ITS
2018
Montreal
CANADA
11-15 JUNE
14TH INTERNATIONAL CONFERENCE ON
Intelligent Tutoring Systems
"A 30 Year Legacy of ITS Conferences"

ITS 2018 Workshop Proceedings

14th International Conference, ITS 2018

Montreal, QC, Canada, June 11–15, 2018

Nathalie Guin and Amruth Kumar (Eds.)

Preface

Four workshops, two tutorials and two industry-track workshops were held on Monday, June 11th and Tuesday, June 12th in conjunction with 14th International Conference on Intelligent Tutoring Systems (ITS 2018) in Montreal, Canada.

The general goal of the workshops was to offer participants an opportunity to engage in professional exchange in an interactive and interpersonal setting. These workshops brought together groups of researchers from around the world who have shared interests. They provided a good, extended opportunity for researchers to get to know each other in a relaxed atmosphere and stimulate the development of ideas and interests, sometimes leading to collaboration, grants and publications.

We are pleased to present the proceedings of the following four workshops:

- C&C@ITS2018: International Workshop on Context and Culture in Intelligent Tutoring Systems
- Learning analytics workshop: Building bridges between the Education and the Computing communities
- Exploring Opportunities for Caring Assessments
- Optimizing Human Learning: Workshop eliciting Adaptive Sequences for Learning

We hope the ideas presented in the proceedings provide the springboard for additional research and collaboration.

Amruth Kumar and Nathalie Guin, Workshop and Tutorial Chairs

Table of Contents

C&C@ITS2018: International Workshop on Context and Culture in Intelligent Tutoring Systems

Learning analytics workshop: Building bridges between the Education and the Computing communities

Exploring Opportunities for Caring Assessments

Optimizing Human Learning: Workshop eliciting Adaptive Sequences for Learning

C&C@ITS2018

International Workshop on Context and Culture In Intelligent Tutoring Systems

Valéry Psyché [1], Isabelle Savard [1], Riichiro Mizoguchi [2,3] and Jacqueline Bourdeau [1]

[1] LICEF Research Center, TELUQ University, Québec, Canada

[2] Research Centre for Service Science, Japan Advanced Institute of Science and Technology (JAIST), Nomi, Japan

[3] Laboratory for Applied Ontology (LOA), ISTC-CNR, Trento, Italy

Workshop Preface: International Workshop on Context and Culture in Intelligent Tutoring Systems

Valéry Psyché[1], Isabelle Savard[1], Riichiro Mizoguchi [2,3] and Jacqueline Bourdeau[1]

[1] LICEF Research Center, TELUQ University, Québec, Canada
[2] Research Centre for Service Science, Japan Advanced Institute of Science and Technology (JAIST), Nomi, Japan
[3] Laboratory for Applied Ontology (LOA), ISTC-CNR, Trento, Italy cc-its2018@teluq.ca

With the internationalization of education, the need for adaptation and flexibility in ITS and other learning systems has never been more pressing, extending to many levels and fields including: the international mobility of learners, teachers and researchers; the integration of international, intercontextual and intercultural dimensions in instructional programs (from primary to higher education and continuing professional development), as well as in the designs, methods, techniques and tools that support them; the international mobility of education viewed through the lens of today's new reality of mass open online courses accessible by a diverse range of learners around the world facilitated by ubiquitous, mobile and cloud learning systems. In this sense, there is a need for more research about context and culture in intelligent tutoring systems. Teachers and researchers need to develop new adaptation skills and embrace diverse contexts and cultures as well as leverage this diversity to foster the transfers that can enhance learning. Clearly therefore, it is important to make room for this diversity in curricula and learning systems and integrate transfer and adaptation concerns into pedagogical practice.

But how can we do this concretely? How can we best manage this complexity and leverage this diversity? How can this materialize in the ITS field, and what are the benefits?

One of the main focuses of current research is to define the boundaries of context and culture (C&C) as a theoretical concept and what constitutes the best methods, techniques and tools in order to collect, analyze and model it from an adaptive learning perspective. Until recently, C&C modelling was considered an intrinsic part of the various classical ITS architecture models. Aspects of C&C were therefore partially covered under the domain, learner, pedagogical and communication models. Now, however, the advent of big data in education and significant innovations in artificial intelligence are opening new doors for us to analyze and model C&C differently, if we are able to take advantage of the information available through the learning analytics process. Big data offers an exciting opportunity for us to look at C&C modelling for ITS through a new lens. Do we need a fifth model? Should we view it as another layer in the ITS architecture? Let's start thinking about it. In today's era of adaptive learning delivering anything learners need, anywhere and at any time, the potential for context and culture-aware ITS could be huge. What would knowledge representation and reasoning mechanisms look like in ITS? What kinds of limits might C&C represent for ITS? How can we identify or measure these limits? Can ocular and biometric measurement play an instrumental role? What are the logical next steps in terms of conducting studies about context and culture-aware ITS and gathering and analyzing data about context and culture?

This C&C@ITS2018 workshop aims to build the foundations of this research stream by forming an international research community and providing new avenues and questions for research. New avenues and questions for research may include the following: Will integrating context and culture mean changing traditional ITS architecture by proposing new models? Is there any interest in using AI innovations (big data, deep learning) with the modelling of context and culture knowledge? Why, knowing that there are many schools of thought? Where do we begin to combine our efforts? Do other modelling methods such as ontological engineering represent a better way to achieve this goal? Is it relevant to use AI techniques for education such as educational data mining or learning analytics to maintain up-to-date knowledge about contextual and cultural diversity? How can an lTS accommodate and leverage this new complexity to gain awareness of contextual and cultural diversity? How can earning analytics support contextual and cultural adaptation, and how can we combine the two? What is the role of the learner in contextual and cultural adaptation? How can contextual and cultural diversity make learning deeper and richer?

