Personalization, Non-Cognitive Factors, and Grain-Size for Measurement and Analysis in Intelligent Tutoring Systems: Implications for GIFT

Stephen E. Fancsali[1], Steven Ritter[1], John C. Stamper[2], Susan Berman[1]
[1]Carnegie Learning, Inc., Pittsburgh, PA, USA
[2]Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA

Introduction

One major goal of our Advanced Distributed Learning (ADL)-funded "Hyper-Personalized Intelligent Tutor" (HPIT) project is to develop an architecture that allows for personalization of educational software based on non-cognitive factors including student preferences, motivation, affect, and other features (Fancsali, Ritter, Stamper & Nixon, 2013). We are particularly interested in using such information to improve instruction within intelligent tutoring systems (ITSs) like Carnegie Learning's Cognitive Tutor® (CT) (Ritter, Anderson, Koedinger & Corbett, 2007) and in educational games like the mathematics "fluency challenges" developed by the Math Fluency Data Collaborative project (http://fluencychallenge.com). To better understand architectural features conducive to such "hyper-personalization," we have built on learning science researchers' recent work focusing on particular aspects of personalization and non-cognitive factors in both the CT and a middle school mathematics ITS based on the CT called MATHia.™

We review results of two recent observational studies with respect to the U.S. Army Research Laboratory's (ARL) Generalized Intelligent Framework for Tutoring (GIFT). In the first, we provide a causal model of measures of goal orientation and self-efficacy (at multiple grain-sizes), behavioral measures in the CT, and learning outcomes (Fancsali, Bernacki, Nokes-Malach, Ritter & Yudelson, in press). These findings are consistent with research (Bernacki, Aleven & Nokes-Malach, 2012) showing that "domain-level" (for the mathematics domain) and "unit-level" (for a CT unit of instruction) measures of goal orientation and self-efficacy provide different information about learners and their outcomes. Such results suggest that these measures are contextually bound and thus need to be assessed relatively frequently.

In the second study, we consider aggregate associations of in-tutor measures of performance with "honoring" student preferences for interest areas (e.g., "sports & fitness") around which mathematics word problems have been authored. We adopted a different approach from past research by considering aggregate associations between honoring preferences and learning outcomes, in order to assess whether, at the timeframe of a school year, we see influences of honoring student preferences. We found little if any association between honoring student preferences and learning outcomes (Fancsali & Ritter, 2014; Ritter, Sinatra & Fancsali, in press). Past experimental studies (Walkington 2013) have found effects at finer levels of granularity (e.g., at the level of problems and skills). We speculate that the failure to find such effects in our study may reflect the fact that, in the context of the full-year course, the opportunities to honor student preferences are limited. This finding suggests that taking advantage of preference honoring in an extended course may require more frequent or overt signaling about attention to student interests.

With respect to the GIFT architecture, we discuss a response to recent work (Otieno, Schwonke, Salden & Renkl, 2013), due to Fancsali et al. (submitted), suggesting that using relatively simple log traces from CT as "online measures" of goal orientation and how the aforementioned goal orientation and self-efficacy study informs this response. Generally, we provide new recommendations for GIFT and a better, data-driven foundation for recommendations concerning GIFT made by Fancsali, Ritter, Stamper & Nixon (2013). We suggest future research in the areas of non-cognitive factors and personalization for ITSs and development areas for HPIT and GIFT. Specifically, we emphasize that the grain-size of measurement and analysis required for adaptation within a particular ITS or instructional system may be different than that at which meaningful inferences (and corresponding adaptive instruction or tutoring decisions) might be made across *different* ITSs or instructional systems.

Preliminaries

We begin by describing Carnegie Learning's CT, the ITS which provides the impetus for our thinking about hyper-personalization of instructional systems as well as the target platform for the research projects we explicate. We then describe the HPIT project as well as two non-cognitive factors and a facet of personalization along which such a hyper-personalized instructional system might tailor instruction. As we proceed, we note an important aspect of adaptation or personalization in any such system: the level of granularity or grain size at which adaptation or personalization occurs.

Cognitive Tutor (CT)

Carnegie Learning's CT (cf. Figure 1 screenshot) is an ITS for mathematics that has been demonstrated effective by one of the largest experimental trials of its kind (Pane, Griffin, McCaffrey & Karam, 2013). CT provides adaptive instruction based on cognitive factors, specifically by tracking learners' mastery of fine-grained knowledge components (KCs) (i.e., skills) using a probabilistic framework called Bayesian Knowledge Tracing (BKT) (Corbett & Anderson, 1995).

