Exploration of Costs and Benefits of Predictive Human Performance Modeling for Design

Bonnie E. John (bej@cs.cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University
5000 Forbes Ave. Pittsburgh, PA 15213

Tiffany S. Jastrzembski (tiffany.jastrzembski@wpafb.af.mil)
Air Force Research Laboratory
2698 G Street, Building 190
Wright-Patterson AFB, OH 45433-7604

Abstract

Human performance modeling promises to be a valuable tool for early evaluation of user interface designs, predicting different performance for different design alternatives and, recently, different performance on a single design between younger and older adults (Jastrzembski & Charness, 2007; Jastrzembski, et al., 2010). When using modeling in the development process, the costs of creating models must be traded-off against the accuracy needed to guide design choices. It is therefore a meaningful exercise to examine and weigh the costs and benefits of different modeling approaches, to provide practitioners information to help them choose the modeling approach best suited for their needs. We compare younger and older adult human performance data captured from dialing and text-messaging tasks, across two mobile phones, against age-specific GOMS (Card, Moran & Newell, 1983) and various CogTool models (John, et. al. 2004), and examine the trade-offs between time and effort required to build those models and the predictive validity each model produces.

Keywords: predictive human performance modeling, design.

Introduction

Research in computational cognitive process modeling continues to progress by creating models able to account for human data on more tasks across more domains, often through years of effort by PhD students, post-docs and/or senior researchers. However, when practitioners wish to use cognitive modeling approaches in user interface (UI) design, issues of costs and benefits become a stark reality. It is therefore often necessary for the practitioner to base the selection of a modeling approach by trading off the costs of producing the human performance models against the desired accuracy of the predictions of those models.

The costs of producing models for design include how much knowledge the practitioner must have to develop an appropriate cognitive model in the task domain of interest for the intended user group, learning and understanding the modeling theory that underlies a modeling tool, learning how to use the modeling tool itself, and the time it takes to accurately implement the models after learning the modeling theory and associated tools. Benefits include the ability of a modeling approach to detect differences between design alternatives and the ability to make accurate predictions of quantitative measures of performance (e.g., time for a skilled user to execute a task or number of errors).

As the consumers of interactive systems age it is becoming economically important to evaluate designs specifically for the older adult. Thus, an additional concern we address in this paper, are costs related to modifying existing modeling approaches and tools to account for the human processing capabilities of the older adult. Given the range of knowledge, time, and effort required to make these model modifications, this paper compares the quality of predictions against the efficiency of each approach.

To put these issues into context, consider a practitioner who is under a tight deadline to choose a final design that is efficient for both younger and older adults from among several design alternatives. A less time-intensive modeling approach may be required to fit into the development life cycle, even if use of that modeling approach comes at the cost of producing less accurate predictions. This paper begins to address cost-benefit concerns by assessing the accuracy of a variety of modeling approach predictions against empirical data, and examining the costs incurred to produce those predictions.

The Designs, Tasks & Empirical Results

We chose to examine two tasks on two mobile phones because Jastrzembski and Charness (2007) provides pre-existing empirical data for younger and older adults. The tasks are dialing a 10-digit phone number (*dialing*) and sending a text message to a person in the contact list (*texting*). Participant groups included a sample of younger adults (M_{age} = 20) and older adults between the ages of 60-75 (M_{age} = 69). The purpose of their study was to validate elemental model human processing parameters updated to account for the older adult, which had been estimated through a comprehensive literature review. These parameter values were then used to build age-specific GOMS (Goals, Operators, Methods, and Selection rules) models (Card, et al., 1983) to predict skilled performance of younger and older "average" adults in the mobile phone tasks.

Figure 1. Mobile phones studied by Jastrzembski and Charness (2007) and used in this analysis: the Nokia 3595 (left) and the Motorola C155 (right).

Predictions were compared to empirical data at each button press required by the task.[1]

Since GOMS models are designed to predict performance of skilled users on routine tasks, the participants were required to complete extensive practice sessions to ensure that they were skilled in the performance of these tasks on these devices. "Skill" was operationally defined as completing three consecutive trials with less than a 1s deviation from each other. Upon successfully achieving criterion in the practice sessions, participants were then given new stimuli to complete three repeated blocks of five different trials for each task. This allowed the authors to average the human performance data for a single stimulus over three trials – thus producing the empirical findings displayed in Figure 2.

