A comparison of the effects of nine activities within a self-directed learning environment on skill-grained learning*

Ari Bader-Natal, Thomas Lotze, and Daniel Furr**
{ari,thomas}@grockit.com

Grockit, Inc.
San Francisco, CA USA
http://grockit.com

Abstract. Self-directed learners value the ability to make decisions about their own learning experiences. Educational systems can accommodate these learners by providing a variety of different activities and study contexts among which learners may choose. When creating a software-based environment for these learners, system architects incorporate activities designed to be both effective and engaging. Once these activities are made available to students, researchers can evaluate these activities by analyzing observed usage and performance data by asking: Which of these activities are most engaging? Which are most effective? Answers to these questions enable a system designer to highlight and encourage those activities that are both effective and popular, to refine those that are either effective or popular, and to reconsider or remove those that are neither effective nor popular. In this paper, we discuss Grockit – a web-based environment offering self-directed learners a wide variety of activities – and use a mixed-effects logistic regression model to model the effectiveness of nine of these supplemental interventions on skill-grained learning.

Keywords: self-directed learning, learner control, skill-grained evaluation

Educational software designed for the classroom is often only effective in the classroom, simply because students use this software only when they are required to do so. For non-compulsory learning software to be effective, being *engaging* is a necessary (but not sufficient) precondition. As the notion of engagement is subjective, one approach to building a system that many learners find engaging is to support a variety of modes and activities and allow each learner to find their preferred niche. Grockit, a web-based learning environment designed for individual students who share a common domain-specific learning goal, takes this approach by incorporating two dimensions of variety/flexibility: context and control. At any point in time, learners can choose from among three contexts of study: individual practice, peer group study, and instructor-led lessons. The learner can also choose the amount of control that he or she wishes to exert

* *To be published in the Proceedings of the 15th International Conference on Artificial Intelligence in Education – June 2011. The original publication will be available at springerlink.com*
** Work done at Grockit. Present address: School of Education, University of California Berkeley.

to define the learning experience [5]: with learner-driven control offered through HCI affordances and system-driven control provided via AI approaches (such as an adaptive problem selection algorithm based on an Item Response Theory model [2]). Grockit pursues an engaging learning experience by means of game design and social interactions both addressed in prior work [1,2], and internal surveys continues to indicate that the vast majority of participants find Grockit's learning environment to be engaging. The variety introduced to increase engagement does, however, add complexity to the attribution of the effectiveness. In this work, we summarize nine of the interventions incorporated into the Grockit system, and evaluate the extent to which each of these is an effective addition to the learning platform.

1 Interventions within Grockit

Grockit provides a place for students to master new skills and exercise what they learn through three contexts for problem solving: (*a.*) *individual study*, which uses an Item Response Theory model to provide that student with appropriate challenges for learning [7], (*b.*) *small group study*, which leverages collaborative learning dynamics to provide students with a social learning network that can help motivate and assist them [2], (*c.*) *instructor-led classes*, which draw on an expert's domain knowledge and experience to provide a guided and structured path for larger groups of learners.

The core activity within all three learning contexts involves answering multiple-choice and numeric response problems in some well-defined learning domain (e.g. an Algebra I course, the GMAT exam, a Grade 8 English Language Arts course), and then reviewing expert-authored solutions and explanations for each of these problems. In the small group and instructor-led settings, all participants see the same question at the same time, enabling group discussion around problems and solutions. In addition to the core problem-solving activity, learners have access to a number of supplemental learning activities motivated by work in prior systems, and introduced to the Grockit environment with the goal of contributing to the learning gains of participating students. In this study, we focus on nine of these activities:

explanation_read: ***Read an explanation of the question immediately after answering it.***
For each question in Grockit's item database, the author of the question prepared an explanation of the solution and, for multiple-choice questions, explanations or comments about each answer choice. After testing a variety of different contexts within the application for incorporating these explanations, we chose to make these explanations available only during individual study sessions.[1] These in-game explanations were introduced in order to provide students with a cohesive expert-authored solution – visible after the student answers the question and sees which answer choice is correct. Viewing these explanations is presented as an optional activity: a "view an explanation for this question" link is displayed above each question, and the student must click the link to reveal the explanation. When given

[1] We found that the time required to benefit from explanations varied widely among students, and was therefore a better fit for self-paced review rather than for the real-time group study. For a more details on decisions around interaction synchronicity, see Bader-Natal [2].

