A Discussion of Gender Diversity in Computer-Based Assessment

David Crusoe
Harvard Graduate School of Education

Consider my surprise when analysis of the *River City* (Harvard Graduate School of Education [HGSE], 2003) user log data – the text files generated by students' synchronous conversations – highlighted possible gender differences in how users utilized the system.[1] Based on evidence within the data, it could be that some students, by gender, extracted different information from their experience than other students. I theorized: if one particular demographic used the system differently than another, how would a teacher equitably assess all students' progress with a single assessment? What form might such an assessment assume?

Collins, Hawkins, and Frederiksen (1993) provide that a well-rounded assessment could be applied through using multiple media to capture different information, such as by using a computer to "capture real-time problem-solving and learning in interactive contexts," with pencil and paper tests or video footage adding additional dimensions of task-related appraisal. This essay will focus on a single aspect of Collins et al.'s (1993) rounded assessment plan: the capture of user data within interactive computer-based learning software.

In considering the differences present in *River City* "action" data, I suggest it's possible there are multiple layers of means and modes that contribute to the inequities that may appear in assessment activities. Perhaps, by necessitating the use of computer-based learning and achievement opportunities, it's possible to artificially disadvantage a population based upon their less-adept manipulation and comprehension of the computer program, their less-thorough synthesis of information, or an ill-match between the user and assessment interface.

Therefore, the *motivation for this short essay is to comprehend the nature of possible media-induced learning and assessment disadvantages, to explore possible gender effects that may exist within media, and to consider the construction of equitable, gender-neutral computer-based learning and assessments.*

One entry into exploring differences between media is to understand the *test mode effect*. The *test mode effect* is the observation that performance tests of similar knowledge yield different results when conducted through computer-based or pencil and paper activities (Clariana & Wallace, 2002). It's possible that learners who are unfamiliar with a particular interface – either written or computer-based – will demonstrate a lesser proficiency when testing with the unfamiliar assessment conduit (Watson, 2001). While much of the focus has been directed toward exploring differences between pencil and paper-based and computer-based assessments, this test-mode effect may also account for differences within or between computer-based activities.

[1] These emerge through exploring data-capture logs of students from a single teacher; n=97. Similarities and differences are present within users' usage, by gender. For example, both male and female citizens (students) spoke with similar proportions of male and female "residents." However, females were much more talkative with in-game bots (mean = 44 conversations, median = 38, mode = 27) than males (mean = 24 conversations, median = 21, mode = 5).

Admittedly, accusations of gender divisions appear frequently in technology discussions; it's difficult to avoid the stereotypical notion that females are less-savvy computer users. Indeed, research suggests that gender gaps have closed rapidly. According to a recent study, 71% of American youths ages 12 to 17 "go online," and there are few statistical differences between the volume of "boy versus girl" Internet use (Lenhart, Rainie, & Lewis, 2001). Yet despite striking similarities in volume of computer use, one must also consider *which activities* are dominant for each gender. For example, the data show that boys are much more likely than girls (75% to 57%) to play online games or look for items to buy (77% to 55%) (Lenhart et al., 2001). A similar finding by Williamson and Facer (2004) illustrates that in a given week, 33% of boys and just 13% of girls are regular computer game players. These findings are significant in that certain genders may be exposed more frequently to conditions that appear in learning and assessment software, and so artificially excel at those activities.

I posit that innate knowledge of the assessment interface, comfort with the system and savvy with the learning and assessment activity would be incredibly helpful in a learner's utilization of a computer-based assessment interface; essentially, that the test-mode effect could be refined for computer-based approaches as the "interface-mode effect." One example of this effect can be found in a study conducted by Littleton and his colleagues. In this study, a learning simulation was developed with two separate interfaces. One interface featured stereotypically female icons; the other featured stereotypically male icons.[2] The study concluded that girls realized a substantial performance improvement within the "female" interface, whereas males demonstrated similar performance results in both (Yates & Littleton, 1999).

This effect might be understood by exploring three key literacies: first, the learner's ability to engage the software interface in a relevant and meaningful way; second, a learner's sociocultural competency in "reading" and analyzing their engagement; and third, the learner's ability to translate learning into assessable output within the software environment.