In light of the above, submissions are welcomed for this workshop on topics including, but not limited to, the following: Contextual theory; Ontological and cognitive modelling of contextual or cultural knowledge/context or culture-aware ITS; Context-aware collaborative learning; Contextual or cultural knowledge in ubiquitous, mobile and cloud learning systems and various application areas.

Context or Culture: What is the Difference?

Isabelle Savard[1] and Riichiro Mizoguchi[2, 3]

[1] TÉLUQ University, 455 Du Parvis, Québec, G1K 9H6, Canada
[2] Research Center for Service Science, Japan Advanced Institute of Science and Technology (JAIST), 1-1 Asahidai, Nomi, 923-1292, Japan
[3] Laboratory for Applied Ontology (LOA), ISTC-CNR, via alla Cascata 56/c, Povo, 38123, Trento, 38123, Italy
isabelle.savard@teluq.ca

Abstract. Literature can sometimes tend to present context and culture almost as synonyms. This creates ambiguity, which can complicate the consideration of contextual and cultural variables in instructional design, learning and teaching. From an ontological point of view, some clarification of these two concepts is essential as each may influence learning and teaching in different ways. Moreover, since context and culture are interconnected to a certain degree, one may influence the other. It is crucial to make a clear distinction between these two concepts in the knowledge models used in Intelligent Tutoring Systems (ITSs) if we want to facilitate 1) their consideration in pedagogical scenarios, and 2) the accumulation of knowledge about different contexts and cultures. This article offers an interpretation of the difference between these two concepts, presenting context as a substrate of culture. Contextual issues in the learning ecology are also discussed, based on this distinction.

Keywords: Context, Culture, Ontology, Learning Ecology

Ontology-Based Context Modelling for Designing a Context-Aware Calculator

Valéry Psyché[1], Claire Anjou[2], Wafa Fenani[3], Jacqueline Bourdeau[1],
Thomas Forissier[2], Roger Nkambou[3]

[1] TÉLUQ University, 5800, rue Saint-Denis, bur. 1105, Montréal (Québec) H2S 3L5, Canada
[2] ESPE Guadeloupe - École supérieure du professorat et de l'éducation - Guadeloupe
[3] Université du Québec à Montréal, 405 Rue Sainte-Catherine Est, Montréal, QC H2L 2C4
valery.psyche@teluq.ca

Abstract. This paper reports on the research conducted by a team from the France-Quebec research project TEEC, and its advances. This team is responsible for modelling and designing of a context gap calculator, the MazCalc. The MazCalc is a computer artifact aimed at measuring the effects of two distinct context with the same object of study. In a Context-Based Teaching project such as the one presented in this paper: Context Modelling is essential in identifying the context parameters needed to include in the design of the context gap calculator in order to predict context differences; At the same time, measurements provided by the MazCalc are essential to guide the design of learning scenarios aiming to produce context effects among learners. The article is divided into three parts. First, the contextual modelling is presented, then we discuss the design of the MazCalc, and finally, we address the challenges of this research, namely: (1) the definition of the didactic context and its modelling, leading to the identification and the prediction of context deviations; and (2) the articulation of this modelling with the specifications of the MazCalc artifact. Context modelling is done using an ontological approach. While the iterative design of the MazCalc in connection with the realization of design experiments is conducted according to the Design Based Research method. At the end, we discuss the next steps to be taken.

Keywords: Ontology-Based Context Modelling; Context-Aware System

1 Introduction

Context effects are pedagogical event occurring when there is a clash between student's conceptions, coming from distinct environmental contexts, and about a shared topic being studied. These effects can arise during communications between individuals involved and it allows them to realize the differences that exist in their conception of a same object depending on the context in which it is studied. Context effects can lead to the construction of richer and more complete conceptions on a given subject. The prior identification of differences in contexts relative to the object of study in the two contexts makes it possible to create collaborative learning scenarios aiming to produce context effects [1]. This model is called the CLASH model [1], and the TEEC project wants to test this hypothesis and validate the model using the Design Based Research (DBR) methodology described in [2]. In order to predict the potential emergence of context effects, a computer artifact was designed to parameterize contexts and calculate their differences. The ultimate ambition of this artifact is to provide input needed for the design of learning scenarios based on the effects of contexts.

Context modelling involves conceptualization, and abstraction; where concepts are specified with their components, properties and relationships among each other. It is, for each iteration of the DBR methodology, the first link in the chain that should produce context effects. The context model therefore, guides the learning scenario which in turn determines the (didactic) design experiments for data collection. It enables the researcher to contrast and contextualize and identify parameters. The first instrument used to model the context is the Meta model (ontology). The second is the context gap calculator which informs the specification of the parameters needed for computing the differences. This paper addresses two questions, then it looks at the challenges of this research, namely: (1) the definition of the didactic context and its modelling leading to the identification of parameters to be used in the prediction of context deviations; and (2) the articulation of this modelling with the specifications of the MazCalc artifact. Furthermore, the context modelling is done using an ontological approach. Finally, the next steps and problems addressed in both the ontology-based context modelling and the design of the MazCalc are discussed.

