# Literacy Design Collaborative - Module Jurying:
# Innovating for High Quality Design

Paper presented at the

Annual Meeting of the American Educational Research Association

Philadelphia, PA

April 2014

Ruth Chung Wei, Stanford University, rchung@stanford.edu

SCALE

Stanford Center for Assessment, Learning and Equity
Stanford University

The Stanford Center for Assessment, Learning, & Equity (SCALE) is an educational research and development laboratory at Stanford University's Graduate School of Education.

http://scale.stanford.edu
Littlefield Management Center
365 Lasuen Street
Stanford, CA 94305

# Literacy Design Collaborative - Module Jurying: Innovating for High Quality Design

## Context

The Literacy Design Collaborative (LDC) was initially conceptualized to support teachers across the content areas in designing their own assignments that would build students' literacy and writing skills aligned to the Common Core State Standards (CCSS). The simplicity and elegance of the templates, as well as their "open-source" availability, has supported the LDC tools "going viral" across the country, with some users in states that have committed to integrating the tools into state supported pilots, some in districts that have committed to piloting and supporting the tools, some in regions supported by strong networks (e.g., SREB), and others where individual teachers have sought out the tools based on word of mouth. The beauty of the LDC tools is that they simultaneously provide a strong architecture for alignment with Common Core expectations, while providing freedom and flexibility so that teachers can adapt the tools to their local curricula and content standards.[1]

While the LDC teaching task and module templates were designed to "hard-wire" the Common Core Literacy Standards into the module, it became apparent after a year of piloting the LDC templates across the country that the templates did not guarantee quality or rigor. While deceptively simple to use, the templates require that module designers make strategic decisions about what content, texts, questions, and instructional strategies to insert into the "fill in the blank" templates. LDC had a limited cadre of staff members with deep understandings of what makes for a high quality module who could traverse the country, teaching practitioners how to improve their use of the templates. In addition, there were few models that could be presented to practitioners as exemplars to emulate. LDC leaders identified a clear need to develop a set of quality criteria and module exemplars that could be used to set the standard for what "high quality" look like.

Beginning in the fall of 2011, the LDC and the Stanford Center for Assessment, Learning,

[1] For more detailed information on the LDC framework and templates, please see http://www.literacydesigncollaborative.org.

and Equity (SCALE) began a collaboration to develop criteria to evaluate the quality of the LDC modules that were being designed by teachers and other users, including their alignment to the CCSS. Another purpose that shaped the design of this tool was the goal of providing fine-grained, constructive feedback that task designers could use to improve their designs. This paper describes the development process and rationale for the design of the "Jurying Rubric" that emerged from that process.

In addition to the rubric, SCALE was commissioned to develop a training protocol and anchor modules to train and calibrate LDC jurors. This juror training protocol has undergone several rounds of pilot testing and refinement. The jurying process that was settled upon is a socially moderated process in which jurors rely on each other's expertise. In this approach, each module is scored independently by at least two jurors who then compare notes and arrive at consensus on the scores. This paper discusses the rationale for the design decisions that were made about the jurying rubric and jurying process, including the decision to rely on ***distributed expertise.*** An electronic scoring and annotation platform designed for the project makes it possible for remotely located jurors to work together on rating modules and for disseminating and distributing expertise to a national community. Evaluation survey results and calibration results are used to reflect on the merits and challenges of particular training designs.

Last, this paper explores the question of how tools like the jurying rubric and protocol can support LDC module quality while also supporting local teacher innovation.

## Theoretical framework

The theory of action underlying the work of the Literacy Design Collaborative is a hypothesis that by providing a set of templates for writing tasks and instructional modules, the quality of teachers' assignments and instruction will improve. The ultimate goal of LDC is to improve student learning, achievement, and readiness for college and careers.