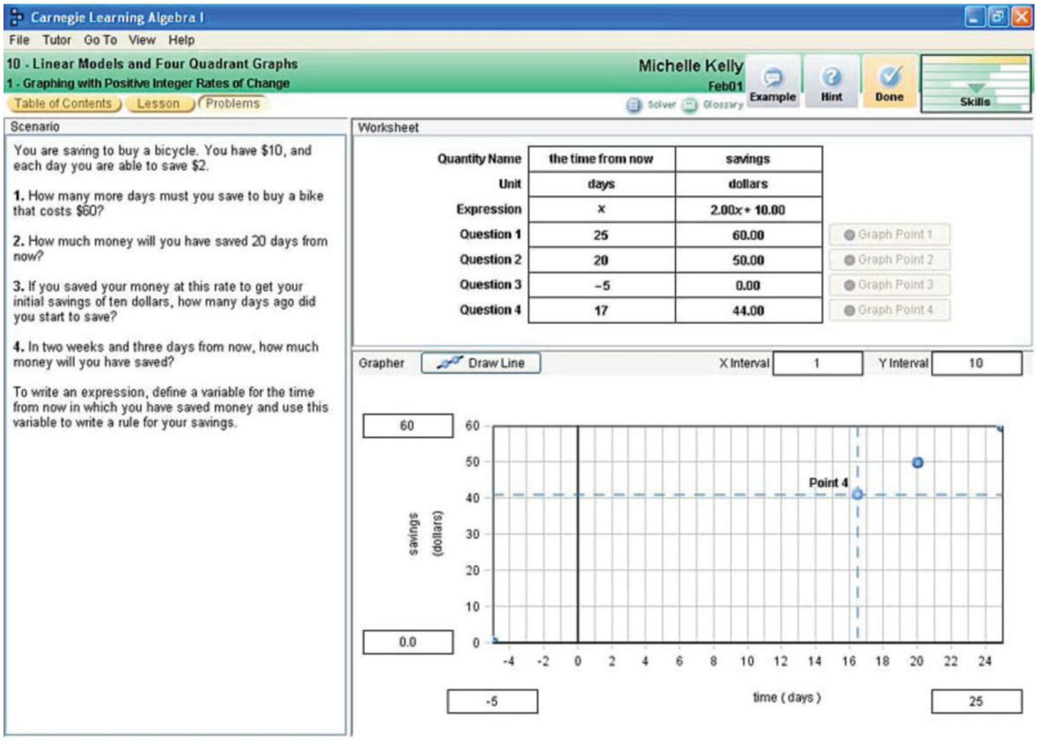

Figure 21. Problem solving in Carnegie Learning's CT

Mathematics curricula are divided into topical sections, each of which is associated with a set of KCs that a student is required to master before progressing on to another section of instruction. Each multi-part problem in a section is associated with a subset of its section's KCs. Students are provided immediate feedback about the correctness of attempts to complete steps in each problem; sometimes CT provides just in time, context-sensitive feedback when students make errors that reflect known misconceptions. At any step of problem solving, students can also request context-sensitive hints. While using ITSs to adapt mathematics instruction based on cognitive factors like skill mastery is relatively well-understand (but still a topic of on-going research), the use of non-cognitive factors to personalize and adapt instruction in ITSs presents a wide variety of unanswered and open questions.

Hyper-Personalized Intelligent Tutors (HPIT)

Our HPIT project (Fancsali, Ritter, Stamper & Nixon, 2013) aims to develop a plug-in based architecture and infrastructure to support personalized and adaptive learning, based on both cognitive and non-cognitive factors, in a variety of computer-based instructional settings, including, but not limited to, ITSs like CT and educational games. To truly "hyper-personalize" ITSs and other instructional systems, it is necessary to better understand how non-cognitive factors can be harnessed to enhance learning in such systems, and the HPIT project will provide a framework within which such research can take place, for example, by allowing developers to integrate plug-ins to drive personalized instructions based on particular non-cognitive factors in which they are interested.

To facilitate the development of such a software framework, a better understanding of particular methodological and measurement issues that will arise in such personalization efforts is also required. This understanding will not only improve the development of HPIT but will also help to drive integration and interoperability (as well as illustrate possible contrasts) between HPIT and projects like GIFT. Findings relevant to HPIT may also drive new recommendations for future development of GIFT. Specifically, we briefly summarize recent work on goal orientation and self-efficacy as they relate to performance in the CT as well as "honoring" student interest area preferences in an ITS for middle school mathematics that is based on the CT. From methodological concerns and findings about specific non-cognitive constructs, we hope to generalize insights to provide recommendations for GIFT. We focus on issues of measurement and analysis grain-size and discuss implications for adaptation and tutoring within an instructional system as well as open problems about making transferable inferences across instructional systems, both of which are important issues relevant to GIFT.