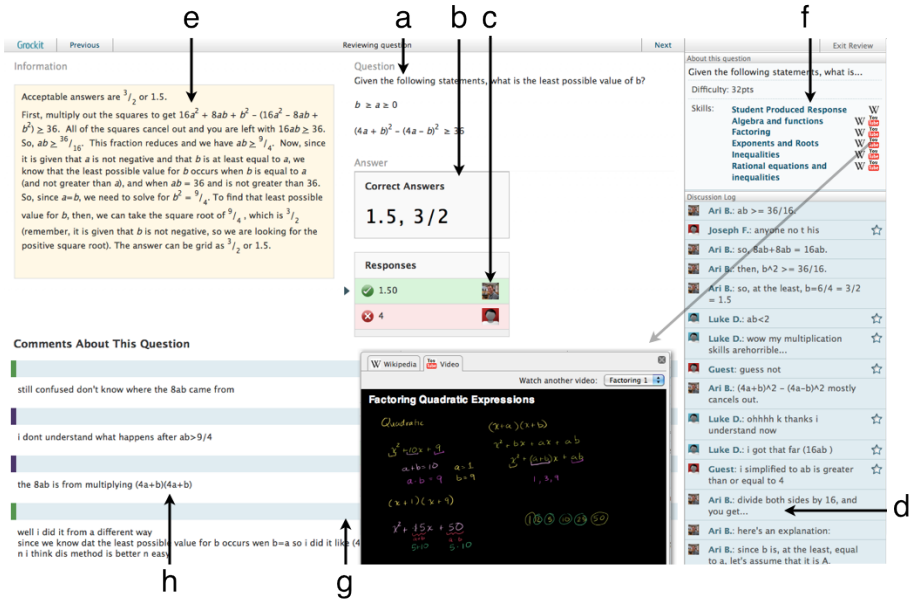

Fig. 1. A Review includes several components, including: (a.) the original question and answer choices, (b.) the correct answer, (c.) each of the answers submitted by the students in the session, (d.) the discussion transcript from that session, (e.) expert explanations of the question and each answer choice, (f.) metadata about the problem including difficulty level and list of associated skills, (g.) access to videos and blog posts discussing each these concepts, and (h.) an asynchronous discussion thread among all students who have reviewed that question.

the opportunity to view an explanation after answering a question and seeing the correct response, 48% of students *who answered incorrectly* and 18% of students *who answered correctly* chose to view the explanation.[2] If we find that viewing an explanation immediately following an incorrect response is an effective intervention, we might start displaying these explanations to all students in individual study sessions following an incorrect response, without them needing to request it.

reviewed: ***Reviewed a question from a study session.*** As mentioned above, a post-hoc review of study sessions is available to students, in which no per-question time constraints are necessary, since the solo nature of the activity means that synchronizing pace with other students is not necessary. Over the past few years, these reviews have grown to include an assortment of resources for the student to draw on, illustrated in Fig. 1. Of these components, three involve actions that are addressed separately in below. Beyond the practical logistics of time necessary to engage in these activities, the reviews serve to distribute skill practice over time (rather than to compress all practice into the initial practice session), an approach that seems to be supported by data on the spacing-effect [4].

[2] Based on item responses from 10/1/2010 - 12/31/2010 from people studying for the GMAT.

watched_video: ***Watched an instructional video about the skill.*** Watching an instructor explain a concept and work example problems is one of the primary modes of face-to-face instruction, and a common component of online learning environments. For each question in Grockit, the set of skills required to solve the problem are listed next to the question in the reviews, along with other question metadata. For each concept listed, the student can choose to watch short videos explaining the concept, embedded from public video sharing sites such as YouTube, with videos selected by content authors based on relevance and quality.

viewed_textbook: ***Read an expert-authored description of the underlying skill.*** Similar to the videos described above, each of the skills associated with the question are correlated with written explanations of those skills (but not of the specific question.) These skill-explanations were originally prepared as a series of blog posts.