The following results were revealed (Figure 2, Table 1).

- Older adults took significantly longer than younger adults to complete both tasks on both phones.
- Dialing completion times were not significantly different across phones for either age group.
- Text-messaging completion times were significantly longer using the Motorola compared to the Nokia phone for both age groups.

These findings give us an interesting spread of results to assess the evaluation of the designs across modeling approaches from a cost-benefit perspective. In order for a model to be useful in practice, it must account for all three results, i.e., detecting a difference between devices and age groups where this is one in the empirical data and detecting no difference where there isn't.

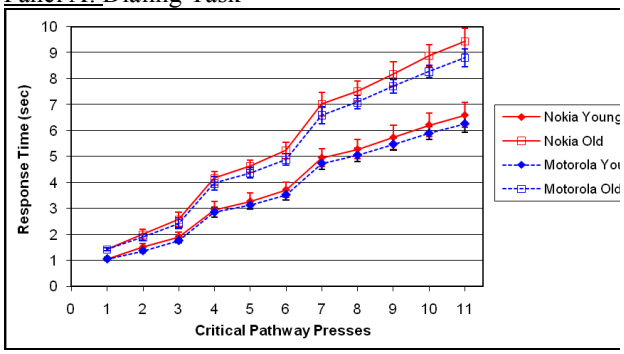
The Modeling Approaches & Their Results

Seven modeling approaches were implemented for the dialing task and four for the texting task, as described below (Table 1 displays completion time results).

GOMS-MHP. A pre-existing model by Jastrzembski & Charness (2007), this approach updated Model Human Processor (MHP) parameters through extensive literature review, to allow GOMS models to predict older adult performance. These models most closely match the "K2" models put forth by Card, et al., (1983, p. 166), where operators are at the level of hundreds of milliseconds, and map closely to MHP cycle times. The cognitive task analysis that underlies these models was informed by observing pilot participants using an eye-tracker while performing the tasks. Eye-fixation operators and subsequent decisions operators were placed in the models guided by these data. These models achieved excellent fits to the human data for tasks, phones, and age groups.

CogTool-OotB. The next modeling approach we examine is CogTool (John, et al., 2004), a tool for prototyping UI designs and automatically producing Keystroke-Level Models (KLM, Card, et al., 1983) through demonstration.

Panel A: Dialing Task



Panel B: Texting Task

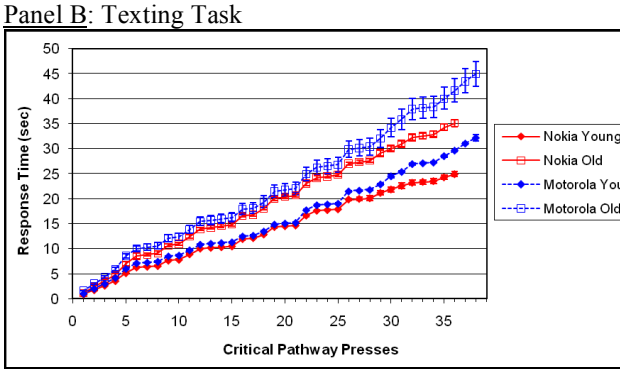


Figure 2. Empirical data for younger and older adults completing tasks on the Nokia and Motorola phones.

KLM is a simplified form of GOMS that sums each key press, *K* (including typing on a keyboard and mouse clicks); pointing movement, *P*; homing movement between devices, *H*; system response time, *R*; and "mental operator", *M* (an averaged amalgamation of visual search, perception and cognitive operations like deciding, recalling commands, etc.), required to do a task.

CogTool automates KLM model construction through a demonstration of a task on a storyboard of a UI, adding perceptual operations in line with Salvucci (2001), and cognitive operations similar to Card, Moran And Newell's Ms[2], called "Think operators." The resulting *script* approximate a KLM produced by an expert modeler. The storyboard and script together compile into an ACT-R model (Anderson & Lebiere, 1998), which runs to produce quantitative predictions of skilled performance time. CogTool allows people with no cognitive psychology or modeling background to make accurate predictions with little variance (John et al., 2004; John, 2010).

This approach used CogTool "out of the box" examining its default predictions without modifications of the script it produced or to any of its parameters. This approach resulted in far better predictions for the texting task than for the dialing task. The remaining approaches progressively add information to this "out of the box" approach.

