment the tool and the data? In what ways?
- What are the features of the artifact and what processes and actions do the features afford and constrain?

We created a chart to record the information garnered from these questions and then identified patterns and opportunities for the use of the tool and the standardization and customization of the interaction and interface design and data teachers collect.

Note/Thoughts: Apply Rich's formative feedback system as framework for analyzing the documents? Need to do a quantitative approach as well? Maybe in the future after the conference and for MAP? Also link to DDIS survey data would be an excellent idea—larger sample to corroborate results.

Results

The assessment survey and design activities have provided insight onto the data use activities of teachers and leaders in schools. The insights about professional practice, however, have helped us understand several key aspects of the local data systems in the school.

The school was already awash in summative student performance data. We counted eight different summative, and summative/formative combination data sources that

leaders, teachers and staff already had to juggle – some integrated with the district data systems, some locally generated and stored, and yet others idiosyncratic to particular teachers, grade-levels, and subject areas. The split between school formative and summative data systems noticed by our prior research was reflected in daily teacher practice. Teachers rarely used and did not personally access the largely summative district-provided data tools to record information about individual or classroom student performance.

Teachers recorded formative information on student learning in a variety of ways. The assessment tools used were all paper-based and either created by the teacher personally or adopted from the math and literacy curriculum materials. Teachers added information in the margins of grade books, recorded information on slips of paper, and often kept student information “in their head.” In one case, a teacher recorded soft skills information on a daily seating chart. These formative data records were certainly not systematic and were often inefficiently redundant; the idiosyncrasies of individual recording reflected a fragmented formative data collection process.

Although teachers thought about data collection in terms of the curriculum and assessments already in place in their school, they wanted data collection tools that could be customized to idiosyncratic needs. Teachers would often design their own formative data collection tools to supplement the tools they used as part of the common school-wide curriculum and then would also customize the data. We found that while there were surface differences in teachers’ paper-and-pencil data collection processes (e.g. annotating, underlining, circling, symbol use, different meaning assigned similar symbols, etc.), there were also deep functional similarities for the purposes of increasing efficiency and effectiveness and reflecting on and informing instruction to personalize learning for students and themselves as practitioners (e.g. chunking, contextualizing, layering, and attention directing techniques). One teacher developed a fairly elaborate symbol system and data augmentation system in that she layered soft skills performance data in the form of checkmarks, pluses, minuses on top of the codes for academic performance in grade sheet cells resulting in a richly dense data artifact. In fact, most text comments were quite brief for all teachers, including no more than 4-5 words (although the text entry areas in the artifacts may have contributed to the constraint of

this action). This finding corroborates Kozma's (2003) research where he found that experts cluster information in a meaningful way, use a variety of representations, and also transform representations.

The majority of the artifacts took the form of spreadsheet-like grids, rubrics, and skills checklists. In many case, the rubrics and checklist format document formats were used by either by the teacher alone during small group conference or assessment or shared with a student during individual meetings during class time. Teachers entered data in the grid document formats in between classes, during whole class work time, and after school. Indeed, we observed that teachers primarily assessed students during class when they met with for one-to-one student-teacher conferences, small group instruction and testing, and while students worked as whole class and in small groups.

We found that in all of these data collected, teachers had little time to a) analyze the information or b) to systematically customize data collection to the specific needs of their curriculum. While one formative assessment document asked the teacher to record "ideas for instructional change" teachers very rarely recorded such ideas on their own in the documents we surveyed.

During the design meetings and one-on-one researcher-teacher conversations in between classroom observations we learned that teachers wanted tools they could use at home and at school. They primarily wanted the tools to help with formative assessments, goal setting, ability grouping. However, they wanted the tools to also assist and integrate with summative assessment and accountability processes. The primary requirement was that the tool be easy to use. The school did not want tools, however interesting the possible functionality, that would require them to add yet another layer of data collection.

KidGrid Early Prototyping Process

Based on the initial document analysis, observations, and design meeting discussions we to create iterative designs for *KidGrid*, a networked formative feedback application that would be student-centric.