To the first and second points, there is some evidence to suggest that both males and females engage in different styles of thinking; boys typically tinker with their available tools to form new constructs, whereas females frequently engage in "relational play" in which "the focus is on the relationship among people or the relationship between people and objects" rather than on modifying the object itself (Caleb, 2000). Additional evidence for stylistic differences is provided in a synthesis of research by Calvert, Mahler, Zehnder, Jenkins, and Lee (2003). Calvert et al. suggests that "boys appear to learn best from playful activities [of content]... and girls learn better from verbal summaries of that content" (2003, p. 630). If these differences are truly existent, then there may be a difference between males' and females' engagement with certain computer-based learning environments.

For instance, the Multi-User Virtual Environment (MUVE) *River City* curriculum requires its users to discover why virtual residents are falling ill; data that supports several competing hypothesis are available for extraction from the environment (HGSE, 2003). Yet, a majority of

[2] The stereotypically "female" version of the learning game featured bears in the likeness of the "Carebears," such as the "Airbear," "Ponybear" and "Waterbear;" the stereotypically "male" version of the learning game featured pilots, drivers and captains.

data provided to learners is made available through conversations with situated virtual agents, who offer descriptions of the epidemic. Log data captured as part of several science classes' interaction (n=96 students) illustrated that girls interacted more heavily with virtual agents than boys (Crusoe, 2005) and that within the limited array of available conversation strings, girls were conversational, while boys were quite direct. If it's the case that females are indeed more apt to engage in "conversational" activities than boys, girls may be afforded a significant performance advantage within this activity.

Another example of the interface-mode effect can be found by examining *GenScope*[3], "…an open-ended exploratory software tool that students can use to investigate a variety of phenomena in genetics" (Horwitz et al., 1998). By manipulating genetic variables of software-simulated creatures, students are expected, for example, to learn specific patterns of reasoning with respect to cause-to-effect variation in genetic evolution. Horwitz (1999) intends for this software to provide a teaching model – a semi-rigid, simplified structure to engage students in "reasoning about the subject matter in unfamiliar ways." This, he reasons, requires an interface that is "not often user-friendly in the normal sense of the word," and objects with preprogrammed "behaviors" (Horwitz, 1999). These complex features, he surmises, help students to acquire critical skills by forcing their thinking in a specific direction. As this software requires skill in tinkering, rather than relational activity, boys might be afforded the performance advantage.

In both instances, it's possible a specific gender might demonstrate proficiency over another; the relational play required by the *River City* interface might favor those who pursue conversational relations, while the *Biologica*[4] interface, the second generation of *GenScope*, might allow those adept at tinkering an advantage[5]. Therefore, it's important to consider the extent to which a learner can meaningfully engage and "read" a software interface, and translate their literacy into actionable output within the given constraints. If it's true that gender differences exist in interface use, one must strongly consider equity, or the establishment of methods to ensure similar learning and assessment results.

One method to ensure equity is through the use of a "cognitive audit trial;" that is, information about where a user has been and what he or she has done. That way, if particular users engaged in using the system differently than assessors might have anticipated, this could be considered for assessment purposes. Audit trails provide a reliable source for exploring a learner's thinking, but do not necessarily ensure equity. For example, a system could engage a learner in *only* "relational" engagements, and maintain a detailed log of interactions. Subtle gender differences might not be highlighted in the audit results, or be recognized by the assessor. Certainly then, software that enables inspection of trails must explicate not simply *what* an assessor (e.g., a teacher) might find, *but also how and why he or she might be interpreting and utilizing the results*. For example, is the assessor using the data for summative evaluation at the end of a unit or to alter ongoing instructional practice? Yet, this is a step after the fact; a summative look at the data presupposes that learners have already engaged (an in some cases completed) the

[3] For more information about *GenScope*, see http://genscope.concord.org
[4] For more information about *Biologica*, see http://biologica.concord.org
[5] One question not addressed in this essay is why; why do these divides exist? Research suggests that media impart stereotypes on children at a young age, and that stereotypes such as "relational" and "tinkerer" are reinforced at a similarly early age.

required tasks, and may not benefit from a reflective look at their experience without the opportunity to modify their interaction.

