Gaming the System: A Retrospective Look

Ryan S.J.d. Baker
Worcester Polytechnic Institute
100 Institute Road
Worcester, MA 01609
+1 508-831-5355
rsbaker@wpi.edu

ABSTRACT

In this article, I present some retrospective thoughts on the methodological directions and decisions that have influenced my research (and that of my colleagues) on gaming the system over the last eight years, focusing on four dimensions: the power of terminology, conducting more open-ended quantitative studies, discovery with models, and conducting research in the real world. I discuss the directions gaming the system research has taken, and some of the factors driving these directions.

1. INTRODUCTION

In the last eight years, research on gaming the system – attempting to succeed in an educational task by systematically taking advantage of properties and regularities in the system used to complete that task, rather than by thinking through the material – has increased greatly in quantity. The phenomenon was first reported by Tait, Hartley, & Anderson (1973), within the context of computer-aided instruction administered by teletype. In the 1990s and early years of the third millennium AD, it was again reported as a phenomenon (Schofield, 1995; Miller, Lehman, & Koedinger, 1999; Aleven & Koedinger, 2000, 2002). In 2004, two articles appeared which coined the term 'gaming the system' (Baker, Corbett, Koedinger, & Wagner, 2004; Baker, Corbett, & Koedinger, 2004); these two articles established the construct of gaming the system (as opposed to the specific behaviors composing gaming), established links between gaming and learning (it is worth noting that these links had been previously established by Aleven and Koedinger in 2000 with regard to help abuse, a specific type of gaming behavior), and established that a machine-learned model of gaming behavior could be developed and validated in terms of its predictive power. 2004 also saw the publication of an article by Aleven and colleagues, which advanced a knowledge-engineered model of two gaming behaviors, along with several help-seeking constructs. After this point, interest in gaming the system emerged as a larger-scale topic in research, leading to dozens, if not hundreds, of articles on gaming behavior and closely related constructs. It is not the goal of this brief article to review all of the research on gaming the system, or to provide a complete review of those articles. Instead, I discuss some of the conceptual innovations and directions that our laboratory and colleagues have taken in these years that in my opinion have beneficially influenced the development of gaming the system as a research area.

2. THE POWER OF TERMINOLOGY

One of the major shifts in 2004, as mentioned in the introduction, was the shift from discussing specific behaviors to discussing the more general phenomenon of "gaming the system." Prior to 2004, specific gaming behaviors were given names such as "hint abuse" (Aleven & Koedinger, 2000), or "standard-goal"/"standard game interaction" (Miller, Lehman, & Koedinger,1999). Some researchers and practitioners referred to the general phenomenon of gaming as "sleazing," a term coined by a high-school student, but which has a pejorative quality that prevented it from being widely used in publication. Shifting to the term "gaming the system" facilitated researchers in seeing the broader context of the construct, and its applicability to types of learning systems beyond the intelligent tutoring systems where it was first coined (cf. Cheng & Vassileva, 2006). I would like to be able to take credit for a rich and conceptually powerful term like "gaming the system," which makes useful connections to similar concepts in other fields (cf. Figlio & Getzler, 2002; Bevan & Hood, 2006). However, I can't. At the time of my first research on this construct, I decided that a new term was needed but could not think of a good term, so I held a contest. "Gaming the system" was proposed by Desney Tan, currently a Senior Researcher at Microsoft Research. He won the prize (I can't remember what the prize was anymore); I've gone on to use the term he recommended. It seems to have been a useful terminological shift.

3. OPEN-ENDED QUANTITATIVE STUDIES

One piece of advice often given in introductory statistics textbooks is to always design an experiment based on a single hypothesis, tightly designed so that you can learn only one thing: whether the hypothesis is right or wrong.

This is common advice. It's also bad advice – or to be more positive, over-generalized advice. A corollary was given to me by a famous senior colleague at Carnegie Mellon University, as he was quitting my thesis committee, less than 72 hours before my thesis proposal was scheduled[1]. He told me that my strategy of finding a phenomenon in the world (in this case gaming the

[1] A good reason to have one more member on your thesis committee than the minimum!

system), and then trying to understand what caused that phenomenon, was a mistake. He said that gaming the system is so complex that it no doubt has many causes and many contributing factors – and the evidence since 2004 appears to bear him out! (e.g. Arroyo & Woolf, 2005; Beal et al., 2006; Baker, 2007; Baker, Walonoski, et al., 2008; Baker et al., 2009; Baker, D'Mello, Rodrigo, & Graesser, 2010; Gong et al., 2010; Muldner et al., 2011). Instead, he recommended that a researcher should decide what mechanism he or she wants to prove exists, and design an experiment tailored to prove that they are correct.