First, we needed to select a technology platform. We needed to design formative data tools that were unobtrusive in a classroom environment, non-redundant, easy to use, and *mobile* as we discovered that teachers rarely stay in one place more than a few seconds. We looked to past successes to guide our designs. Rochelle, Patton, & Tartar

(2007) described three success stories of technology-enhanced learning that drew upon properties of networked handhelds that do not particularly characterize personal computers. In particular, Probeware's major benefits for students is that this technology eases collecting and recording accurate data, provides the ability to collect time series data, allows for instant graphing and analysis of data, and provides the opportunity for students to exchange or pool data sets. Additional research has shown that two of the advantages to networked, handheld computers in the classroom include portability and easy integration into existing classroom infrastructure (Roschelle & Pea, 2002; Tatar et al., 2003; Vahey & Crawford, 2002; Tinker & Krajcik, 2001; Soloway et al., 2001; Staudt and His, 1999). **Check the references!** Sharples, Taylor, & Vavoula (2005) suggest that "For the era of mobile technology, we may come to conceive of education as conversation in context, enabled by continual interaction through and with personal and mobile technology." In addition, Vahey, Tatar & Roschelle (2007) focus on one particular social set of handheld affordances: "the ability for students to engage in, and move seamlessly between, *private* interactions with their computational environment, and *public* interactions for face-to-face collaboration around the computational environment;" leveraging these affordances provides a platform for advancing research as well. Mobile technology certainly seemed to be the way to go.

Roschelle, Penuel, Yarnall, & Tatar (2004) found that science teachers needed support to conduct ongoing formative assessments to track students' evolving content and conceptual knowledge in the science classroom. Their work has revealed that teachers need ways to increase the quality of assessment information and need tools that "informate" rather than "automate" existing assessments. Their Wireless Handhelds Improving Reflection on Learning (WHIRL) project is exploring the use of handhelds to provide support for better assessment in science classes. However, with a few exceptions like the WHIRL study, most of the literature about handheld tools in education focuses on mobile learning and students' use of handheld technologies for learning situated in subject-matter. But this research maps directly on the use of handhelds by teachers for data sense-making and formative feedback purposes as well. These studies also influenced our data tools for teachers research as well in that they led us to focus on handheld, mobile tools for teacher and to include teachers in the data system design

process in order to take advantage of the information revolution sparked by the school accountability movement.

KidGrid: the early design process

With the teachers' requirements and research data in mind, we brainstormed concepts on the tools we wanted to create for teachers that facilitated student-centric formative data collection and reflection and took advantage of the media rich features of current mobile devices. We came up with 1) a student performance tracker where the teacher evaluates an individual student's action by selecting a general evaluation area, selecting an evaluation code, recording voice notes, typing text notes, and/or selecting pre-determined notes, 2) a student intervention tracker where the teacher enters information about an intervention performed on an individual student, 3) a group performance tracker, 4) an interaction recorder for observers, 5) a teacher self-assessment tool that could facilitate teacher evaluation of their own original lessons, or implementation of otherwise new curricular material, 6) a data dashboard that represented student data with intuitive graphical formats and 7) a group manager and data report application for a desktop computer synced with the handheld tools. We then wrote a design document for each tool that included primary and secondary audience descriptions, the purpose of the tool, features and functionality, use cases, narrative walk-through descriptions of the tool in action in context, a list of data content that would be collected, and wireframes of the interface. Figure X shows a wireframe of the initial design for a student and small group formative feedback management tool. The main menu is a grid of student pictures in a class the order of which can be customized. A teacher could choose to view her students based on her seating chart or student groups, for example. Tapping on a student's picture brings up a separate data collection window.