Therefore, cognitive audit trails might play one role, but not the entire role, in ensuring equitable assessments. At the heart of the "interface mode effect" would be a prevailing supposition for how users should learn and interact within the system. Since learning must be extracted (and later assessed) it would be important to build redundancy in knowledge. If an answer could be gained by speaking with a situated agent (e.g., Dr. Jones in *River City*), perhaps it could be similarly discovered through exploration in the MUVE and combining the correct visual information. This might agree with Collins et al. (1993) that a testing system provides all learners areas in which to "excel."

As Horowitz (1999) concludes, "the most important question that still confronts us in the use of computer based manipulative is, 'What are students learning?'." In light of possible gender differences in interaction styles, there's strong need to ask a second, equally important, question: "Which students are learning?" Software design must transcend the paper-based versus computer-based test-mode effect analysis; computer-based learning and assessment activities are clearly gaining ground. What must be understood more readily, as the research is quite thin, is how different learners will engage computer-based assessments, and how computer-based learning systems and assessments should best ensure equitable experiences for all involved. Clearly, it's possible that even broad demographic measures such as gender could, under the influence of strong cultural bias, highlight differences in need and interaction.

Therefore, to develop equity-minded computer-based learning and assessment environments, developers should pause to consider which learners would best engage the software interface, including methods to ensure the interface is legible for all cultural demographics. Development should also require learners to glean information and translate their learning into meaningful output through several methods, rather than to rely on a single input/output model that might preclude subgroups from equal success. By constructing software packages and computer-based assessments that adequately address these critical questions, one can begin to address possible inequities that exist within current learning and assessment systems.

References

Caleb, L. (2000). Design technology: Learning how girls learn best. *Equity and Excellence in Education, 33*(1), 22–25.

Calvert, S. L., Mahler, B. A., Zehnder, S. M., Jenkins, A., & Lee, M. (2003). Gender differences in preadolescent children's online interactions: Symbolic modes of self-presentation and self-expression. *Journal of Applied Developmental Psychology, 24*(6), 627-644.

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology, 33*(5), 593-602.

Collins, A., Hawkins, J., & Frederiksen, J. R. (1993). Three different views of students: The role of technology in assessing student performance. *Journal of the Learning Sciences, 3*(2), 205-217.

Crusoe, D. (2005). *Big Talk in a Small Town: An analysis of Human-Computer and Human-Human Communication in the River City Multi-User Virtual Environment.* Unpublished manuscript, Cambridge, MA.

Harvard Graduate School of Education. (2003). *Multi-User Virtual Environment Experiential Simulator*. Retrieved May 5, 2005, from http://muve.gse.harvard.edu/muvees2003/index.html

Horwitz, P. (1999). Designing computer models that teach. In W. Feurzeig & N. Roberts (Eds.), *Modeling and simulation in science and mathematics education* (pp. 179-196). New York: Springer.

Horwitz, P., Schwartz, J., Kindfield, A. C. H., Yessis, L. M., Hickey, D. T., Heidenberg, A. J., et al. (1998). *Implementation and Evaluation of the GenScope Learning Environment: Issues, Solutions, and Results.* Paper presented at the Third International Conference on the Learning Sciences, Charlottesville, VA.

Lenhart, A., Rainie, L., & Lewis, O. (2001). *Teenage life online: The rise of the instant-message generation and the Internet's impact on friendships and family relationships*. Washington, DC: Pew Internet and American Life Project.

Watson, B. (2001). Key factors affecting conceptual gains from CAL. *British Journal of Educational Technology, 32*(5), 587-593.

Williamson, B., & Facer, K. (2004). More Than 'Just a Game': the implications for schools of children's computer games communities. *Education, Communication & Information, 4*(2/3), 255-270.

Yates, S. J., & Littleton, K. (1999). Understanding computer game cultures: A situated approach. *Information, Communication and Society, 2*(4), 566-583.