[1] The original GOMS parameters were set with data from younger adults, therefore we will use the original GOMS parameters for younger adults unless otherwise noted in this paper.

[2] Card, et. al. (1983) set the duration of *M* to 1.35s, but CogTool uses 1.2s because it has separate processes for eye movement and visual perception, which require about 0.15s processing time.

CogTool+KLM. To improve predictions for the dialing task, our third modeling approach brought knowledge of the KLM to bear, editing out Think operators where they violated Card, Moran & Newell's M-placement rule concerning *cognitive units*. We deemed this approach reasonable because people separate US telephone numbers into cognitive units consisting of a 3-digit area code, a 3-digit exchange, and a 4-digit station number. Because *CogTool-OotB* does not automatically identify these units, analysts must use their knowledge to delete unnecessary Think operators from the scripts. (Such modification was reasonable for the dialing task, but not for the texting task where *CogTool-OotB = CogTool+KLM*.)

CogTool+KLM+RatioThink. Since CogTool generates predictions specific to younger adults, it cannot make predictions for older adults without modifications. Therefore, our fourth modeling approach augments CogTool+KLM by applying Hale and Myerson's (1995) findings that older adults take 1.5 times as long as younger adults to process linguistic information. This means that the analyst simply copies the original CogTool+KLM script for a task and edits each Think to be 1.5 times as long as the standard younger adult time (i.e., 1.8s v 1.2s). This resulted in an average absolute percent error of less than 10% for the texting task, but 36% for the dialing task – vastly over predicting the time it takes both young and old to dial a phone number (see Table 1).

CogTool+KLM+RatioThink+ExtremePractice. Reflecting on the previous method's poor fit to the dialing task data, we realized that participants in 2005 would have had almost a lifetime of experience dialing touch-tone phones and substantially less practice sending text messages on mobile devices. Prior research in extreme practice has shown that pauses indicating mental operations almost disappear. Thinking is both getting shorter with practice and also presumably happens in parallel with the perceptual and motor actions necessary to do the task (e.g., Card, et al., 1983, pp. 279-286). Simulating extreme practice is an easy process in CogTool; the analyst simply deletes every Think step in the script except the first (which is still required because the digits must be visually acquired from a sheet of paper). This resulted in predictions that better fit the younger and older adult data. However, these predictions were within 10% of each other, meaning that these models no longer detected the main effect of age.

CogTool+KLM+RatioThink+ExtremePractice+OlderWMcapacity. Our next approach examines the accuracy of a CogTool model created by analysts possessing additional information about older adult performance, as was uncovered by Jastrzembski & Charness' (2007) literature review. That review revealed that the working memory (WM) capacity of older adults is smaller than that of younger adults. This may cause a strategy change in older adults; they may spend more time with written instructions than younger adults, trading off time for accuracy. With this insight, we put the Think steps associated with looking at the paper for the area code, exchange and station digits, back into the older adult dialing task models. This reduced

Table 1. Modeling approach predictions for the mobile phone dialing task with percent deviations from empirical data.

		Nokia				Motorola			
	Abs Avg	Younger		Older		Younger		Older	
Source of data or predictions	**%diff**	**Time (s)**	**%diff**	**Time (s)**	**%diff**	**Time (s)**	**%diff**	**Time (s)**	**%diff**
Dialing Task									
Human Data		6.606		9.442		6.268		8.812	
GOMS-MHP	0.6%	6.559	-0.7%	9.369	-0.8%	6.228	-0.6%	8.804	-0.1%
CogTool-OotB	169.9%	16.451	149.0%	n/a	n/a	18.227	190.8%	n/a	n/a
CogTool+KLM	44.1%	9.171	38.8%	n/a	n/a	9.359	49.3%	n/a	n/a
CogTool+KLM+RatioThink	36.0%	9.171	38.8%	11.571	22.6%	9.359	49.3%	11.759	33.4%
CogTool+KLM+RatioThink +ExtremePractice	15.5%	5.976	-9.5%	6.576	-30.4%	6.302	0.5%	6.902	-21.7%
CogTool+KLM+RatioThink +ExtremePractice +OlderWMcapacity	5.0%	5.976	-9.5%	8.092	-14.3%	6.302	0.5%	8.387	-4.8%
CogTool+KLM+RatioThink +ExtremePractice +OlderWMcapacity +LitReviewACT-Rparameters	6.4%	5.830	-11.8%	9.505	0.7%	6.205	-1.0%	9.520	8.0%
Texting Task									
Human Data		24.905		35.127		32.186		44.991	
GOMS-MHP	0.0%	24.901	0.0%	35.126	0.0%	32.153	0.1%	44.989	0.0%
CogTool-OotB (=CogTool+KLM)	13.9%	27.582	10.7%	n/a	n/a	37.664	-17.0%	n/a	n/a
CogTool+KLM+RatioThink	9.4%	27.582	10.7%	35.382	0.7%	37.664	-17.0%	49.064	9.1%
CogTool-KLM+RatioThink +LitReviewACT-Rparameters	11.7%	27.148	9.0%	37.442	6.6%	37.118	-15.3%	52.177	16.0%