Grain size and non-cognitive factors

We now turn our focus to analyses of process data and learning in the CT environment relative to two non-cognitive factors: goal orientation and self-efficacy[1]. We begin by briefly introducing these constructs before considering issues about their measurement and grain-size before considering relationships of these factors with learning outcomes in the CT in the following section.

Goal orientation and self-efficacy

Students' achievement goals for learning are a key component of their motivation for learning. Dweck (1986) identifies performance and achievement as two particular goal orientations. Students with a performance goal orientation tend to define their goals relative to the performance of their peers or some other group (i.e., to perform better than their peers) while those with a mastery goal orientation take understanding a particular task or aspect of a domain as their goal. Elliot & McGregor (2001) later distinguish between two "valences," approach (success) and avoidance (of failure), for mastery and performance goal orientation. For example, performance avoidance as a goal orientation is attributed to students who seek not to perform worse than their peers (i.e., to demonstrate that they are no worse than their peers). A performance approach orientation is attributable to students who seek to demonstrate their competence by out-performing their peer group. Students with mastery approach goals tend to seek to develop competency for a learning task by achieving at levels above their past performance or by setting other criteria for judging increased or developing mastery (Ames, 1992; Elliot, 1999).

Learner self-efficacy, or beliefs a learner has about his or her abilities in performing in a particular domain or task may also play a role for setting goals, self-regulation, and other factors important for learning (Bandura, 1994). Learners with high self-efficacy, for example, may be more willing to set more difficult or less easily attainable goals that require working through difficult tasks. Self-efficacy is also related to persistence and better performance on learning tasks (Bandura, 1997).

[1] We here provide a brief summary of the explication of these non-cognitive factors found in Fancsali et al. (in press).

Self-report questionnaires, especially those given before or after a particular learning task, are frequently used measure goal orientation and self-efficacy. However, several authors note that factors like goal orientation can change as learners progress in a learning experience or over longer periods of instructional time (e.g., an academic year) (Richardson, 2004; Fryer & Elliot, 2007; Muis & Edwards, 2009). A propos, we consider work that takes such differences in measurement grain size into account. Later, we consider the possibility of "online" measures (e.g., log traces from ITSs) of such non-cognitive factors.

Grain size

Given the possibly dynamic nature of non-cognitive factors, including goal orientation, Bernacki, Nokes-Malach & Aleven (2013) raise concerns for the granularity at which factors like goal orientation are measured in conjunction with use of ITSs like CT. Beyond reliable changes in learners' goal orientation from unit to unit of CT instruction found by Bernacki, Nokes-Malach & Aleven (2012), the same authors also found that domain-level measures of goal orientation (e.g., for the mathematics domain) versus unit-level measures of goal orientation (e.g., for CT units of instruction) have different levels of association with various behaviors, which provides for the possibility that different measurement grain sizes are providing different information in possibly important ways (Bernacki, Nokes-Malach & Aleven, 2013). We report recent work due to Fancsali et al. (submitted) that provides a linear structural equation model relating data collected on self-efficacy at the unit- and domain-levels and learners' goal orientations. This work suggests that the move from self-report questionnaires to relatively simple log traces to measure achievement goals, including counts of hints and glossary access in the CT, is likely more nuanced than suggested by recent work of Otieno, Schwonke, Salden & Renkl (2013). Before describing this work, we briefly consider one facet of personalization based on non-cognitive factors that provides the basis of the second study we report.

Student preferences, interest area personalization, and grain size

Computer-based ITSs like CT can provide functionality that allows for context personalization, roughly aligning instructional content to interests of an individual learner (Anand & Ross, 1987; Cordova & Lepper, 1996; Walkington, 2013). For instance, Carnegie Learning's MATHia product, an ITS for middle school mathematics based on CT, provides for mathematics word problems to be tailored to specific interest areas of learners for domains outside of the classroom (e.g., "sports & fitness") as well as based on student-entered names of their favorite classmates or friends. Personalizing instruction based on both of these non-cognitive factors (i.e., problem content and names) has been demonstrated effective by previous experiments (Anand & Ross, 1987; Cordova & Lepper, 1996; Ku & Sullivan, 2002). With respect to CT, research using an experimental version of this ITS finds that personalizing mathematics word problems based on student interest areas improves learning outcomes (i.e., performance) at the problem-level, especially on problems with a relatively high reading level that involve difficult KCs (Walkington, 2013). One posited mechanism at the level of KCs is that personalization improves students' ability to symbolically represent word problems at a higher reading level. However, as we describe later, at a higher level of granularity, when we have considered the aggregate association of "honoring" student preferences with learning outcomes, our results are less clear (Fancsali & Ritter, 2014; Ritter, Sinatra & Fancsali, in press).