2 Ontology-Based Context Modelling

Ontological modelling dealing with contextual issues is a well-studies research topic[3-7]. However, so far, none of already existing studies have met the challenge of modelling the didactic context. The didactic context of a learning scenario is influenced by sociolinguistic, environmental or socioeconomic factors and their subsequent impact in the learning process. The theoretical framework of the didactic context has been described in [8]. In the TEEC project, our focus has been on studying the external context which concerns the impact of the environment and authentic situations on learning.

Vision and purpose of ontological engineering. Although ontology was initially defined by Gruber as “an explicit specification of a conceptualization” [9], other authors have sought to emphasize essential features of ontology that we feel are important to recall. First, we agree that an ontology be “a formal system with an explicit specification of a shared conceptualization” [10]. This means that an ontology is an abstract model of a world phenomenon whose appropriate concepts are identified (conceptualization). The type of concepts used and the constraints related to their use are defined declaratively (explicitly). In addition, ontology can be translated into interpretable language by a (formal) machine. Finally, an ontology captures consensual knowledge, that is, not reserved for a few individuals, but shared by a group or community (shared).

Moreover, when we speak of articulating ontology to the digital artifact design model, it is to these two definitions that we refer: “an ontology is a hierarchically structured set of terms for describing a domain that can be used as a skeletal foundation for a knowledge base” [11]; which “provides the means for describing the conceptualization explicitly behind the knowledge base” [12]. These definitions recall us that ontological engineering must be based on the final purpose and use of ontology, and on the services it will ultimately render. The purpose of this ontological engineering is therefore to specify a conceptualization (level 1) of the domain of didactic contextualization shared by the members of TEEC, then to formalize it (level 2) and then make it operational (level 3) in the context deviation calculator [13]. And that of context ontology is to describe the skeleton of the MazCalc knowledge base.

Ontological Modelling Process. The goal of this article is not to explain the ontological engineering method used. We rely on the MI2O method [14].

Among preliminary pilots, we selected geothermal energy as a topic that was subject to a detailed analysis [8] and led to MazCalc 1 (1st generation). This created a list of candidate terms. These terms discussed with the team were retained or not depending on their potential to correctly represent the field, that is, to become concepts. At this point, they were inserted into a concept dictionary (Table 1).

Table 1. Excerpt from the MazCalc Ontology Concept Dictionary

Concept	Definition	Property (part-of)	Relation (is-a)
Didactic Context	It is a sub concept of context. It can be social, internal or external (environmental). It is defined by a set of context parameters.	Has set of context parameters.	Is a Context. Is created by someone Is related to a learning scenario.
External Context	It is composed of a set of context parameters. We model the external context (not the social or internal ones).	Has set of context parameters.	Is a Didactic Context.
Context of study	It is an external context which is based on an object of study.	Has one or many context parameters clusters.	Is an External Context.
Context parameter cluster	It is part of Context of study. It is a non-exclusive set of context parameters from various themes. It was formally called: Family.	Has one or many context parameters.	Is a (sub) Context of study.
Learning Domain	Example: geothermal energy, language.	Has many Object of study	Is a Domain
Object of study	It is related to the learning domain and theme. It is dependent on the domain but not on the theme. e.g. in the domain of biology, an object of study is “frog”, and a theme is “nutrition”.	Has one or many themes. Has many contexts of study.	Is a (sub) Domain
Context parameter	A set of context parameters defines a context of study (the state of the context). Each context parameter belongs to one or more clusters. e.g.	Has a list of possible context parameter values.	

Concept	Definition	Property (part-of)	Relation (is-a)
	In the domain of geology, a context parameter is “type of roc.”	Has many types.	
Context gap	It is the gap between two context parameter values due to two distinct given contexts. Context Gap is the result of gap computing.	Has computed values	Is a gap

It should be noted that ontological engineering does not consist of creating a collection of terms (which are polysemous), but rather in extracting the concepts (which are explicit). This is an abstraction exercise that is essential for ontological modelling, and it involves the specification of concepts with their properties, as well as their relationships with other concepts within a conceptual network. In parallel to this process, several versions of an initial conceptual ontology (Figure 1), in the sense of [13], were created using GMOT software [15] and shown to experts in different didactic fields (geothermal energy, socio-history, language/French, environment and sustainable development [ESD]). It should be recalled that four design experiments are context modelling based.

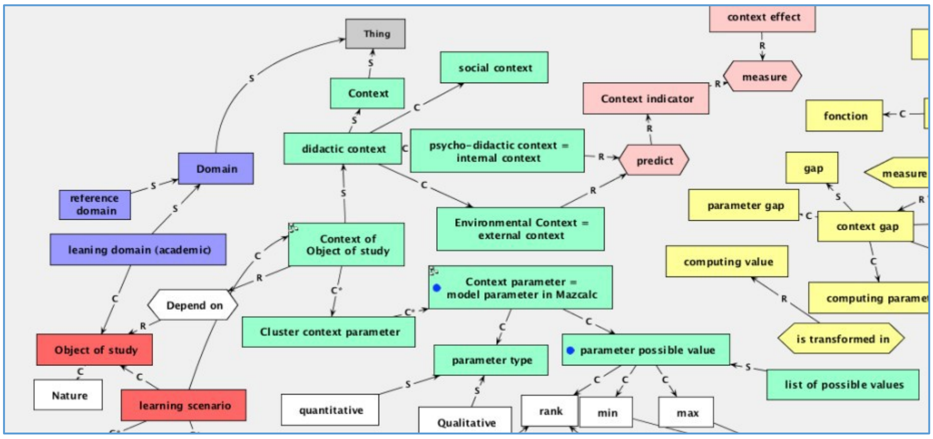


Figure 1. Graphical Representation of the MazCalc Ontology

The evaluation of the conceptual ontology was completed through several collaborative activities with different stakeholders. First of all, the ontology was explained to the content experts in order to verify that we had a common representation of the didactic context. Then, we addressed their feedback on the contextual representation of their didactic domain by replacing the ontology concepts by instances taken from the different versions of MazCalc 1 (MazCalc 1 applied to geothermal energy 2, language, socio-history and ESD). We also consulted about the ontology with the analyst responsible for the MazCalc 2 specifications. This third phase’s purpose was to compare the MazCalc 2 class diagram, a kind of skeleton of its database, with the ontology.