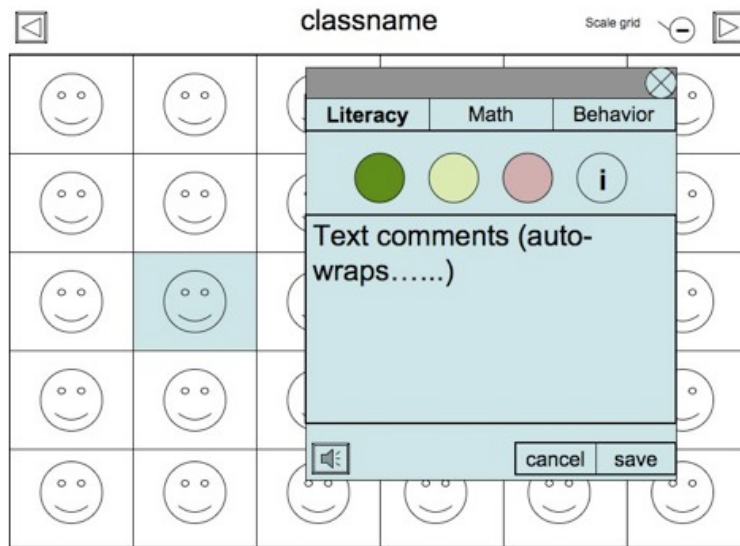


Figure X: Interface Wireframe

We posted all documentation to the project website for the teachers to review and also fine-tuned the interface (Figure X) and features and functionality during several design meetings. For example, the grid menu could now also be viewed in a list format and we included a feature where teachers could “check-off” students. However, Figure X: Data collection and review screen, depicts a rather standardized interface where performance and intervention data is “hard coded” and based on a common and static list of assessment items and codes that teachers would agreed to use that may not be supported by a database. This is because we intended to create a common vocabulary about classroom assessment with tools that could both informate and automate formative feedback processes. Our goal was to *reduce* the data fragmentation at the teacher level and bridge the gap between formative and summative systems through standardization.

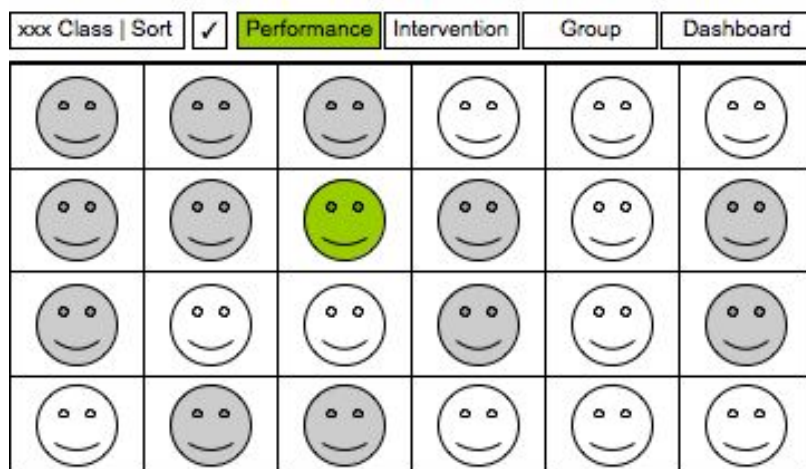


Figure X: Main menu screen

Figure X: Data collection and review screen

KidGrid v 1.0

At this point in our design process, the much awaited software development kit for the highly interactive, media enabled Apple iPhone had just been released and we selected this device as our platform. We decided to develop a native iPhone application rather than a mobile version of a website as teachers may not have regular wifi access in the classroom. We immediately purchased iPod Touches and signed up for the University iPhone development program in order to explore the SDK, database options, software applications, and interface. Armed with greater knowledge of the limitations and affordances of the iPhone/iPod Touch as a handheld device we revised our designs

beginning with a final version of the information architecture (Figure X), which drove the finalization of the interface elements.