Goal Orientation, Self-Efficacy, Behavior, and Learning

To better understand, among other things, the proposal of Otieno, Schwonke, Salden & Renkl (2013) that "online" measures from ITS log data, simple counts of hints requested and glossary use, be used to assess achievement goals or goal orientation (performance approach and mastery approach goals, respectively), Fancsali et al. (in press) learned qualitative causal structure, using data-driven procedures based on conditional independence tests and background knowledge (Spirtes, Glymour & Scheines, 2000), to specify a linear structural equation model (estimated and illustrated as Figure 2[2]) over features representing goal orientation, self-efficacy, math grades (pre/post), and process data from CT.

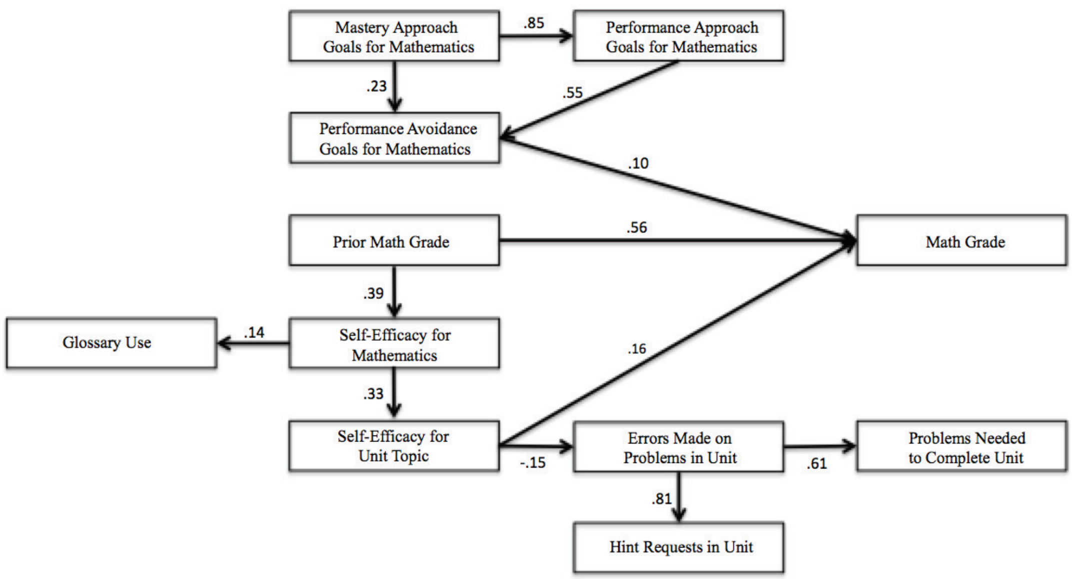

Figure 22. Estimated linear structural equation model reported by Fancsali, Bernacki, Nokes-Malach, Ritter & Yudelson (submitted), including standardized path coefficients.

Data were collected for 273 pre-algebra, algebra, and geometry students using the CT during an experiment run by Matthew Bernacki, Timothy Nokes-Malach, and Vincent Aleven in a school district in western Pennsylvania to investigate issues of grain-size and measurement of performance goals and self-efficacy. Self-report survey items[3] for achievement goals were original items from subscales for performance approach, performance avoidance, and mastery approach on the Achievement Goals Questionnaire – Revised (Elliot & Murayama, 2008). The study designers crafted self-efficacy survey items according to guidance provided by Bandura (2006). Domain-level survey items (i.e., with respect to learner's feeling about mathematics) were administered within the CT software at the beginning of the school year, while unit-level items (with language adapted to refer to the particular CT unit of instruction students had just completed) were presented after every other unit of instruction (alternating with unit-level self-efficacy items).

[2] Their reported model fits the data well according to the appropriate chi-square statistical test for such a model comparing its implied covariance matrix to the observed covariance matrix ($\chi^2(43) = 49.19$, $p = 0.239$) (Bollen, 1989).

[3] For complete details of survey items, see Fancsali et al. (in press).