3 Context Gap Calculator: Models and Design

Consistent with Tchounikine's [16] views, MazCalc can be considered as a component of an intelligent tutoring system (ITS) [17] called CAITS, given that CAITS is "a system that works on knowledge," those specific to setting the context of an object of study in a given context, and "that manipulates symbolic representations." In this sense, the problems related to the design of the MazCalc are ITS engineering problems. It is therefore from this angle that we approached the design of the MazCalc and the challenges that flow from it.

MazCalc 1 and 2: genesis of context calculator. The MazCalc's engineering process was carried out in conjunction with design experiments in a connected classroom with collaborative learning, in order to test it. Several iterations of design and design experiments were set up jointly and informed the knowledge used to guide the project. Four phases illustrating the evolution of the project are detailed here.

Phase 1—Ideation during the GOUNOUIJ project: First design experiment whose scenario was based on differences in conceptions of the frog between primary school pupils in Guadeloupe and Quebec [18].

Phase 2—First iteration of MazCalc: MazCalc prototype, the MazCalc 1. First development of a computational tool in the form of a spreadsheet. This prototype enabled the creation of a learning scenario about geothermal energy during the GEOTREF project [8].

Phase 3—Second iteration—alpha version of the MazCalc: Launch of the TEEC project [2]. Creation of a web version of the MazCalc 2 (alpha version).

Phase 4 — Third iteration — MazCalc Beta version (in progress) : MazCalc 3.

MazCalc 3 Modeling. MazCalc 3 is a web computer tool that has been proposed to calculate the differences between contexts and predict their effects. But to successfully design such a tool, context modelling is very necessary to cover all cases and states of any context. The more detailed and clear the specifications, the higher the quality of the software.

Design specification. The specification definition consisted of describing the actors who will use this artifact (Table 2) and three types of design models: the use case diagram, the class diagram (Figure 2) and the sequence diagrams. The use case diagram showing how each actor is involved in a specific part of the calculator development and implementation. The class diagram shows all the objects that the MazCalc 3 tool will contain. The starting point of our work was to consider the assertion [19] that "the context of the study is described using context objects". Thus, modelling a study object amounts to modelling a context relative to its object (Table 2).

Table 2. Actors using the MazCalc

Actors	**Roles**
Actor 1: Cognitionist	Model a Meta model (Ontology, class diagram); Update the parameters of the Meta model.

Actor 2: Expert Designer of the Study Object	Model an object of study (related to the didactic field); Specify the parameters of an object of study; Specify the properties of parameters; Update the parameters of a study object.
Actor 3: Specialist of the object of study in its context	Instantiates an object of study in a given context = create a context; Assigns parameter values for a context model; Add a context parameter Update the values of the parameters.
Actor 4: Instructional Designer	Access the deviation calculation of each parameter; Access the result of the global calculation of the difference between the contexts.

Class diagram. The diagram that has caught our attention the most is the class diagram, as we see it as the design model for an ITS [16]. This model is the most important, it is the one that will be used as a comparator with the ontology of the didactic context, and how the two can be linked (see section 4). The object of study is defined by a set of parameters. These parameters are of the "qualitative" or "quantitative" type with "continuous" or "discrete", "bounded" or "not bounded" values. Each parameter belongs to one or more clusters (families). It can have a list of possible value. A parameter can derive from another parameter [8]. These specifications have been grouped into "Models" and "ModelParameters" tables, as well as their link with the "Family", "paramfamily", "paramValueTypes" and "ParamPossibleValues" tables (Figure 2). The table "Models" represents the model of an object of study and not its instance (with actual values). That is to say, Model is the skeleton of an object of study only. The field referenced in the "ModelParameters" table refers to its parent parameter. Here, the model of an object of study is constructed independently of the context to be studied.