The point I'd like to make is not that researchers should avoid cleanly defined, unconfounded experimental studies; clearly, this type of study has a place (and arguably needs to occur more often in evaluation of educational programs, a point made in the creation of the Society for Research on Educational Effectiveness). It is also not to say that a researcher with a strong vision about a mechanism's existence should not attempt to design a study that provides a conclusive example.

But these approaches have limits, and there is also a place for more exploratory research – the kind of research that opens new possibilities and directions. Starting from open questions about what causes a poorly-understood phenomenon and examining which of many possible explanations appears to have potential – this may lead to faster progress than laboriously constructing single-hypothesis experiments when there are a wide range of possible hypotheses. In these cases, it may be better to investigate many hypotheses in a first study, and then conduct a more conclusive experimental study afterwards.

At the same time, we have found it advantageous to use quantitative methodologies even when conducting exploratory research. Qualitative research, for all its interpretive power, is limited to individual cases, and may lead to researchers focusing on intriguing single cases as opposed to patterns that show up across students or situations. My colleagues and I often do quick qualitative exploration prior to more quantitative and larger-scale data collection and/or analysis. Quantitative methodologies, though more frequently discussed for experiments with pre-designed hypotheses, can still be conducted in a valid fashion in open-ended research paradigms. One method for accomplishing this goal is to conduct limited numbers of statistical tests, or using post-hoc corrections or checks of various sorts. This can include the Tukey post-hoc test (Jaccard, Beck, & Wood, 1984), Monte Carlo simulations (Metroplolis & Ulam, 1949), and False Discovery Rate corrections (Benjamini & Hochberg, 1995). Bonferroni tests, by contrast, are extremely conservative and typically indicate that nothing is significant, even when all other methods agree there is something present (Perneger, 1998). Another strategy is to use methods from the data mining literature; validation methods such as cross-validation can establish the degree of generalizability of findings in a clear fashion (Efron & Gong, 1983). It is worth noting that conducting cross-validation at the correct level is essential. Cross-validating at the level of individual actions – ignoring which student made them – can lead to a model only being validated for generalizability to new actions from the original set of students. Student-level cross-validation is more useful, typically, as it establishes generalizability to new students; classroom-level or lesson-level cross-validation can be even stronger evidence, when feasible.

Our use of more open-ended quantitative methods shows up in our first observational study of gaming (Baker, Corbett, Koedinger, & Wagner, 2004), where we studied not just gaming, but also on-task conversation and several forms of off-task behavior. A study designed to test just one hypothesis would not even have included gaming the system; our original goal was to study off-task behavior (a focus that shows up in the title of that first paper – Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game The System"[2]). However, including gaming the system and several other behaviors in our coding scheme was very little additional effort, and turned out to be very productive.

A more extreme example can be seen in our group's work to determine which factors of an intelligent tutoring system lead to different amounts of gaming the system (Baker et al., 2009). In that study, we annotated 79 potential differences between tutor lessons, using a combination of human labeling and data mining. We then conducted factor analysis, and found a factor that predicted a considerable amount of the variance in gaming. We are currently conducting an experimental study where we systematically re-designed a tutoring system to eliminate tutor features found to be associated with greater gaming. Our hypothesis is that the modified tutor will be gamed less often (and hopefully, that students will correspondingly demonstrate greater learning), without requiring the active and often disruptive types of intervention previously used to address gaming (e.g. Walonoski & Heffernan, 2005; Baker et al., 2006; Arroyo et al., 2007; Roll et al., 2011). This pair of studies forms an example of following an exploratory investigation with a more controlled experiment.