the average absolute percent error to 5% for the dialing task.

CogTool+KLM+RatioThink+ExtremePractice+OlderWMcapacity+LitReviewACTRparameters. The last modeling approach modifies the ACT-R models running under the hood of CogTool. This approach requires both more cognitive psychology knowledge and programming skill. It leverages the aforementioned literature review as well as Jastrzembski, et al.'s (2010) translation and extension of age-specific parameters to ACT-R. We ran CogTool in a development environment rather than as an executable, and edited four specific underlying ACT-R parameters identified by Jastrzembski, et. al. (2010), in order to account for age. We modified the best of the CogTool approaches previously mentioned (*CogTool+KLM+RatioThink+ExtremePractice+OlderWM capacity* for dialing and *CogTool-OotB* for texting), but results produced overall goodness-of-fit values slightly less than other approaches, for both dialing and texting tasks.

Cost and Benefit Metrics

We now assess the costs each modeling approach would incur, based upon the estimated amounts of knowledge, time, and effort required to produce predictions using each method. Benefits are assessed relative to the empirical data collected by Jastrzembski and Charness (2007), which will be considered "*the gold standard*" - that is, the "truth" against which the models will be compared. Costs are assigned a value pertaining to the length of time required to attain the appropriate knowledge base and perform the modeling itself. A *large* cost entails months of experience to learn and/or use the method; a *medium* cost requires weeks of training and use; a *small* cost requires days.

Of course, actual costs to an organization depend on both workforce and resources. For instance, empirical collection of human data is characterized as having a *large* cost in this analysis because many practitioners are not trained in experiment design, they lack data collection laboratories, and they often do not possess statistical packages or analytic know-how to properly interpret the empirical data. These costs may be much smaller for organizations like Google or Microsoft, which already have highly equipped labs, PhD-level experimentalists and statisticians, and a network for recruiting appropriate participants.

In addition, the costs are estimated for moving into a new domain or user group where parameters are not already routinely used in models or built into tools. Many of these estimates would reduce as modeling knowledge increases and tool functionality is enhanced to embody that knowledge. Given these caveats, we identified the following costs for the analyses described in this paper.

Collect Human Data. Cost = Large because of expertise and resource issues discussed above, and because participants must be trained to a skilled level of performance on the tasks and devices studied.

Literature Review. Cost = Large for a full review and meta-analysis (it took Jastrzembski approximately nine months to complete the parameter estimation alone). Cost = Small if only a rule-of-thumb 50% increase (as reported by Hale & Myerson (1995)) is used.

Program a running prototype. Cost = Large due to required programming skill expertise (UI designers often possess graphic design backgrounds rather than a computer science backgrounds to compound the problem).

Measure for Fitts's Law. Cost = Small because estimates of size and distance between all keys are required for movement times to be integrated into models. (Although it does not take days to learn or accomplish this, the sheer tedium bumps this, in our estimation, into a real cost).

Build a Storyboard. Cost = Low because building a storyboard in CogTool (John, et. al., 2004) involves only creating a frame using a picture of what the device looks like, placing button widgets on that frame, and drawing transitions to represent user actions required to accomplish the task. Storyboards for the two phones used in this investigation took the first author about 15 minutes to build.

Know GOMS/MHP. Cost = Large. In the first author's 25 years of experience teaching GOMS, it takes engineers several sessions to learn the typical version of GOMS but requires feedback on multiple exercises and often an apprenticeship with an expert GOMS model builder to be able to produce high-quality models. The GOMS-MHP models assessed here were built with PhD-level knowledge of cognitive psychology guided by eye-tracking observations (Jastrzembski, 2006).