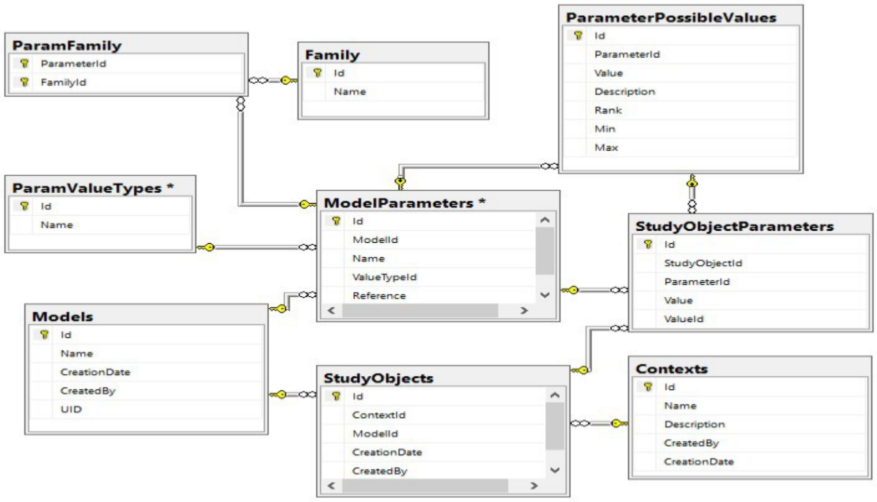


Figure 2. MazCalc3 Class and Object Diagram

The object of study in a context must have only one value for each parameter. Therefore the model is developed to produce to an object of study defined in the "StudyObjects" table, which is relative to a context. This relationship is respected by the link between the "Models", "StudyObject", and the "Contexts" tables (figure 2). Each parameter of the model of an object of study must have a unique value among its list of possible values. This value, for each parameter, is stored in the "StudyObjectParameters" table and is extracted from the existing values in the "ParamPossibleValues" table. This explains the link between the "StudyObjects", "StudyObjectParameters", "ModelParameters", "ParamPossibleValues" tables (Figure 2).

MazCal 3 Conception and Implementation. The MazCalc 3 database is created based on the class diagram. It allows to define, via MazCalc 3, all types of study objects independently of the context, which makes MazCalc a generic tool. It allows to create several objects of study, and to instantiate several contexts in relation to a single object of study. In order to calculate the difference between two contexts, we calculate the difference between each parameter of these two contexts. The formulas for calculating the context gap are under discussion.

The MazCalc 3 tool is still under development. And, yet many tasks have been completed. For instance, the database is implemented, but it can evolve according to the evolution of the modelling of the objects of studies as well as the formulas for the gap computing, as stated by the DBR methodology [2]. The main human-machine interfaces have also been created: the one for the generation of models, one for the definition of parameters and their value types, one for the definition of all possible values for each parameter as well as the instantiation of contexts with respect to the object of study.

4 Challenges in Modelling and Articulating its Models

4.1 Models to Understand Theories and to Design Artifacts

On the one hand (Challenge 1), we had to model to understand what is meant by "didactic context" in order to serve the needs of the TEEC project, i.e. to measure contextual gaps. Starting from the concept dictionary (Table 1), we now wish to give an overview of the discussions conducted to reach a consensus during the modelling. Especially around terms which have been difficult to define such as the term "Family".

Examples of problems related to Metamodel modelling. "Family" Case.

For some members of the Modelling team, "Family" was understood as a theme, a learning area, or a scale. But, for others, it was seen as a grouping of context parameters. For them, the concept of "Learning Domain" which is a well-defined concept, could not be associated with "Family", since in an ontological view, it is quite clear whether a term corresponds to a concept or not: one tries to construct the specification with components, properties and relationships, and if one does not succeed, then this term probably does not have the status of a concept in this ontology. Thus, if the term does not pass the test of conceptualization, this is probably because it is already taken into account somewhere else with another label.

Examples of problems related to domain context modelling. “Language” Case. Let us take the case of the design experiment “Language”. This experiment is experimental in the sense that it is more difficult than others to quantify in order to calculate the differences in context. Thus, we encountered the problem of representing the “quantification” of context parameters in order to calculate the context gap.

Other very beautiful problems of transposition of theories into models have also arisen. For example, the “oral nature of the narrative situation” cannot be modelled as a sub concept of “Intrigue”. We must therefore find another idea to place orality in ontology. To better understand the problem, let us try to explain it differently: in ontology, we have the concept “object of study”. In the case of the didactic situation Language, perhaps the object of study is “the story”. For the “object of study” concept to respond well to the principles of ontological engineering, a sub concept of the “Object of study” concept would have to be created.

Table 3. Illustration of a modelling problem

Concept = Object of study= tale;
o Subconcept = oral story (=orality, event, actors, space-time dimensions, unforeseen);
o Subconcept = written story (=document, whether or not a transcription of the oral story).

With this example, we see that we can, in the written tale, make a reference to the oral tale. It must therefore be included in the ontology so that it is representative of all possible cases of the target domain to represent. The two previous examples clearly show the similarity between the modelling problems of the class diagram and those of ontological modelling. This brings us to our challenge: articulating these two types of resulting models.

4.2 Models to Design Artifacts

On the other hand (Challenge 2), we had to define and model the design intent of the artifact [16]. This is software engineering work leading, among other things, to the production of a class diagram.

Example of a problem related to challenge 2. Modelling of the “Parameter (context implied)” class. One of the main problems encountered concerns the modelling of context parameters, the latter leading to the calculations of context deviations. In particular, we have tried to answer the following questions: What defines a parameter? What are its attributes (type, nature, properties)? Should the parameters be prioritized? Should parameter values be differentiated according to their type (constant or variable)?

4.3 Articulation of Models

Articulate models to understand theory and models to design the artifact (challenge 3) [20]. The difficulty was to completely transpose the “theoretical” model, the ontology resulting from the work of the “Context Modelling” team, to the design model, the class diagram, resulting from the “Context Calculator Development” team. However, we soon realized that we were facing the same modelling problems. Before we spoke, we had encountered problems in representing certain concepts/classes. A concrete example of a common problem we faced was to represent the concepts of “Context parameter”, “Parameter value” and “Possible parameter value”. Questioning each other and sharing our representations has allowed us to improve both models.