4. DISCOVERY WITH MODELS

One direction which I am convinced has great potential for the future of research in the learning sciences is discovery with models. Discovery with models has been around for a while within work at the intersection between data mining/computing and other scientific fields, but is a considerably newer development in the learning sciences. It is defined in Baker and Yacef (2009) as when "a model of a phenomenon is developed through any process that can be validated in some fashion… and this model is then used as a component in another analysis…"

As an example, the gaming detector for Cognitive Tutors which our laboratory developed (cf. Baker, Corbett, & Koedinger, 2004) and validated (cf. Baker, Corbett, Roll, & Koedinger, 2008) has turned out to be a useful tool for studying many questions using discovery with models:

- Do situational factors or individual differences predict more of the variance in gaming the system? (Baker, 2007; also see Gong et al., 2010; Muldner et al., 2011)
- Why do students game the system? (Baker, Walonoski, et al., 2008; also see Beal et al., 2007)
- Do gaming the system and off-task behavior impact learning in different ways? (Cocea, Hershkovitz, & Baker, 2009)

[2] A title which I have regretted ever since; my colleagues and I abandoned the theoretical lens that gaming is a type of off-task behavior after this single paper, but it has shown up in citations of gaming the system to the present day.

- Do urban, rural, and suburban students differ in their degree of disengaged behavior? (Baker & Gowda, 2010)
- Does personalization in learning software improve engagement? (Walkington & Maull, 2011)
- Are fast non-gaming actions predictive of robust learning? (Baker, Gowda, & Corbett, 2011)

The studies listed above would have been much more difficult and time-consuming to conduct without an automated and validated detector of gaming the system; the development of this detector has been a very useful tool for speeding and facilitating research in this area.

5. RESEARCH IN THE REAL WORLD

A final point I would like to make is that conducting research in real-world settings, as opposed to in laboratory settings, is a decision that leads to many additional difficulties, in terms of administering and controlling the research, as well as in finding sites and getting approval to conduct the research. However, there are several advantages to conducting research in real-world settings, when studying learning and related phenomena. Simply put, people behave differently when they are taken out of the natural context of a task, and when their motivation for completing the task is artificial (though the second of these limitations can be mitigated by conducting laboratory research on people with genuine motivation for the task – cf. D'Mello, Lehman, & Person, 2010).

To illustrate this, it is worth telling an anecdote. One colleague asked me, several years ago, why it was worth studying gaming the system at all. She explained to me that when she brought students into the lab to use her intelligent tutor, if the student started gaming, she would tell them to stop, and they would. This can be seen as clear (if qualitative) evidence that gaming is different between the lab and the field – during our observations, when a teacher asks a student to stop gaming, the student typically resumes gaming as soon as the teacher walks away. A phenomenon like gaming could be expected to vary considerably in many ways in the lab; it might occur in different circumstances, for different reasons, and for different lengths of time. Hence, it's questionable what value lab research would have for helping us understand gaming as a phenomenon.

But even in considering phenomena that seem much more general – such as affect – it is not clear that results from laboratory studies can be assumed to be representative of the real world. For example, there have been several lab studies in the USA and classroom studies in the Philippines, showing evidence for "vicious cycles" during affect, where a student becomes bored and stays bored (cf. D'Mello, Taylor, & Graesser, 2007; McQuiggan, Robison, & Lester, 2008; Baker, D'Mello, Rodrigo, & Graesser, 2010). However, recent studies conducted in classrooms in the USA suggest that in these settings, students regulate their boredom with off-task behavior, considerably reducing the incidence of these vicious cycles (cf. Baker, Moore, et al., in press). This finding, if replicated, would suggest that there are considerable challenges to trusting laboratory findings in the real-world, in this domain – and also that there are considerable challenges to generalizing real-world findings on affect across cultures.

6. CONCLUSION

In this article, I have presented some retrospective thoughts on the methodological directions and decisions that have influenced my laboratory's research on gaming the system over the last eight years, focusing on four dimensions: the power of terminology, conducting open-ended quantitative studies and analyses, discovery with models, and conducting research in the real world. Research on gaming continues to change over time. Several directions currently seem to be emerging: the consideration of gaming the system as one disengaged behavior among many (and correspondingly, the consideration of how gaming differs from other disengaged behaviors), the analysis of the links between gaming and affect, and the integration of gaming the system into broader models of meta-cognition and motivation (a direction present since Aleven et al., 2004), among many directions. I hope that this retrospective is useful to those interested in gaming the system and related areas, and look forward to what the next eight years bring, in terms of research in this area.