Know KLM. Cost = Small. In the first author's 25 years of experience teaching GOMS, KLM can be taught in a single class session but requires feedback on several exercises to be able to remove mental operators appropriately to account for "cognitive units" (John, 1994). The cost increases to Medium when knowledge of different strategies due to older adults' smaller WM span and the effects of extreme practice are required in the model.

Know CogTool. Cost = Small. Recent research has shown that CogTool can be taught in one class session and novice modelers building their first model produced predictions within 4% of an expert's model prediction, with a CV of only 7% (John, 2010).

Edit ACT-R. Cost = Large. In the final approach we studied, the practitioner must edit an ACT-R file to modify specific parameters to those established for younger and older adults (Jastrzembski, et al., 2010). This requires accessing CogTool's open source code, editing the code in the Eclipse programming environment, and knowing how and where to change the parameters. Although it is only four lines of Lisp code, the knowledge necessary to perform this procedure is, in our estimation, daunting, and would be required until CogTool could be enhanced to provide a GUI to switch between user groups.

There are two types of benefits possible in our analysis: the ability to correctly detect a difference between devices or user populations, and the numeric accuracy of its predictions. An approach is assigned a *large thumbs-up* when it correctly detects a statistically-significant difference present in the human data and, just as important for design,

Table 2. Assessment of costs and benefits of empirical data collection and seven modeling approaches.

Costs									Approach	Benefits					
										Detecting Differences		Match to Observed Times			
Collect Human Data	Program a running prototype	Literature Review	Measure for Fitts's Law	Build Storyboard	Know GOMS/MHP	Know KLM	Know CogTool	Edit ACT-R		Detect difference between devices	Detect difference between age groups	Dialing-Young	Dialing-Old	Texting-Young	Texting-Old
									Empirical: Train participants to skilled level, then collect data *(Human Data)*						
									GOMS + literature parameters for all operators *(GOMS-MHP)*						
									CogTool "Out of the box", naïve use *(CogTool-OotB)*						
									CogTool + KLM knowledge to delete Think operators *(CogTool+KLM)*						
									CogTool+KLM + 1.5:1 ratio (Hale & Myerson, 1995) for Think operators *(CogTool+KLM+RatioThink)*						
									CogTool+KLM+RatioThink + deleting all but the first Think operator to account for extreme practice dialing *(CogTool+KLM+RatioThink+ExtremePractice)*						
									CogTool+KLM+RatioThink +ExtremePractice + WM capacity of Older Adults causes more looking at task description in Dialing task *(CogTool+KLM+RatioThink+ExtremePractice+OlderWM)*						
									Best CogTool from Tables 1&2 + literature review parameters for all ACT-R operators *(CogTool(Best)+LitReview ACT-R parameters)*						

Key

- Months to acquire knowledge or do this work
- Weeks to acquire knowledge or do this work
- Days to acquire knowledge or do this work
- Detects difference when there is one and not when there isn't, or numerically within 5% of human data
- Does not (or cannot) detect difference when there is one
- Numerically within 5-10% of human data
- Numerically within 10-20% of human data
- Numerically greater than 20% of human data

does not claim a difference when there is no statistically-significant difference in the human data; a *large thumbs-down* is assigned when this is not the case. With respect to numeric accuracy, we assigned each prediction to one of four categories as shown in the key in Table 2.

Discussion of Costs & Benefits

The results of our assessments appear in Table 2. As mentioned before, collecting human data is considered the gold standard in UI design practice, but its cost is high, particularly for organizations with little staff or resources for experiment design, collection and analysis. Jastrzembski and Charness's GOMS-MHP modeling produced excellent predictions, but required eye-tracking and PhD-level understanding of the psychology literature and the Model Human Processor in order to attain those levels of predictive accuracy.

CogTool-OotB is less costly to learn and use, even for people with no psychology background. It correctly detected the difference between the devices when there was one in the data (for texting), but it was not designed to detect age-related performance differences, as it applies only to the performance of younger adults. Only by augmenting that tool with levels of knowledge of KLM and age-related literature, do models constructed within CogTool approach the level of accuracy useful for UI design if age is a factor. In fact, when only consideration of extreme practice is taken into account, the CogTool models produced fail to detect the

age-related differences in the dialing task. Only when the combination of extreme practice and WM capacity for older adults were incorporated, did the predictions fall into alignment with the empirical results. This requires substantial knowledge of the psychology literature that many practitioners would likely not possess.