5 Next Step in an ITS Point of View

Next steps concerning the context modelling. The problem of merging between the Context Modelling team and the design Experiment teams is still to be developed in TEEC. It is a weak link in the TEEC project, which is engaged in a chain of production of context effects: modelling with calculation of the gap and probability of context effects, learning scenarios, experiments and data analysis. Fortunately, with the DBR methodology, we are able to deal with “real life” and learn from each iteration of the production chain for the next.

In addition to the context ontology, we plan to construct a domain ontology for each contextualized domain. Next, the line between the meta-model (ontology) of the context and the domain model must be drawn. Normally, ontology governs models as instantiation, which inherit them. If this is not possible, it is because either the Meta model has a flaw, or the domain model must conform to it.

We also plan to build an ontology of context effects. Next, the line between the meta-context model and the meta-context effects model must be drawn.

Next steps concerning the context gap calculator. So far, MazCalc has been developed as an independent tool, and will remain like this until its design and implementation are completed. But ultimately it will be part of a context-sensitive learning software suite (with authoring and tutoring services), and it is the core of the CAITS, a “Context-Aware Intelligent Tutoring System” [21]. The CAITS comprises three main components: The Context-Sensitive Domain Model (CSDM); the Context-Sensitive Teaching Model (CSTM) and the Context-Sensitive Learner Model (CSLM). MazCalc will share its results with the CAITS component by connecting with its CSDM; this connection will make it possible to provide the ITS with context effect information which will drive the domain model behaviour [22]. This is why the MazCalc 3 was designed as an API web application (to exchange services to the CAITS), rather than a simple web application.

Ultimately, once the development of the MazCalc is completed, it should be able as well to provide a service to the learning designer to specify and adjust the instructional scenario (Actor 4); and serve as a reference in the analysis of experimental data to validate the CLASH model [1]. Indeed, one of the mandates of the Data Analysis team is to detect weaknesses in the elements of our causal chain that are supposed to produce context effects: the context modelling for each iteration, the scenario, the experimentation, and the data collection device. So, the quality of the MazCalc is essential, since it conditions the other elements.

References

1. Forissier, T., et al., *Modeling Context Effects in Science Learning:The CLASH Model*, in *CONTEXT 2013*, P. Brézillon, P. Blackburn, and R. Dapoigny, Editors. 2013, Springer. p. 330-335.
2. Bourdeau, J. *DBR, une Méthodologie de Recherche pour le Design d'Environnements d'Apprentissage*. in *Context 2017*. 2017.
3. Gu, T., et al. *An ontology-based context model in intelligent environments.* . in *Communication networks and distributed systems modeling and simulation conference*. 2004.
4. Krummenacher, R. and T. Strang. *Ontology-based context modeling*. in *Proceedings*. 2007.
5. Ejigu, D., M. Scuturici, and L. Brunie. *An ontology-based approach to context modeling and reasoning in pervasive computing*. in *PerCom' 07*. 2007. IEEE.
6. Strang, T. and C. Linnhoff-Popien. *A Context Modeling Survey*. in *First International Workshop on Advanced Context Modelling, Reasoning And Management at UbiComp 2004*. 2004. Nottingham, UK.
7. Bettini, C., et al., *A survey of context modelling and reasoning techniques.* Pervasive and Mobile Computing, 2010. **6**(2): p. 161-180.
8. Anjou, C., et al., *Elaborating the Context Calculator: A Design Experiment in Geothermy*, in *International and Interdisciplinary Conference on Modeling and Using Context*. 2017.
9. Gruber T., *A Translation Approach to Portable Ontology Specifications.* Knowledge Acquisition, 1993. **5**(2): p. 199-220.
10. Studer R., Benjamins V. R., and Fensel D., *Knowledge engineering: Principles and methods.* Data Knowledge Engineering, 1998. **25**(1-2): p. 161-197.
11. Swartout B., et al., *Towards Distributed Use of Large-Scale Ontologies.* Spring Symposium Series on Ontological Engineering, 1997: p. pp.138-148.
12. Bernaras A., Laresgoiti I., and Corera J. *Building and Reusing Ontologies for Electrical Network Applications*. in *Proc. of the 12th ECAI96*. 1996.
13. Mizoguchi R. *A Step Towards Ontological Engineering.* in *12th National Conference on AI of JSAI*. 1998.
14. Psyché, V., *Rôle des ontologies en ingénierie des EIAH : Cas d'un système d'assistance au design pédagogique*, in *Informatique*. 2007, Université du Québec à Montréal: Montréal. p. 527.
15. Paquette, G., *Visual Knowledge and Competency Modeling - From Informal Learning Models to Semantic Web Ontologies*. 2010: Hershey, PA: IGI Global.
16. Tchounikine, P., *Educational Software Engineering*, in *Computer Science and Educational Software Design*, P. Tchounikine, Editor. 2011, Springer. p. 111-122.
17. Nkambou, R., R. Mizoguchi, and J. Bourdeau, eds. *Advances in intelligent tutoring systems*. Vol. 308. 2010, Springer Science & Business Media.
18. Forissier, T., J. Bourdeau, and S. Fécil, *Interfaces Elève-Machine pour apprendre à partir des contextes*, in *IHM'14*. 2014. p. 38-43.
19. De Lacaze, T., *Caractérisation de particularités environnementales liées au développement durable en Guadeloupe : conceptions d'acteurs locaux.* 2015.
20. Baker, M., *The roles of models in AIED research: a prospective view.* IJAIED, 2000. **11**(2): p. 122-143.
21. Nkambou R., Bourdeau J., and Psyché V., *Building Intelligent Tutoring Systems: An Overview*, in *Advances in Intelligent Tutoring Systems*, Nkambou R., Bourdeau J., and Mizoguchi R., Editors. 2010, Springer. p. 16.
22. Bourdeau, J., et al. *Web-Based Context-Aware Science Learning*. in *WWW'15*. 2015. ACM.