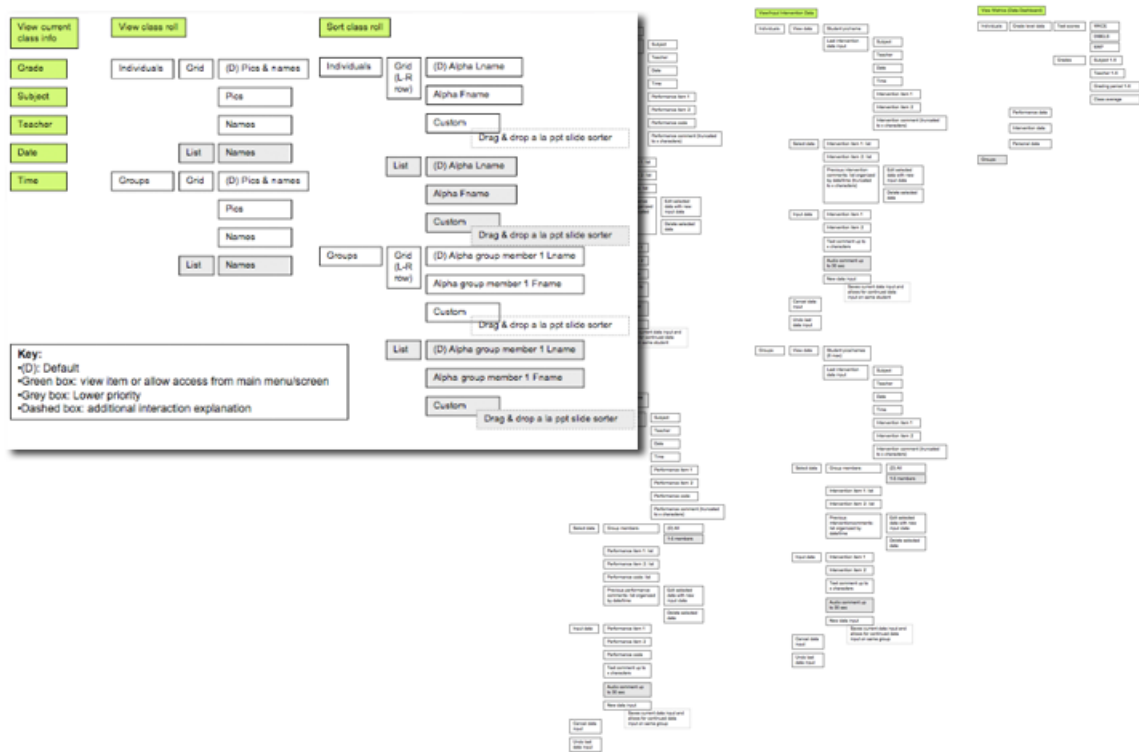


Figure X: Information Architecture

We decided to take full advantage of the extant interface elements extant via the SDK. Our intent was to provide teachers with a tool that contained functionality similar to the iPhone operating system and applications in order to flatten their learning curve. Figure X shows the final version of the data entry and collection screens for *KidGrid 1.0*. The teacher logs in to the application and views a grid of all students in the class she is teaching for that time period as the tool is synched to the district curriculum database. The teacher can also view other classes and create groups of students for each class or subject. To enter data about a student, the teacher taps on a colored square icon above a student's name on the main menu grid (we were unable to use student pictures for the icons due to student safety and privacy issues). This action brings up the data collection screen where the teacher can enter text annotations about a student and access the formative feedback “wheels” that contain curriculum benchmarks, units, and performance codes as well as instructional feedback content. These “wheel” data are

curriculum-driven and standardized as per the district requirements. However, as can be expected from the results of our assessment document survey, teachers really required individualized data systems for the tool to work for them in a formative classroom context. Therefore, the wheel data can also be customized by each teacher individually on another screen and saved to a local database. (We hope that this data will also reveal deeper and more authentic data on teacher’s formative feedback practices). After the teacher selects performance or feedback data or enters text, the data are time-stamped and saved in list format associated with each student for each class/subject. The saved data can be accessed and modified or viewed any time, anywhere.



Figure X: *KidGrid*: student performance and instruction data collection

As an example use case of *KidGrid* in practice, a teacher may be teaching a reading unit focused on students’ “Making connections.” The teacher may rotate to each student in the class during reading time for individual consultations. She may talk to the student about the content of the book he is reading and may ask the student to make connections from themes of the story to the student’s real life. The teacher could enter in brief text notes about questions she asked as well as student responses, select the “making connections” benchmark item from the top wheel and then a performance code of “2” if the student made a moderate connection from the theme to real life, and also select “Suggestion” from the “feedback” wheel to track that she, as the teacher, provided a

suggestion as to how the student might make a stronger connection. In essence, the teacher has collected data about a teacher-student formative feedback interaction, a concrete teaching-learning experience, and that data is saved to a database along with automated contextual, situated cues. The teacher would then move on to the next student. After class, the teacher may review the data she entered about the student and the instruction or feedback she provided to the student and compare that data to other data entered previously and reflect on the student's learning, her instructional practices, and her instruction over time. She may also use the data for planning her next lesson and future individual consultation sessions with the student. The teacher may bring the tool to a meeting with the student's parent and, because the tools are networked and contain standardized content features, can share the data before, during or after a meeting about curricula or assessment to collaboratively plan and evaluate practice and student achievement with other teachers in the local school community.