This premise is not new. Prior research has found the quality of teacher assignments to be predictive of student achievement. For example, research on the Instructional Quality Assessment (IQA) (Matsumura, Garnier, Pascal, & Valdés, 2002) has found that the quality of assignments given to students by teachers is a strong predictor of student achievement. In a study of teachers in the Los Angeles Unified School District, Matsumura and colleagues found that a measure of high school teachers' assignments, which were scored on their cognitive challenge, clarity of learning goals, clarity of grading criteria, and overall quality, was significantly related to the quality of student work produced by their students and reading/language scores on a standardized achievement measure. Research on the quality of intellectual work rubric developed by Fred Neumann, Anthony Bryk and colleagues (Neumann, Lopez, and Bryk, 1998), and further developed by the Center for Authentic Intellectual Work (King, Schroeder, & Chawszczweski, 2001) has also found that when teachers are provided with clear criteria for high quality assignments and engage in professional development focused on these criteria, the quality of their assignments improve and their students' achievement improves accordingly (Iowa Dept of Education, 2012).

A second theory of action underlying the jurying work is that expertise can be distributed and disseminated. The goal of the LDC module jurying work is to build a non-exclusive national cadre of professionals who are equipped to not only jury new modules that are submitted for review, but to also deepen their expertise about high quality modules so that they are better equipped to provide professional development and coaching to users of the LDC templates and tools. Another goal is to directly disseminate clear criteria for "exemplary" writing tasks to teacher practitioners, the designers of LDC modules. Here, an understanding of how expertise and knowledge are disseminated, how innovative tools "go viral", and a new understanding of how expertise and tools get shaped comes into play.

In his book chapter, "Innovation and Diffusion as a Theory of Change," Tom Bentley (2010) contrasts the current bureaucratic model of school reform (which seems to impede diffusion of good ideas) with the rapid diffusion of innovations in ICTs (Information and Communication Technologies), citing the Linux operating system, the Internet, Wikipedia, and Facebook as innovations that began in educational institutions, but experienced rapid and wide-spread diffusion. Bentley calls out three common features of innovations that experienced rapid and large-scale diffusion:

1) open source -- allowing access to users and participants with few restrictions;
2) highly networked, enabling rapid lateral transfer across institutional, sectoral, or geographical boundaries; and
3) user driven, empowering participants instead of relegating them to a role as passive consumer or inexpert student.

The first two characteristics are ones that have clear resemblance to LDC's mode of operation up to now. Because the LDC templates and tools have heretofore been made available to the public freely, without charge and with limited restrictions, as well as through grant support from the Bill and Melinda Gates Foundation for states that have committed to large-scale piloting, LDC has seen a rapid expansion across the country since it was first introduced in 2010. The LDC community is also highly networked, operating through school networks (e.g., New Visions Public Schools in New York City) and regional hubs (e.g., Southern Regional Education Board), and supported by national professional development centers such as the National Writing Project and Paideia Schools, among others. But the Gates Foundation's investment in state-sanctioned initiatives (working through more traditional top-down implementation strategies) has also worked to support dissemination.

The third characteristic - user driven - is one that is of particular interest for the LDC jurying work. It suggests that rather than focusing on traditional "fidelity of implementation" criteria to evaluate the success of an innovation, that instead, a key indicator of an innovation's successful diffusion is the engagement and empowerment of users to drive the design and content of the innovation itself.

As a professional development approach, LDC has some aspects that could be said to focus on "fidelity of implementation". Even the jurying work calls for users to use the templates as intended (without modification to the template language) and for the modules to have complete components. However, the jurying system depends on participants contributing to an understanding of exemplary modules, both through their engagement in the rich conversations around modules, but also by contributing the modules themselves, which then come to serve as the exemplars of different score levels. Until recently, there was a paucity of modules that could be held up as exemplars, partially because of poor access to a wide array of modules, but also because the definition of exemplary had not been formulated and disseminated. It was a "cart before the horse" phenomenon. Until now, the "exemplary" level has been defined rather narrowly because the number and variety of anchor modules that we had available to us to define and illustrate the "exemplary" level has been limited. As we begin to accumulate more and more modules that we can recognize as "exemplary", we anticipate that the definition of exemplary will also be broadened and at the same time, become more specified and transparent.