Ontological Support for the Cultural Contextualisation of Intelligent Learning Environments for Adaptive Education

Phaedra S. Mohammed

Department of Computing and Information Technology,
The University of the West Indies, St. Augustine, Trinidad and Tobago
Phaedra.Mohammed@sta.uwi.edu

Abstract. Within ITS research, most systems rely on data in order to train models for decision making and for customising system behaviour. The inherent bias has been traditionally in favour of developed nations. This paper examines the issues involved in contextualising interactive intelligent educational systems using a semantic approach that leverages the meaning of data rather than common patterns within data. It presents a trio of ontologies for relating conceptual knowledge to sociolinguistic terms in the context of a student's cultural influences and background. The paper argues that if an ITS can model students culturally, model their languages, and model their cultural concepts, then it would be possible for an ITS to start communicating with students socially and conceptually in a culturally appropriate way. The paper explains the rationale behind the need for ontological concepts when adapting aspects of instruction, how they relate to cultural lexical terms, and examples of when these terms may be suitable for use in educational content and instructional events.

Keywords: Ontologies, Cultural Semantics, Student Modelling, Sociolinguistic Contexts, Content Adaptation, Semantic Analysis

1. Introduction

In 2010, there were approximately 1,991 million Internet users worldwide [11]. Compared to 2016, that figure increased to 3,385 million. Not only has the sheer volume of users increased, the cultural backgrounds of these users are being quickly diversified. In just under 10 years, the proportion of Internet users from the developing world has almost doubled in relation to those from the developed world. In 2008, the ratio of developed world users to developing world users was approximately 4.2. In 2017, that ratio is now 2.0. Moreover, 70% of the world's youth (aged 15-24) are online and they make up the largest group of Internet users [11]. Two interesting points arise from these statistics. Firstly, a lot of data is being generated daily and this will continue to increase. Secondly, as the human sources of this data change, so does the quality of the data, and more importantly the cultural bias.

Within ITS research, most systems rely on data in order to train models for decision making and for customising system behaviour. The inherent bias has been traditionally in favour of developed nations [2] and this makes sense since most users in the past have been predominantly from these areas. ITS research would have therefore been driven by the cultural backgrounds and biases of the researchers who produced the systems and the student users who produced data that fed the research. The problem here is that data biases affects the design of an ITS and the eventual decisions made by the system. The bias can be positive or negative, and educational systems need to be more acutely aware of this because of the impact on learning and rates of success. For instance, statistical analysis of large amounts of data allows prediction of various types of instructionally relevant events that might take place next with a fair level of accuracy. This allows models to be built based on the observation of patterns in the data which help to give an indication of the details of some domain of interest. The flexibility of the patterns that are detected however, depend heavily on the kinds of data that the models are trained on which in turn affects the scaleability of the system overall [8].

Culturally-aware ITS design is a reasonable way of dealing with this lack of flexibility since, as the statistics show, the landscape of the student audience is changing and systems need to evolve or risk irrelevance. It is difficult however to transfer and extend intelligent learning environments to different cultural contexts for several reasons [14,19]. Diversity arises from differences between cultures. While tangible and concrete in many instances, such as language, dress, food, gestures, and music, culture at its deepest level is intangible and non-deliberate. Furthermore, the multiple factors and influences that shape an individual person's cultural awareness come through interactions, perceptions and knowledge of other cultural groups. Culture itself is therefore challenging to model computationally in a holistic sense and even more complex when aiming to do this for an individual learner within an ITS. It necessitates organising cultural semantics and data from heterogenous sources to reduce bias and also because individual data points such as country of origin or language are insufficient for meaningful modelling.

Semantic web technologies have been around for many years but widespread uptake has not been achieved [18]. This is subject to change in the upcoming years as the importance of linked data becomes evident with the need to organise and structure data [5]. This paper argues that rather than taking a data centric approach towards cultural inclusiveness, a semantic approach is preferable since it allows the meaning of the data to be leveraged rather than common patterns. Ontological modelling of cultural contexts would allow data from heterogenous sources to be filtered, disambiguated and combined. The paper describes a trio of ontologies that were developed for modelling cultural contexts in intelligent learning environments. The ontological representations covers three main areas: modelling a student's cultural context, modelling a student's language and cultural expressions, and modelling the cultural concepts (metaphors, idioms, concepts) that are relevant to a student. Each ontology is useful in isolation for various purposes, however when all three are merged, they give insight regarding how to communicate with a student using appropriate sociocultural concepts and language.

The rest of the paper is organised as follows. Section 2 defines the process of cultural contextualisation. Section 3 describes the trio of ontologies: CSM, CERA and

VELO. Section 4 illustrates how concept chains produced when the ontologies are merged result in the identification of appropriate cultural terms and concepts for a given students. It also gives examples of how these may be used in instructional events. The paper concludes in Section 5.