While the proposal of Otieno, Schwonke, Salden & Renkl would have us count *Glossary Use* as a measure of mastery approach, we find that it is weakly linked only to domain-level learner self-efficacy (i.e., for mathematics); the model of Figure 2 implies that conditional on *Self-Efficacy for Mathematics*, *Glossary Use* is independent of all other variables included in the model. Similarly, *Hint Requests in Unit*, a proposed measure of performance approach goal orientation, is strongly correlated with *Errors Made on Problems in Unit*, which in turn is only weakly linked to *Self-Efficacy for Unit Topic* (i.e., unit-level self-efficacy). The model is inconsistent with any strong link between hint use and achievement goals.

The model posits that three variables are directly linked to *Math Grade*: *Prior Math Grade*, *Self-Efficacy for Unit Topic*, and *Performance Avoidance Goals for Mathematics*. Notably, this set of variables includes both a unit-level and domain-level variable, and we see that these different levels of granularity do provide different information about behavior in CT. Direct links in this model, and especially the weak correlation of *Performance Avoidance Goals for Mathematics* with *Math Grade* should be interpreted cautiously, as the algorithm used to infer the structure of the model in Figure 2 assumes there are no unmeasured common causes of measured variables, and this assumption is unlikely to obtain in this application domain (i.e., the adage to not confuse correlation and causation is operative). Nevertheless, the model is useful for providing basic structural relationships from correlational data and providing information about what relationships *are not consistent* with observed data. Further, the pattern of high correlations among goal orientation variables is consistent with previous literature positing a "goal complex" (Barron & Harackiewicz, 2001; Senko, Hulleman & Harackiewicz, 2011); one explanation of such a "complex" would include one or more latent variables representing some more complex factor(s). We return to the problem of developing online measures for important non-cognitive factors like goal orientation and self-efficacy as well as issues of grain size and implications for GIFT in the discussion section.

"Honoring" Student Preferences and Learning

While previous CT experiments had demonstrated improved problem-level learning outcomes from personalizing mathematics word problems based on student interest areas, less clear was whether "honoring" interest area preferences, in the aggregate, is associated with better performance in an ITS like CT. To begin considering this question, we analyzed observational data from Carnegie Learning's MATHia ITS, which is based on CT (Fancsali & Ritter, 2014).[4] While MATHia allows students to (optionally) rate four interest areas with one to five "stars," (cf. Figure 3) we found, in a sample of 104,197 learners, 62,168 learners (59.7%) set interest areas, but 16,003 learners (25.7% of those who set interest areas) provided the same rating to all interest areas.

Consequently, we employ a notion of strong student preferences according to which a student has strong preferences if (and only if) they rate at least one interest area with five stars and at least one area with one star (or leave at least one area un-rated). From a smaller sample of 1,230 learners (from eight randomly selected schools that completed at least five instructional sections that contain preference-tailored problems and upon whom our detailed analysis focused), we found that 518 students had strong preferences

[4] Results reported in this section are described in detail by Fancsali & Ritter (2014).

according to this definition. We then calculated the proportion of problems presented to each of these students (in sections that contained preference-tailored word problems) that corresponded to their highly rated interest area as a measure of the extent to which their preferences were "honored."