### Key design features of the LDC module jurying rubric and jurying process

The LDC module jurying rubric was developed through a socially moderated process that harnessed the expertise of core LDC staff (founders Marilyn Crawford and Eleanor Dougherty, LDC staff member Stacy Kaliatsos), early adopters such as Susan Weston, Janet Price, Barbara Smith, and others) and included individuals with a broader base of expertise in literacy and performance assessment (David Pearson, Ray Pecheone, Stuart Kahl). Convened in early 2012, this group together described the key elements of an acceptable LDC module, which was used to generate a scoring guide (a rubric) with two levels -- 1) Work in Progress; 2) Good to Go. In the spring and summer of 2012, this two-level version of the rubric was piloted in that year and feedback from the national community was collected through national conferences and other meetings. In January of 2013, a third level was added to the rubric -- "Exemplary" -- to describe the features of high quality LDC teaching tasks and instructional modules. Again, the three-level rubric was piloted with a national audience and additional feedback was collected on the rubric during the spring of 2013. The rubric had two overall scoring domains - 1) the Teaching Task; and 2) the Instructional Module. The three level rubric had eight scoring dimensions as below:

A. The Teaching Task
   1. Task Clarity and Coherence
   2. Content
   3. Texts
   4. Student Product

B. The Instructional Module
   1. What Skills?
   2. What Instruction?
   3. What Results?
   4. Teacher Work

Note: The full LDC Module Jurying Rubric can be accessed at the LDC website: http://www.ldc.org/how-ldc-works/tools-to-ensure-ccss-alignment

***A discipline-specific approach to rigor***

Several key features of the Exemplary level of the jurying rubric include: a) an attention to discipline-specific big ideas and enduring understandings; b) complex, higher-order thinking skills central to the discipline; c) a task pattern that can be generalized to other tasks within the discipline; d) authentic and engaging texts and writing products in the discipline; and e) instructional strategies that are specific to the discipline. These key features of the Exemplary level of the jurying rubric reflect a stance about rigor that the

Stanford Center for Assessment, Learning, & Equity (SCALE) brought to the table. The stance is that rigor cannot be present without attention to discipline-specific content, thinking skills, and instructional strategies.

The LDC templates and scoring rubrics are content-neutral, in the sense that they are designed so that teachers across the disciplines can use the same templates and apply them in different ways to create tasks in their disciplines. However, SCALE believes that rigorous learning and demonstration of that learning cannot be assessed without attention to the particular definitions of rigor that are embedded in each discipline. For example, the literacy skills that are applied to an analysis of a literary text (e.g., a novel) are quite different from the literacy skills that are applied to a scientific analysis of a science journal article or an informational website. Likewise, historical analysis of primary sources involves more than summarizing the author's message - it means looking carefully for information about the author, date and place of publication, audience and purpose, and the historical, political, cultural, or personal context in which the source was created, and using those pieces of information to make inferences, interpret, or draw conclusions about the author's message. By incorporating discipline-specific criteria into the Exemplary level of the jurying rubric, LDC also took a stance that rigorous modules must attend to the specific literacy skills and strategies of each discipline.

This attention to disciplinary rigor has repercussions on the jurying process, the juror training design, and WHO is qualified to jury modules across disciplines. In our initial piloting of the juror training protocol in May 2013, we invited the original members of the team that created the module jurying rubric, as well as members of the broader LDC community who were leaders within their contexts. The training protocol was not discipline-specific, and used training materials (anchor modules) that represented three of the most frequent content fields in which the LDC templates are used -- ELA, history, and science. Following the training, we asked participants to jury as many modules as they could within a one day period. While we tried to match the content expertise of participants with the content of the modules, this was not always possible. We did not have any science content experts in the group; nor did we have many career and technical education specialists. So science and career-pathway modules (e.g., law, health professions, agricultural science) were assigned somewhat randomly to jurors without subject area expertise. What we found is that a lack of content knowledge as well as content pedagogical knowledge made it more likely that jurors would inflate the scores of the modules. Later, when we had science specialists review the science modules that had been scored as Exemplary, their scores on the same modules were not as high.

What we realized is that to be an accurate juror, one must have sufficient content expertise and pedagogical content knowledge to jury modules because without that knowledge, you would not know what is a big idea or enduring understanding central to the discipline, or whether a particular set of literacy scaffolds or instructional activities is appropriate for the discipline. Having a content-area match between the content addressed in the module and the juror scoring it improves not only the accuracy and consistency of scoring, but also raises the bar for what can be considered an "Exemplary" module. While the rubric level descriptors for Exemplary modules always had elements of discipline-specific criteria embedded in them, how one judges whether those criteria are met varies based on how much content knowledge and pedagogical content knowledge one has.