7. ACKNOWLEDGMENTS

I would like to thank Lisa Rossi and Mercedes Rodrigo for their comments on this paper, and Desney Tan for coining the term "gaming the system," back in 2003. I would also like to thank all of my colleagues and collaborators in the last eight years, who worked with me in researching gaming the system; this work could not have happened without them.

8. REFERENCES

[1] Aleven, V., Koedinger, K. R. (2000). Limitations of Student Control: Do Students Know when they need help? *Proceedings of the 5th International Conference on Intelligent Tutoring Systems*, ITS 2000, 292-303.

[2] Aleven, V., Koedinger, K. R. (2002). An effective meta-cognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*, 26 (2), 147-179.

[3] Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2004). Toward tutoring help seeking: Applying cognitive modeling to meta-cognitive skills. In J. C. Lester, R. M. Vicario, & F. Paraguaçu (Eds.), *Proceedings of Seventh International Conference on Intelligent Tutoring Systems*, ITS 2004 (pp. 227-239). Berlin: Springer Verlag.

[4] Arroyo, I., Ferguson, K., Johns, J., Dragon, T., Meheranian, H., Fisher, D., Barto, A., Mahadevan, S., and Woolf. B.P. (2007) Repairing Disengagement with Non-Invasive Interventions. *Proceedings of the 13th International Conference on Artificial Intelligence in Education*, 195-202.

[5] Arroyo, I., Woolf, B. (2005) Inferring learning and attitudes from a Bayesian Network of log file data. *Proceedings of the 12th International Conference on Artificial Intelligence in Education*.

[6] Baker, R.S.J.d. (2007) Is Gaming the System State-or-Trait? Educational Data Mining Through the Multi-Contextual Application of a Validated Behavioral Model. Complete On-Line *Proceedings of the Workshop on Data Mining for User Modeling at the 11th International Conference on User Modeling 2007*, 76-80.

[7] Baker, R.S., Corbett, A.T., Koedinger, K.R. (2004) Detecting Student Misuse of Intelligent Tutoring Systems.

Proceedings of the 7th International Conference on Intelligent Tutoring Systems, 531-540.

[8] Baker, R.S.J.d., Corbett, A.T., Koedinger, K.R., Evenson, S.E., Roll, I., Wagner, A.Z., Naim, M., Raspat, J., Baker, D.J., Beck, J. (2006) Adapting to When Students Game an Intelligent Tutoring System. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 392-401.

[9] Baker, R.S., Corbett, A.T., Koedinger, K.R., Wagner, A.Z. (2004) Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game The System". *Proceedings of ACM CHI 2004: Computer-Human Interaction*, 383-390.

[10] Baker, R.S.J.d., Corbett, A.T., Roll, I., Koedinger, K.R. (2008) Developing a Generalizable Detector of When Students Game the System. *User Modeling and User-Adapted Interaction*, 18, 3, 287-314.

[11] Baker, R.S.J.d., de Carvalho, A.M.J.A., Raspat, J., Aleven, V., Corbett, A.T., Koedinger, K.R. (2009) Educational Software Features that Encourage and Discourage "Gaming the System". *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 475-482.

[12] Baker, R.S.J.d., D'Mello, S.K., Rodrigo, M.M.T., Graesser, A.C. (2010) Better to Be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners' Cognitive-Affective States during Interactions with Three Different Computer-Based Learning Environments. *International Journal of Human-Computer Studies*, 68 (4), 223-241.

[13] Baker, R.S.J.d., Goldstein, A.B., Heffernan, N.T. (in press) Detecting Learning Moment-by-Moment. To appear in *International Journal of Artificial Intelligence in Education.*

[14] Baker, R.S.J.d., Gowda, S.M. (2010) An Analysis of the Differences in the Frequency of Students' Disengagement in Urban, Rural, and Suburban High Schools. *Proceedings of the 3rd International Conference on Educational Data Mining*, 11-20.

[15] Baker, R.S.J.d., Gowda, S., Corbett, A.T. (2011) Towards predicting future transfer of learning. *Proceedings of 15th International Conference on Artificial Intelligence in Education*, 23-30.