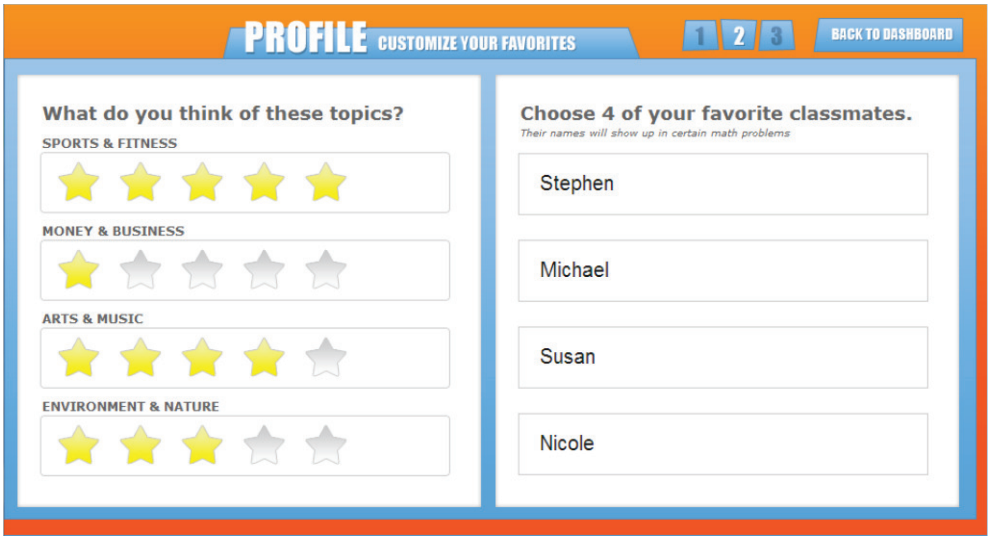

Figure 23. MATHia's preference dashboard

We found that the probabilistic provision of preference-tailored problems led to a relatively restricted range of proportions of preference-honoring per student, with the majority of students seeing somewhere between 10%–40% of problems in preference-tailoring sections that corresponded to their interests. Given that MATHia does not explicitly signal to students that problems have been provisioned according to their preferences, there is some question as to whether students generally realize that the software is adapting to their interests. Further, we considered correlations of proportion of preference-honoring problems to four process variables found in past work to be predictive of standardized test scores (Ritter, Joshi, Fancsali & Nixon, 2013): assistance per problem (i.e., number of hints requested & errors committed per problem), number of MATHia sections encountered, number of sections completed per hour (i.e., roughly efficiency), and amount of time logged into MATHia.

While three of four correlations with proportion of preference-honoring problems were significant (at the $\alpha = 0.05$ level), no correlation had an absolute value greater than 0.14, and it is unclear that there is any substantive use to be made of such weak, aggregate correlations. Further, the directions of these correlations imply a slight negative relationship between preference honoring and MATHia performance, but it is unclear whether this also has any substantive significance. Interestingly, we found that there were statistically significant differences (at the $\alpha = 0.05$ level) between strong and weak preference students, as strong preference students worked through MATHia material more efficiently (i.e., greater sections completed per hour) (Welch two-sided, two sample $p = 0.01$, Cohen's $d = 0.2$)[5], and students who merely set preferences require (statistically) significantly less assistance per problem than those who do not set

[5] We omit mean values since Fancsali & Ritter (2014) normalize and transform process variables, following Ritter, Joshi, Fancsali & Nixon (2013) in such a way that mean values would be relatively un-interpretable without more significant explanation, which we omit for brevity.

preferences ($p = 0.03$, $d = -0.12$). Perhaps even more interestingly, students who set names of friends/classmates in MATHia outperformed those who did not set such names on all four process variable measures, requiring less assistance per problem ($p < 0.001$, $d = -0.27$) and encountering more MATHia sections ($p = 0.009$, $d = 0.16$) while working more efficiently ($p < 0.001$, $d = 0.12$) in less time ($p < 0.001$, $d = -0.23$).

These results raise questions about the nature of the factor(s) (either cognitive or non-cognitive) for which use of the preference dashboard is serving as a proxy; possibilities include conscientiousness or attentiveness to the MATHia environment, but since total login time is not significantly different in two out of the three comparisons, possible notions of engagement for which dashboard use might serve as a proxy may be constrained. It is also possible that students may simply appreciate the opportunity to set preferences to which the ITS responds and adapts. Future versions of MATHia should make more explicit the prevision of problems to students based on their interests, while also honoring preferences overall more frequently and with greater variability to avoid the restricted range pitfall that we have discovered in our current implementation.

Discussion and Implications for GIFT

One immediate implication, if our hypothesis about learners' appreciation of an instructional system's personalization to their interest areas or adaptation based on other non-cognitive factors is true, is that GIFT might also benefit from providing means by which training applications can not only provide adaptive instruction but also make explicit (at least aspects of) the decision making process that led GIFT to "choose" a particular adaptive, instructional strategy. We now discuss several other implications derived from the above presentation of recent research on specific non-cognitive factors in practice.

Online measures

We agree with the argument of Otieno, Schwonke, Salden & Renkl (2013) that developing appropriate online measures of non-cognitive and other factors is important to improving the state of the art of ITS research because surveys tend to be (at least) time-consuming while providing only noisy measures of intended phenomena. One implication for an intelligent tutoring architecture, then, is that the tutor may, in addition to its role as a source of instruction, play a role as a "sensor" for various characteristics and states. As noted by Fancsali, Ritter, Stamper & Nixon (2013), developments in sensor-free, data-driven "detectors" (e.g., Baker 2007; Baker & de Carvalho 2008) of both relatively complex behavior and affective states might be extended to factors like achievement goals, providing important avenues for future research into data-driven, online methods for making inferences about such phenomena.

Transferable inferences and grain size

We have provided concrete examples of (still unresolved) issues raised by Fancsali, Ritter, Stamper & Nixon (2013) in the ever-growing body of research on ITSs. Differing levels of granularity in measurement and analysis provide disparate but possibly important information about learners" interactions with ITSs like CT, and the effects of some types of personalization (e.g., preference-tailored word problems) may only be seen at the fine-grained level of individual problems, based upon their composition in terms of KCs.