Conclusion

Our data collection survey revealed teacher-level data collection practices such as grade sheets, literacy and math rubrics, observation checklists, student-teacher meeting organizers and idiosyncratic assessment symbol systems that varied across teachers. Our design-based research process highlighted a tension between standardization and customization. Although we found patterns in how teachers collected and recorded data, we also found that teachers valued the ability to customize their data-collection tools as well as the data within them. Teachers pushed back on our efforts to build a common tool data system and encouraged us to enable as much customizability as possible. Therefore, the fragmentation of the teacher-level of a school data-system can be seen as a feature rather than a bug. **Need more here. Lessons learned from the prototyping process? Necessary to mention limitations: small sample, expert teachers?**

Application/Implications

We hope that this line of research informs efforts to understand how data tools actually operate in schools and contribute to our understanding of how to bridge the gap between the formative and summative aspects of school data systems and inform designs for handheld formative feedback tools that alleviate the tension between standardization

and customization in ways that stretch practices and initiate dialogs about data and student learning in interesting ways.

Future design possibilities: KidGrid v 2.0

Ainsworth (1999, 2005) identified three key functions of multiple external representations (MERs): to complement, constrain, and construct. When MERs compliment each other, each representation contains some different information. A complementary function can be appropriate for an experienced learner audience as redundant information can be reduced while presenting the learner with complex, layered representations of data. When the function of MERs is to constrain interpretation “a known representation supports the interpretation of an unfamiliar abstract representation” and the goal is to “exploit learners' understanding of the relation between the representations” (Ainsworth, 1999). Therefore, “a learning environment should make very explicit the relation between representations... achieved by automatic translation or dynamic linking as a learner manipulates one representation, another one changes” (Ainsworth, 1999). When the function of the MERs is to construct understanding, multiple representations can be holistically linked together cognitively as a “gestalt reification” where the learner can then reflect upon the abstraction and visualize complexities (Ainsworth, 1999). In addition, when MERs function to construct understanding, knowledge is not necessarily reorganized but extended and thus generalized (Ainsworth, 1999).

In the current version of *KidGrid*, teachers construct concrete and symbolic representations of their interpretations of student knowledge and performance and their instructional feedback when they modify the wheel data according to their own preferences, select data from the “wheels,” and add text annotations associated with a student. However, currently *KidGrid* compiles that teacher-entered and selected data into basic viewable text-based list-type data representations. Textual lists and annotations may be fine for expressing “ambiguity in a way that graphics cannot accommodate” (Stenning & Oberlander cited in Ainsworth (1995)). However, lists are certainly not the only or necessarily the best way to represent data to aid experiential learning and support efficient reflection-in-action as well as active learning via reflection-on-action for experienced practitioners. By considering Ainsworth’s three MER functions, the limited

screen real estate of the Apple iPod Touch and the audience needs, the following illustrates example MERs mock-ups as two potential additional data representations for *KidGrid*.

Figure 2 illustrates the grid menu buttons. Colors could represent a student's average performance score for knowledge of a content benchmark or unit lesson (i.e. a scale of 3-1 represents “the student's got it!”, “she's getting close to understanding”, and “not adequate knowledge, needs support” associated with blue, yellow, and red). The student's average score number for the current subject unit and lesson could be displayed on the button and underneath it, the number of performance code entries and the number of times the instructor provided instructional feedback. This grid student icon data would be dynamically linked to the “wheel” data and automatically updated via teacher data entry actions. This representation of student performance and instructional feedback is one the teacher could process at a glance during reflection-in-action while she is meeting with a student. There is some redundant data (i.e. color and the number for an average performance score) but the combination of colors, numbers, and averages of data provide a simple yet holistic picture of the student's and the instructor's past actions that would inform the action of the teachers while interacting with the student during, for example, a reading conference.