We subsequently redesigned the juror training protocol to include anchor modules and calibration modules that are discipline-specific, and we provided discipline-specific break-out groups so that we could assign participants to their areas of content expertise. We believe that the content-area match is extremely important for participants to become accurate jurors, and to provide targeted feedback and support to teachers designing discipline-specific modules.

In a more recent juror training (January 2014), during which we piloted the discipline-specific protocols, we had an opportunity to test out our theory about the importance of this content-area match. In three out of four groups, participants were placed in the content field for which they had expressed having content area expertise. In the last group, most of the participants were English language arts or literacy specialists placed in a history-social studies group. What we found is that this last group had the lowest level of calibration on the final calibration module. While there are other factors that might explain why this group was the least calibrated (including differences in facilitation, a difficult calibration module), we believe that part of the reason for this lower level of calibration was the content-area mismatch.

When we reviewed the calibration levels overall across groups, we were somewhat disappointed with the calibration of the trained participants. Part of the problem was with the modules that were selected to be the calibration modules, but we believe that the lower than anticipated levels of calibration also reveals an important limitation of a one-day training. What we realized is that to truly build expertise in jurying LDC modules, it takes time and practice with many modules.

### *A socially moderated jurying system*

A second design decision that shaped the LDC jurying system is that, true to the spirit of the Literacy Design ***Collaborative***, the juror training design and jurying process was deliberately designed to include social moderation processes. By "social moderation", we mean that participants experience the training as a collective and contribute to each other's understandings of the meaning of the rubric and the anchor modules. In addition, when participants make decisions about what scores to assign, they make those decisions collectively. Throughout the training process, participants share their observations and perspectives on the modules they have read, and engage in evidence-based discussions in which they are asked to negotiate and come to agreement on what scores to assign each module. Finally, during the calibration phase, participants first score individually, come together with a partner to discuss their scores, and create a "consensus score" for the module. In live jurying, this process of pair-wise consensus scoring is an inherent part of the jurying process. Participants first score the module on their own and generate their own set of scores. Then they connect with a partner who has also been assigned the same module, through Skype, telephone or another social media tool, to discuss their scores. They create a third set of scores - the consensus scores - which reflect their joint decisions about what scores should be given to the module.

This means that every module that is juried has had the benefit of three sets of scores - two independent scores, and a consensus score. We believe this pair-wise consensus process supports score reliability by moderating the impact of lenient and stringent raters. We also believe that there are professional benefits of scoring in this way. It requires that professionals share their thinking with each other and refine their understandings together, like two stones that are rubbed together to polish and sharpen the other. To improve juror reliability even further, the plan is to conduct periodic "read-behinds" for each rater to ascertain their accuracy in scoring and to check for "rater drift", as well as to have everyone in the juror pool score the same module. This allows for a periodic calibration check, and through "crowd sourcing", builds a pool of modules that have been scored by the national community and can be used to refresh the anchor modules.

**Technology tools for diffusion of expertise and local innovation**

Given the geographically dispersed nature of the LDC community, the process of jurying through consensus scoring has been facilitated through an online jurying platform. This online jurying platform allows LDC to assign modules for rating (or modules can be selected by jurors to rate), to jurors who may be located on opposite coasts of the country. The platform stores each of the independent scores, and allows jurors to create a consensus score without overriding their independent scores. The platform can produce data on each module scored (both individual and aggregate scores), and captures multiple levels of annotation on the module itself and on the scoring rubric, as well as commentary and rationale for each score given. This output can be shared with task authors as feedback, to support revisions on the module to move it to the next level. (Jurors are trained during the juror training on how to provide appropriate and constructive feedback, and to write their annotations and comments in ways that may be helpful for the task authors.)