question_comment: ***Appended a message to a question during a review session.*** Within group study sessions, students are able to discuss questions as they work on them, in real-time. In reviews, students can read their past discussions, but cannot get real-time answers to their questions. We introduced an asynchronous discussion thread for each question to allow students to discuss with others who have seen the question, even if at a different time.[3]

discussed: ***Typed a message after answering a question in group study.*** In group study sessions, a chat box is displayed next to the question that the students are attempting to solve. While discussions about a question may include no participants with knowledge or expertise, studies by Smith et al. suggest that small group discussions following a question can be beneficial even when none of the participants had correctly answered the question initially [9].

questioned: ***Asking questions, in game discussions.*** We use the presence of a question-mark in the discussion as a low-fidelity indicator of a request for help. The discussions that transpire are generally a combination of on-task peer-assistance and off-task conversations.[4] While this signal is clearly quite noisy and the outcome not definitive, we prefer to include this rather than nothing at all.

tutor_led: ***Participating in an instructor-led lesson.*** The three modes of study in Grockit – instructor-led sessions, group study, and individual practice – have parallels in Dron and Anderson's distinction between groups, networks, and collectives [6]. Of the three, the instructor-led sessions most closely resemble a traditional classroom: The instructor schedules a session and some number of students attend. The instructor can incorporate slides, whiteboards, and shared text editors into the session, and while practice problems are done, the primary focus is on instruction. We include this to determine if these structured lessons are of measurable value.

with_tutor: ***Participated in a group study session in which a tutor was present.*** The instructors who lead lessons also frequently join ongoing peer-group study sessions. Instructors generally participate and encourage discussion in these sessions, but

[3] Comments in the asynchronous discussion threads are often more thoughtfully prepared and are less context-dependent than the more casual and interactive discussion messages in synchronous group games. We include comment authoring in this analysis because we wish to see if taking the time to participate in this forum has an effect on skill learning outcomes.

[4] The casual nature of off-task discussions serves to reduce the stress associated with studying, so we do not discourage these discussions.

they do not lead them in the formal way that they lead lessons. We include this to determine if this more casual participation in group study is beneficial.

2 Methods

We formulate the effect of available interventions on skill-grained learning as follows: *After a student incorrectly answers a question involving some skill, engages in an intervention involving that skill, and then attempts a subsequent question involving that same skill, what effect does that intervention have on second response accuracy?*

For this analysis, we consider data collected during a two-month period (October 1 - December 1, 2010). We consider two types of data: item responses and item interventions, and exclude item responses and interventions from all user accounts belonging to teachers, tutors, system administrators, and anonymous guests. Each item in the Grockit database is associated with one or more skill tags describing the concepts required to solve the problem. Both responses and interventions can be associated with skills, and here we use skills as the granularity for analysis.

Each student's performance on a specific skill can be organized into a timeline of *item responses* on that skill, which may be correct or incorrect, and *item interventions* on that same skill, which are intended to improve the student's performance. When an item intervention is followed by an item response, we have the opportunity to see how the intervention impacted the user's performance on that skill.

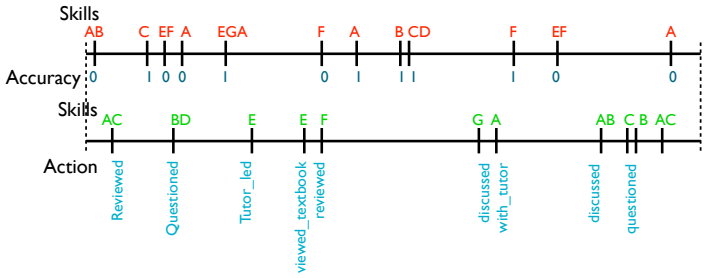

Fig. 2. An dual-timeline example for a student. The upper line contains skill-tagged item responses and the lower line contains skill-tagged interventions.