Figure X: Student grid menu buttons

A more complex representation of descriptive statistical student data is illustrated in Figure 3 that could be easily viewed when the teacher taps a button to “flip” the student data entry screen over. This data “dashboard” represents a variety of metrics, which illustrate current and past student performance and instructional data. The top left area is a basic “tag cloud” visualization of the words the teacher has entered in the text annotation area of the data entry screen; the text size and location might represent the frequency the teacher included in the text entry area on the student data entry screen and how recently she entered that word. Font style and color could be applied to provide other layers of representation. The area located in the upper right tracks the total types

and number of instructional feedback the teacher has provided the student. The areas under the text cloud show the mean and mode benchmark or unit and lesson performance scores for the student and class for comparison purposes. The graph at the bottom depicts colored points as the performance scores the teacher entered for the student for each benchmark unit and lesson over time; this graph would include a line of regression. All graphical statistical visualizations would be dynamically linked to the student data entry screen and updated automatically. The entire data visualization screen might scroll horizontally to provide additional dashboard data. Certain data components could be interactive, where the teacher could tap on an area in order to “drill down” to view more detailed information, or tap certain hotspots to annotate symbols (e.g. the teacher could tap the points on the bottom graph to change the color of the point in order to make that point more visually explicit as a reminder to herself).

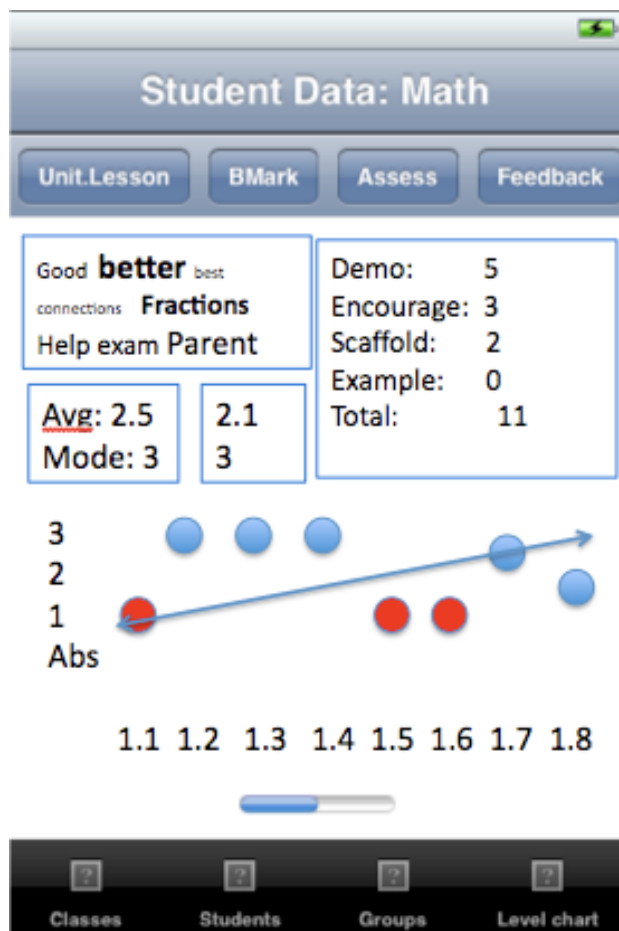


Figure X: Data representation screen

This statistical data dashboard of student performance and instructional feedback is merely a potential example of the MER features and functionality of *KidGrid*. This feature requires further collaborative design with target audience. However, this screen demonstrates MERs that are multi-functional in that the data representations complement, constrain, and also afford construction of knowledge via interactivity. As this MER *KidGrid* feature represents student and teacher actions and interactions, provides contextual information about the data to foster reflection, and affords efficient functions that teachers already perform via paper-and-pencil according to the needs analysis, we would expect that a feature similar to teachers' current practices will support their experiential learning in multiple activity settings. Rather than presenting iconic representations where the audience must make sense of and thus learn the data representations, the strategy for *KidGrid* MERs is to provide multiple concrete representations of implicitly symbolic data with constraining functions individually that, when layered, together complement each other and serve to focus the teacher's attention on the data as representational of multiple actions to foster the teacher's knowledge construction and reflection about each student and the class as a whole.

References