[16] Baker, R.S.J.d., Moore, G., Wagner, A., Kalka, J., Karabinos, M., Ashe, C., Yaron, D. (in press) The Dynamics Between Student Affect and Behavior Occuring Outside of Educational Software. To appear in *Proceedings of the 4th bi-annual International Conference on Affective Computing and Intelligent Interaction*.

[17] Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., Koedinger, K. (2008) Why Students Engage in "Gaming the System" Behavior in Interactive Learning Environments. *Journal of Interactive Learning Research*, 19 (2), 185-224.

[18] Baker, R.S.J.d., Yacef, K. (2009) The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, 1 (1), 3-17.

[19] Beal, C. R., Qu, L., & Lee, H. (2006). Classifying learner engagement through integration of multiple data sources. *Proceedings of the 21st National Conference on Artificial Intelligence*, July 16-20, 2006, Boston MA.

[20] Benjamini, Y., Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57 (1), 289-300.

[21] Bevan, G., Hood, C. (2006) What's measured is what matters: Targets and gaming in the English public health care system. *Public Administration*, 84 (3), 517-583.

[22] Cheng, R., Vassileva, J. (2006) Design and evaluation of an adaptive incentive mechanism for sustained educational online communities. *User Modeling and User-Adapted Interaction*, 16 (3-4), 321-248.

[23] Cocea, M., Hershkovitz, A., Baker, R.S.J.d. (2009) The Impact of Off-task and Gaming Behaviors on Learning: Immediate or Aggregate? *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 507-514.

[24] D'Mello, S. K., Lehman, B., & Person, N. (2010). Expert Tutors Feedback is Immediate, Direct, and Discriminating. *Proceedings of the 23rd Florida Artificial Intelligence Research Society Conference (FLAIRS-23)*. (595-604). AAAI Press

[25] D'Mello, S. K., Taylor, R., & Graesser, A. C. (2007). Monitoring Affective Trajectories during Complex Learning. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 203-208). Austin, TX: Cognitive Science Society.

[26] Efron, B., Gong, G. (1983) A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *The American Statistician, 37* (1), 36-48.

[27] Figlio, D.N., Getzler, L.S. (2002) Accountability, Ability and Disability: Gaming the System, *National Bureau of Economic Research*.

[28] Gong, Y., Beck, J.E., Heffernan, N.T., Forbes-Summers, E. (2010) The impact of gaming (?) on learning at the fine-grained level. *Proceedings of the 10th International Conference on Intelligent Tutoring Systems*, 194-203.

[29] Jaccard, J., Becker, M.A., Wood, G. (1984) Pairwise multiple comparison procedures: A review. *Psychology Bulletin*, 96 (3), 589-596.

[30] McQuiggan, S.W., Robison, J.L., Lester, J.C. (2008) Affective transitions in narrative-centered learning environments. *Intelligent Tutoring Systems*, 5091, 490-499.

[31] Metropolis N., Ulam, S. (1949) The Monte Carlo Method. *Journal of the American Statistical Association*, 44 (247), 335-341.

[32] Miller, C.S., Lehman, J.F., Koedinger, K.R. (1999) Goals and learning in microworlds. *Cognitive Science*, 23 (3), 305-336.

[33] Muldner, K., Burleson, W., Van de Sande, B., VanLehn, K. (2011) An analysis of students' gaming behaviors in an intelligent tutoring system: Predictors and impacts. *User Modeling and User-Adapted Interaction*, 21(1-2), 99-135.

[34] Perneger, T.V. (1998) What's wrong with Bonferroni adjustments. *British Medical Journal*, 316, 1236-1238.

[35] Roll, I., Aleven, V., McLaren, B.M., & Koedinger, K.R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, 21, 267-280.

[36] Schofield, J. W. (1995). Computers and Classroom Culture. Cambridge, UK: *Cambridge University Press*.

[37] Tait, K., Hartley, J.R., Anderson, R.C. (1973) Feedback procedures in computer-assisted arithmetic instruction. *British Journal of Educational Psychology*, 43 (2), 161-171.

[38] Walkington, C., Maull, K. (2011) Exploring the assistance dilemma: The case of context personalization. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 90-95.

[39] Walonoski, J.A., Heffernan, N.T. (2006) Prevention of Off-Task Gaming Behavior in Intelligent Tutoring Systems. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 722-724.