For *item responses*, we use $r_n^{(s,k)}$ and $t_n^{(s,k)}$ to denote the response accuracy and timestamp, respectively, for the n^{th} response by to skill k by student s (where $r_n^{(s,k)} \in \{0, 1\}$). For *item interventions*, we use $T_{(j,u)}^{(s,k)}$ to denote the time at which student s participated in their u^{th} intervention of type j (where $j = 1..9$ for the nine intervention types) on skill k. We may then determine, for each user response, which interventions the student participated in before that response. If the student participated in a certain

person	skill	first_response_time	second_response_time	second_difficulty	reviewed_explanation	discussed_read	questioned_tutor	watched_video	viewed_textbook	question_comment	tutor_led_with_tutor	first_accuracy	second_accuracy
u_1	s_a	2010-10-22 18:19:20	2010-10-22 18:21:38	-0.68	0	0	1	1	0	0	0	1	1	0	0
u_2	s_a	2010-10-22 18:21:38	2010-10-22 18:23:12	-1.09	0	1	0	0	0	0	0	0	0	0	0
u_2	s_b	2010-10-22 18:23:12	2010-10-22 18:25:17	-0.98	1	0	0	0	1	0	0	0	0	0	1

Table 2. Example rows from the combined dataset used for analysis.

type of intervention for the skill between two subsequent item responses on that skill, we record this as a 1. If there was no such intervention, we record it as a 0:

$$i_{(j,n)}^{(s,k)} = \begin{cases} 1 & \textit{if } \exists T_{(j,u)}^{(s,k)} : t_{n-1}^{(s,k)} < T_{(j,u)}^{(s,k)} < t_n^{(s,k)} \\ 0 & \textit{otherwise} \end{cases}$$

We only measure interventions after the user's previous response on this skill, as we consider these to be the strongest indicators of an improvement due to the intervention. Table 2 illustrates a few example rows from the resulting data set.

We use a mixed-effects regression to model the second response accuracy. As we are looking for evidence of learning, we consider only those records for which the previous response was incorrect (i.e. responses $r_n^{(s,k)}$ where $r_{n-1}^{(s,k)} = 0$); we view a correct response to the second item to be an indicator of learning. Among these records, we treat the nine interventions as fixed effects. We also include the difficulty of the second item (d_{q_2}) as a fixed effect, as we expect the second question's difficulty to (negatively) impact the person's response accuracy on that question. We treat the variance between students as a random effect in this model, $\alpha_s \sim N(0, \psi^2)$:

$$\textit{logit}\left\{P\left(r_n^{(s,k)} = 1\right)\right\} = \beta_0 + \beta_d d_{q_2} + \beta_1 i_{(1,n)}^{(s,k)} + \cdots + \beta_9 i_{(9,n)}^{(s,k)} + \alpha_s$$

We note a few weaknesses in this approach. This adjacent-pair analysis provides insight into short-term effects of individual interventions. Learners generally respond to a sequence of items for each skill, and these cumulative effects are not captured in this model, resulting in a weak signal of learning. Additionally, most questions are tagged with more than one skill, and an incorrect response cannot be attributed to a single skill. Finally, we recognize that when a student engages in a particular intervention, they are both benefitting from it and signaling that they believe that they will benefit from it. The benefits may therefore be affected by the biased sample. Overall, since this is not a randomized controlled experiment and learners can self-select their interventions, we can attribute correlation but not causation.

3 Results

Table 3 reports the coefficients estimated from the mixed-effects logistic regression model, obtained using the *lme4* package for the R statistical environment [3,8]. The difficulty of the second item (*second_difficulty*) has a statistically significant effect on the second response accuracy, as was expected. The more difficult the item, the lower

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.17	0.01	-14.10	0.00	*
reviewed	0.04	0.02	2.51	0.01	*
explanation_read	0.04	0.01	2.77	0.01	*
discussed	0.05	0.01	5.18	0.00	*
questioned	-0.02	0.01	-1.55	0.12	
watched_video	-0.82	0.52	-1.57	0.12	
viewed_textbook	-0.36	0.17	-2.12	0.03	*
question_comment	0.14	0.10	1.36	0.17	
tutor_led	0.22	0.11	1.97	0.05	*
with_tutor	0.10	0.02	5.85	0.00	*
second_difficulty	-0.68	0.00	-223.07	0.00	*

Table 3. Model coefficients from the Generalized Linear Mixed Model. The student is treated as a random effect (variance: 0.98). Stars indicate significance at the $\alpha = 0.05$ level.