While much of the work we have presented and our previous recommendations describing the integration of HPIT and GIFT (Fancsali, Ritter, Stamper & Nixon, 2013) has focused on relatively fine-grained measurements and inferences necessary for a particular ITS or instructional system to provide appropriately adaptive educational experiences, relatively coarse-grained assessments and inferences may be most appropriate for GIFT to provide rich learner models for multiple instructional systems or training applications. However, it is not, at this stage, clear how, for example, non-cognitive (and meta-cognitive) factors like achievement goals or affective states like boredom in an ITS for mathematics might transfer to other training applications (when even factors like achievement goals, as noted above, have variability at different grain sizes during instruction that can provide important information about learning outcomes). It isn't necessarily true that, for example, detection of a student as bored in one tutor has any relevance to another tutor; perhaps the second tutor is more exciting to the student. For this reason, it makes sense to consider communicating not just conclusions (e.g., "boredom") to the *Learner module* but also some of the evidence leading to that conclusion. A subsequent system or training application, coordinating with the *Pedagogical module* and *Domain module*, could then judge whether the conclusion is likely to obtain, given evidence.

Contrast these factors with what could plausibly be transferable assessments of other factors like learner interest areas, assuming such factors are relatively stable over time. Nevertheless, a great deal of factor-specific research across ITS platform, educational games, simulation-based training environments, and others is required to determine those factors (and at what level of granularity) that will allow for transferable inferences across training applications and at what level of granularity these factors are sufficiently "domain-independent" to be tracked by GIFT's *Learner module*. The type of factor-specific research described above (though only in the CT and MATHia platforms) and also pursued by a wide variety of researchers in the educational data mining community and elsewhere is a step in the right direction (cf. the brief review for ITSs like CT in Fancsali, Ritter, Stamper & Nixon, 2013), but determining how findings transfer across domains and training applications is vital.

This problem is not limited to non-cognitive factors: Even certain cognitive factors like low-level skill models may not transfer from one ITS for mathematics to another. For example, the types of errors that students can make in a desktop version of the CT's equation solver units are different than those that can occur in a tablet-based version of these units we are currently implementing as a part of the HPIT project that will use handwriting recognition. This has important implications for the types of competencies and/or skills that can be usefully tracked by the *Learner module* and the *Persistent Learner Model* in GIFT. Finding appropriate levels of granularity (i.e., grain sizes) for various factors that must be measured, assessed, and tracked by systems with ambitious goals like HPIT and GIFT is paramount to delivering intelligent tutoring in a variety of educational and training contexts.

References

Ames, C. (1992). Classrooms: goals, structures, and student motivation. *Journal of Educational Psychology, 84*(3), 261 – 271.

Anand, P.G. & Ross, S.M. (1987). Using computer-assisted instruction to personalize arithmetic for elementary school children. *Journal of Educational Psychology, 79*(1), 72 – 78.

Baker, R.S.J.d. (2007). Modeling and understanding students' off-task behavior in intelligent tutoring systems. In *Proceedings of the 2007 Conference on Human Factors in Computing Systems* (pp.1059 – 1068). ACM: New York.

Baker, R.S.J.d., de Carvalho, A. M. J. A. (2008). Labeling student behavior faster and more precisely with text replays. In R.S.J.d. Baker, T. Barnes & J.E. Beck (Eds.), *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 38 – 47). International Educational Data Mining Society.

Bandura, A. (1994). Regulative function of perceived self-efficacy. In M. G. Rumsey, C. B. Walker & J. H. Harris (Eds.), *Personal selection and classification* (pp. 261 – 271). Hillsdale, NJ: Erlbaum.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Self-Efficacy Beliefs for Adolescents* (pp. 307 - 377). IAP.

Barron, K. E. & Harackiewicz, J. M. (2001). Achievement goals and optimal motivation: testing multiple goal models. *Journal of Personality and Social Psychology, 80*, 706 – 722.

Bernacki, M.L., Nokes-Malach, T.J. & Aleven, V. (2012, April). Investigating stability and change in unit-level achievement goals and their effects on math learning with intelligent tutors. Paper presented at Annual Meeting of the American Educational Research Association, Vancouver, Canada.

Bernacki, M.L., Nokes-Malach, T.J. & Aleven, V. (2013). Fine-grained assessment of motivation over long periods of learning with an intelligent tutoring system: methodology, advantages, and preliminary results. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 629 – 644). New York: Springer.