Finally, the addition of specialized ACT-R parameters for younger and older adults in fact *increased* the average absolute percent error, demonstrating that utilization of substantially increased requirements of knowledge and skill (ACT-R, Eclipse & Lisp) does not always improve predictions sufficiently to warrant the increased effort.

Conclusions & Future Work

This research compares the *efficiency* and *effectiveness* of a variety of modeling approaches across tasks, designs, and user populations. There is no "right answer" for any particular development project, as each will vary in their need for accuracy, the current knowledge and skill of their team, and the value placed on acquiring modeling skill for future use. For example, if a design project must have predictions for all tasks within 5% of the "gold standard", the only approaches we examined achieving that criterion are empirical data collection[3] or GOMS-MHP modeling, with their associated high costs. However, if slightly less accurate predictions are acceptable, CogTool models augmented with some knowledge of KLM and psychology may be useful. Table 2 should be considered a guide when considering modeling, not a table of definitive recommendations.

Furthermore, advocates of using models in the development process always suggest that modeling can be used in conjunction with empirical testing, i.e., quick and easy CogTool modeling could be used as a means of weeding out detectibly poor designs from an assortment of design options in a tractable amount of time, so that empirical data collection may then be used to evaluate the few remaining candidates where accuracy is of high value. No one method need stand alone.

Several areas of future tool development are suggested by this investigation, pending, of course, repeatability of these results. First, if age-specific Think values detect age-related differences on other tasks on other devices, it would be a simple matter to put a radio button in the CogTool UI to allow analysts to select younger or older adults and attain appropriate predictions without editing scripts. Likewise, if future research showed that age-specific ACT-R parameters increased accuracy in the majority of cases, they also could be brought into play without analysts touching the underlying ACT-R Lisp code. Thus, it is beneficial to examine the costs and benefits of modeling approaches periodically, because such examinations may be used to improve model tool development, and allow us, as a field, to change the costs associated with the most useful approaches.

Acknowledgments

This research was supported in part by funds from IBM, NASA, NEC, PARC, and ONR N00014-03-1-0086. The views and conclusions in this paper are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of IBM, NASA, NEC, PARC, ONR, AFRL, or the U.S. Government.

References

Anderson, J. R., & Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Card, S. K., Moran, T.P. and Newell, A. (1983) *The Psychology of Human-Computer Interaction.* Lawrence Erlbaum Associates, Hillsdale, NJ.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology, xlvii,* 381–391.

Hale, S., & Myerson, J. (1995). Fifty years older, fifty percent slower? Meta-analytic regression models and semantic context effects. *Aging and Cognition, 2,* 132–145.

Jastrzembski, T. S. (2006). The Model *Human Processor and the Older Adult: Validation and Error Extension to GOMS in a Mobile Phone Task.* Unpublished doctoral dissertation. Psychology Department, Florida State University, Tallahassee, FL.

Jastrzembski, T. S., & Charness, N. (2007). The Model Human Processor and the older adult: validation in a mobile phone task. *Journal of Experimental Psychology: Applied*, 13, 224-248.

Jastrzembski, T. S., Myers, C., & Charness, N. (2010) A principled account of the older adult in ACT-R: Age-specific model human processor extensions in a mobile phone task. To appear in *Proceedings of the Human Factors and Ergonomics Society 54nd Annual Meeting*, (San Francisco, CA, September 27-October 1, 2010).

John, B. E. (1994) Toward a deeper comparison of methods: A reaction to Nielsen & Phillips and new data. In *Proceedings Companion of CHI, 1994* (Boston, MA, April 24-28, 1994) ACM, New York, NY. 285-286.

John, B. E., (2010) Reducing the Variability between Novice Modelers: Results of a Tool for Human Performance Modeling Produced through Human-Centered Design. *Proceedings of the 19th Annual Conference on Behavior Representation in Modeling and Simulation* (Charleston, SC, March 22-25, 2010).

John, B. E., Prevas, K., Salvucci, D. D., & Koedinger, K. (2004). Predictive human performance modeling made easy. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '04.* ACM, New York, NY, 455-462.

Salvucci, D. D. (2001). An integrated model of eye movements and visual encoding. *Cognitive Systems Research*, *1*, 201-220.