the expected response accuracy.[5] Of the nine interventions examined in all, five had a statistically significant positive effect (at the $\alpha = 0.05$ level), one had a statistically significant negative effect, and three had no statistically significant effect. The interventions with the highest coefficients involved the expert instructors, with a 0.22 increase in the log odds of learning in instructor-led lessons (*tutor_led*) and a 0.10 increase in group games in which an instructor participates (*with_tutor*). Reviewing items (*reviewed*) also has a significant effect, with an estimated coefficient of 0.04. This coefficient represents the increase in the log odds of success (i.e. a correct to the following attempt at a question of the same skill) for this student, if this student reviewed a question involving that skill prior to the second response. Participating in group game discussions was estimated to increase the log odds (logits) of learning by 0.05. Choosing to view an explanation (*explanation_read*) after answering a question in a individual practice session increased the outcome by 0.04 logits, and reviewing a question (*reviewed*) increased the outcome by 0.04 logits. Neither watching a video (*Watched_video*) nor leaving a comment (*question_comment*) had a statistically significant effect (beyond that of reviewing itself). Unexpectedly, viewing the "textbook" concept explanations *viewed_textbook* had a statistically significant negative effect. Asking a question within a group discussion (*questioned*) was not found to have a significant effect.

4 Discussion

This analysis represents our first effort to quantify and evaluate the learning outcomes associated with individual activities available within Grockit. The variety of available tools in the learning environment adds both richness to the experience and complexity to the attribution of learning gains. The results here suggest which of the interventions analyzed were most effective and, coupled with an understanding of how engaging each activity is, these results can inform decisions around which interventions to highlight, which to refine, and which to reconsider. Given the positive effect observed among students who choose to view an question explanation in an solo practice after an incorrect response, we might automatically show these, rather than requiring students to opt-in

[5] Item difficulty is estimated based on a three-parameter item response theory model.

each time. As for activities displaying no statistical significance, we are now discussing modifications expected to make them more effective.[6]

In another study currently in progress, we use a randomized controlled design to evaluate overall learning gains from participation, without attribution to interventions by type. Where the current analysis only examines select interventions, the A/B design is more comprehensive, incorporating the core problem solving practice and intermittent assessments that are not captured in the present analysis. To understand the effect of a complex learning environment, we believe that both approaches are valuable.

While students are generally required to use (and continue using) educational software introduced in a formal learning setting, no such obligation governs use of educational software by self-directed learners. In order to be capable of impacting learning for these students, a system must be *both* sufficiently engaging for students to continue using it *and* effective. Different people find different learning contexts and activities engaging, so Grockit chose to introduce and leverage *variety* – learner choice and control over how, when, and with whom one learns – to address the assorted needs and preferences of self-directed learners. A large (and growing) number of students do, in fact, find the platform engaging, as evidenced by internal survey data and observed time-on-task. In this analysis, we find that several of the learning interventions incorporated into the platform are effective, with participation associated with skill-grained learning. By building a platform that is engaging and incorporates effective interventions, Grockit has created an environment uniquely-suited to the needs of the self-directed learner.

References

1. A. Bader-Natal. Incorporating game mechanics into a network of online study groups. In Scotty D. Craig and Darina Dicheva, editors, *Supplementary Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED-2009)*, volume 3, Intelligent Educational Games workshop, pages 109–112, Brighton, UK, July 2009. IOS Press.
2. A. Bader-Natal. Interaction synchronicity in web-based collaborative learning systems. In Theo Bastiaens, Jon Dron, and Cindy Xin, editors, *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2009*, pages 1121–1129, Vancouver, Canada, October 2009. AACE.
3. D. Bates and M. Maechler. *lme4: Linear mixed-effects models using S4 classes*, 2010.
4. J.J. Donovan and D.J. Radosevich. A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, 84:795–805, 1999.
5. J. Dron. *Control and constraint in e-learning: Choosing when to choose* . Information Science Publishing, 2007.
6. J. Dron and T. Anderson. Collectives, networks and groups in social software for e-Learning. In *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education Quebec. Retrieved Feb*, volume 16, page 2008, 2007.
7. Frederic M. Lord. *Applications of Item Response Theory to practical testing problems*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1980.
8. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010.

[6] We suspect that the non-significant effect of asking a question during discussion may be due to imperfect identification, which includes both on-task and off-task (e.g. social) questions. This could be clarified if questions were coded as such and tested separately.

9. M. K. Smith, W. B. Wood, W. K. Adams, C. Wieman, J. K. Knight, N. Guild, and T. T. Su. Why peer discussion improves student performance on in-class concept questions. *Science*, 323(5910):122–124, January 2009.