Bollen, K. (1989). *Structural Equations with Latent Variables*. Wiley.

Corbett, A.T. & Anderson, J.R. (1995). Knowledge tracing: modeling the acquisition of procedural knowledge. *User Modeling and User Adapted Interaction, 4*(4), 253 – 278.

Cordova, D.I. & Lepper, M.R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology, 88*(4), 715 – 730.

Dweck, C.S. (1986). Motivational processes affecting learning. *American Psychologist, 41*(10), 1040 – 1048.

Elliot, A.J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist, 34*(3), 169 – 189.

Elliot, A.J. & McGregor, H.A. (2001). A 2X2 achievement goal framework. *Journal of Personality and Social Psychology, 80*(3), 501 – 519.

Elliot, A.J. & Murayama, K. (2008). On the measurement of achievement goals: critique, illustration, and application. *Journal of Educational Psychology, 100*(3), 613 – 628.

Fancsali, S.E., Bernacki, M.L., Nokes-Malach, T.J., Ritter, S. & Yudelson, M. (in press). Goal orientation, self-efficacy, and "online measures" in intelligent tutoring systems. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Fancsali, S.E., Ritter, S. (2014). Context personalization, preferences, and performance in an intelligent tutoring system for middle school mathematics. In S. Teasley & A. Pardo (Eds.), *Proceedings of the 4th International Conference on Learning Analytics and Knowledge* (pp. 73 – 77). ACM: New York.

Fancsali, S.E., Ritter, S., Stamper, J., & Nixon T. (2013). Toward “hyper-personalized” Cognitive Tutors. In R.A. Sottilare & H.K. Holden (Eds.), *AIED 2013 Workshops Proceedings Volume 7: Recommendations for Authoring, Instructional Strategies and Analysis for Intelligent Tutoring Systems (ITS): Towards the Development of a Generalized Intelligent Framework for Tutoring (GIFT)* (pp. 71 – 79). Sun SITE Central Europe (CEUR).

Fryer, J.W. & Elliot, A.J. (2007). Stability and change in achievement goals. *Journal of Educational Psychology, 99*(4), 700 – 714.

Ku, H.Y. & Sullivan, H.J. (2002). Student performance and attitudes using personalized mathematics instruction. *Educational Technology Research and Development, 50*(1), 21 – 34.

Muis, K.R. & Edwards, O. (2009). Examining the stability of achievement goal orientation. *Contemporary Educational Psychology, 34*(4), 265 – 277.

Otieno, C., Schwonke, R., Salden, R. & Renkl, A. (2013). Can help seeking behavior in intelligent tutoring systems be used as an online measure for goal orientation? In M. Knauff, M. Pauen, N. Sebanz & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 1103 – 1108). Austin, TX: Cognitive Science Society.

Pane, J.F., Griffin, B.A., McCaffrey, D.F., & Karam, R. (2013). Effectiveness of Cognitive Tutor Algebra I at scale. *Educational Evaluation and Policy Analysis*. Advance online publication. doi: 10.3102/0162373713507480.

Richardson, J.T.E., (2004). Methodological issues in questionnaire-based research on student learning in higher education. *Educational Psychology Review, 16*(4), 347 – 358.

Ritter, S., Anderson, J.R., Koedinger, K.R. & Corbett, A.T. (2007). Cognitive Tutor: applied research in mathematics education. *Psychonomic Bulletin and Review, 14*(2), 249 – 255.

Ritter, S., Joshi, A., Fancsali, S.E. & Nixon, T. (2013). Predicting standardized test scores from Cognitive Tutor interactions. In S. K. D’Mello, R. A. Calvo & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 169 – 176). International Educational Data Mining Society.

Ritter, S., Sinatra, A.M. & Fancsali, S.E. (in press). Personalized content in intelligent tutoring systems. In *Design recommendations for adaptive intelligent tutoring systems: Volume 2 - adaptive instructional strategies*. Orlando, FL: U.S. Army Research Laboratory.

Senko, C., Hulleman, C.S. & Harackiewicz, J.M. (2011). Achievement goal theory at the crossroads: Old controversies, current challenges, and new directions. *Educational Psychologist, 46*(1), 26 – 47.

Spirtes, P., Glymour, C. & Scheines, R. (2000). *Causation, Prediction, and Search*. 2nd Edition. Cambridge, MA: MIT.

Walkington, C.A. (2013). Using adaptive learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance learning outcomes. *Journal of Educational Psychology*, *105*(4), 932 – 945.