Because of limited resources and time, it is difficult to convene national jurying training events on a frequent basis. For this reason, LDC has undertaken, with SCALE developing the content, to create an online juror training system that will potentially simulate the experience that jurors have in face-to-face training. The biggest challenge in designing this system is devising a way to simulate the rich conversations that participants experience during "table talk" and negotiations during the pair-wise consensus scoring. These conversations are key to making sense of the jurying rubric and the anchor modules. Other online training systems that have been developed to train scorers have been observed to be lengthy, lacking in the engagement factor, and having poor calibration results. LDC and SCALE are working on developing a learning system that is more interactive, that simulates the "table talk", and capitalizes on the "game-ification" of learning that is becoming a more common strategy in online learning. A hybrid approach, in which most of the learning modules can be completed asynchronously is combined with a more interactive approach, in which participants complete some modules in partnership with other local participants, is being considered.

## Conclusion - A tension between expert and user definitions of "Exemplary"?

Up to now, LDC has successfully created a national community around a promising innovation that has already transformed how many teachers around the country design writing assignments, engage their students in developing literacy and writing skills, and organize themselves across local, regional, and national networks to implement the Common Core and improve the quality and rigor of their assignments.

Upholding a high standard in the definition of "Exemplary" work might also move teachers' design work and instruction toward greater levels of disciplinary rigor. However, we wonder whether there is a difference between the way our "experts" (SCALE content area specialists with deep knowledge of the disciplines) define rigor and the way that practitioners (with expertise and deep knowledge of students and teaching contexts) might define it or apply it. We find that in general, after being trained on the anchor modules through a one-day training protocol, practitioners still view the modules through a generic "gestalt" (an overall impression of rigor) and with less precision (for example, a module that includes a high level of detail in describing instructional activities might be scored Exemplary, even if the instruction on the whole is not discipline-specific or tightly aligned with the goals of the teaching task). This results in practitioner scores being generally higher than that of our "expert" jurors. We do not know yet if it this is a problem of needing more experience with jurying, whether this is a problem with the design of the rubrics and training protocols, or whether this is a fundamental tension between "expert" and practitioner views of rigor. We plan to take advantage of additional training opportunities and data gathering in the coming months to capture the differences between expert jurors and practitioner jurors' ratings (and their rationale for ratings) to determine whether there is a fundamental difference between "expert" and practitioner jurors, or whether this is mainly an experiential effect.

Tom Bentley (2010) suggests that what will truly lead to widespread and large-scale diffusion of educational innovation is a continued commitment to being user driven. This would mean that LDC as an organization should work toward a mode of working in which the definition and models of "exemplary" work are crowd-sourced. Expert-driven definitions and models may be limited, static, and lack generalizability, though they do serve to uphold a high standard. On the other hand, user-driven definitions might contribute to a more practice-driven definition of "exemplary" and broaden the ways in which a module could be considered exemplary. Though it is certainly laborious to continuously refine a tool and while there is an appeal to holding something still for practical reasons, involving users in the definition of "exemplary" could potentially build a more useful set of tools (e.g., rubrics and anchor modules representing different score levels) that are more generalizable and more scalable. The definition of "rigor" would continuously be refined and the anchor modules selected to represent different types of rigor refreshed on a more frequent basis. By engaging teachers to be more than "passive consumers or inexpert students" (Bentley, 2010, p.31), their collective contributions might impact the effectiveness, usability, and viability of the LDC templates and tools.

## References

Bentley, T. (2010). Innovation and diffusion as a theory of change. In Hargreaves, A., Lieberman, A., Fullan, M., & Hopkins, D. (Eds.) *Second International Handbook of Educational Change*. Dordrecht, Heidelberg, London, New York: Springer (pp.29-46).

Iowa Department of Education (2012, Jan). Authentic Intellectual Work Evaluation. Overview and Impact. Iowa City: author.

King, M.B., Schroeder, J., & Chawszczweski, D. (2001, September). Authentic Assessment and Student Performance in Inclusive Schools. Brief No. 5. Madison, WI: Research Institute on Secondary Education Reform for Youth with Disabilities.

Matsumura, L.C., Garnier, H., Pascal, J., & Valdes, R. (2002). Measuring Instructional Quality in Accountability Systems: Classroom Assignments and Student Achievement, *Educational Assessment, 8*(3), 207-229. Accessed at: http://dx.doi.org/10.1207/S15326977EA0803_01

Neumann, F.M., Lopez, G., Bryk, A. (1998). The quality of intellectual work in Chicago schools: A baseline report. Chicago, IL: Consortium on Chicago Schools Research.