2. Defining Cultural Contextualisations

Culture refers to a cognitive and linguistic framework within which humans interact with and relate to their environment [10,13]. Interactions are governed by societal and ideological systems of thought [12] and result in the construction, distribution and assimilation of shared meanings that originate from individual and group level perceptions. These shared meanings, also called *cultural conceptualisations* [17], result from human cognitive processes of categorising observations and experiences under familiar conceptual categories. These categorisations are intrinsically linked to language which conveys cultural knowledge and allows individuals to understand each other's perspectives when communicating. *Contextual groups* are defined as collections of individuals with common beliefs, characteristics and values who reference cultural conceptualisations through shared linguistic terms. *Cultural contextualisation* is therefore defined as the process of integrating one or more cultural conceptualisations into aspects of a digital learning environment [16]. Cultural conceptualisations manifest as concrete representations of abstract concepts and are comparable to *cultural elements*. Defined in the literature as an observable manifestation of culture, *cultural elements* are categorised as material artefacts or non-material cultural products which represent or embody the shared meanings of a cultural group [4]. For the purposes of this paper, cultural elements and contextual elements are used interchangeably.

3. Ontological Descriptions of Cultural Context

An intelligent learning environment that aims to model cultural contexts will rely heavily on semantic metadata. This is necessary in order to reason about the cultural contexts of educational resources and relate these contexts to a student's cultural background. Many standard upper-level ontologies define general knowledge concepts that relate to cultural descriptions of real-world phenomena and provide foundational semantic bridges between intermediate levels of cultural knowledge abstraction. Upper ontologies have not been designed with the intention of structuring cultural knowledge in particular. Recent work by Blanchard and Mizoguchi [3] describes high-level cultural conceptual entities in an upper ontology of culture (MAUOC) and identify several categories of cultural elements that manifest in a culture. In addition, ontological concepts should be defined such that lexical entries irrespective of the source language are all accessible by these concepts, that is, through ontological mapping and merging. The following subsections describe the trio of ontologies introduced in this paper using UML notation.

3.1. Contextual Student Model (CSM) Ontology

The ontological structure of the CSM is extensible for capturing and modelling multiple cultural backgrounds. Figure 1 shows the main concepts and relationships in the CSM ontology. It is partitioned into three layers consisting of factors and influences originating from various sources. The first layer stores personal demographic data that define a student's core identity. The second layer consists of dimensions from immediate socio-cultural units that play formative roles in a student's life such as family members and close friends. The third layer consists of dimensions from neighbouring socio-cultural units that are of lesser influence but still contribute towards a student's awareness of and exposure to cultural contexts. This is possible because the *Guardian* and *Contextual_Group* concepts (and related attributes) and relationships can be instantiated any number of times with dimension data. This implies that a student's cultural background can be modelled not only from a single temporal perspective indicated by the student's age, but also from a chronological perspective where his/her cultural background may change with age.

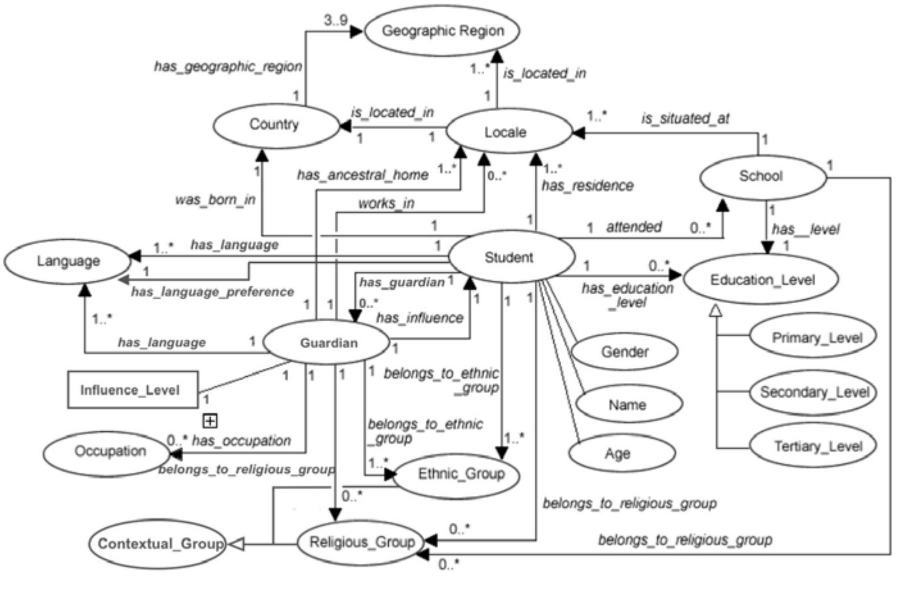


Fig. 1. The Contextual Student Model Ontology

3.2. Contextual Element Resource Annotation (CERA) Ontology

Observable manifestations of culture have been referred to as cultural elements, or more generally, as contextual elements [4]. High level categories that represent language independent abstractions of real world phenomena are described in [3, 15]. Based on these abstractions, the Contextual Element Resource Annotation (CERA) ontology specifies the ontological concepts and relationships that describe the nature and background of a contextual element which is referred to as an *Entity* in Figure 2 which shows the ontological signature of CERA. The More Advanced Upper Ontology of Culture (MAUOC) [3] and SUMO[1] (Suggested Upper Merged Ontology) were used to build the semantic backbone of CERA. SUMO provided a comprehensive hierarchy of spoken human languages used by members of a contextual group and helped to define the language origin of linguistic concepts that are used to describe one or more contextual elements (identified as dark grey concepts in Figure 2). The MAUOC on the other hand, provided high-level classifications of entity abstractions (identified as light grey concepts in Figure 2) namely *Physical Entity, Continuant Entity, Abstract Entity, and Semi-Abstract Entity* concepts which were subsumed by the *Entity* concept in CERA. The Entity concept is linked to a *Contextual_Group* concept.

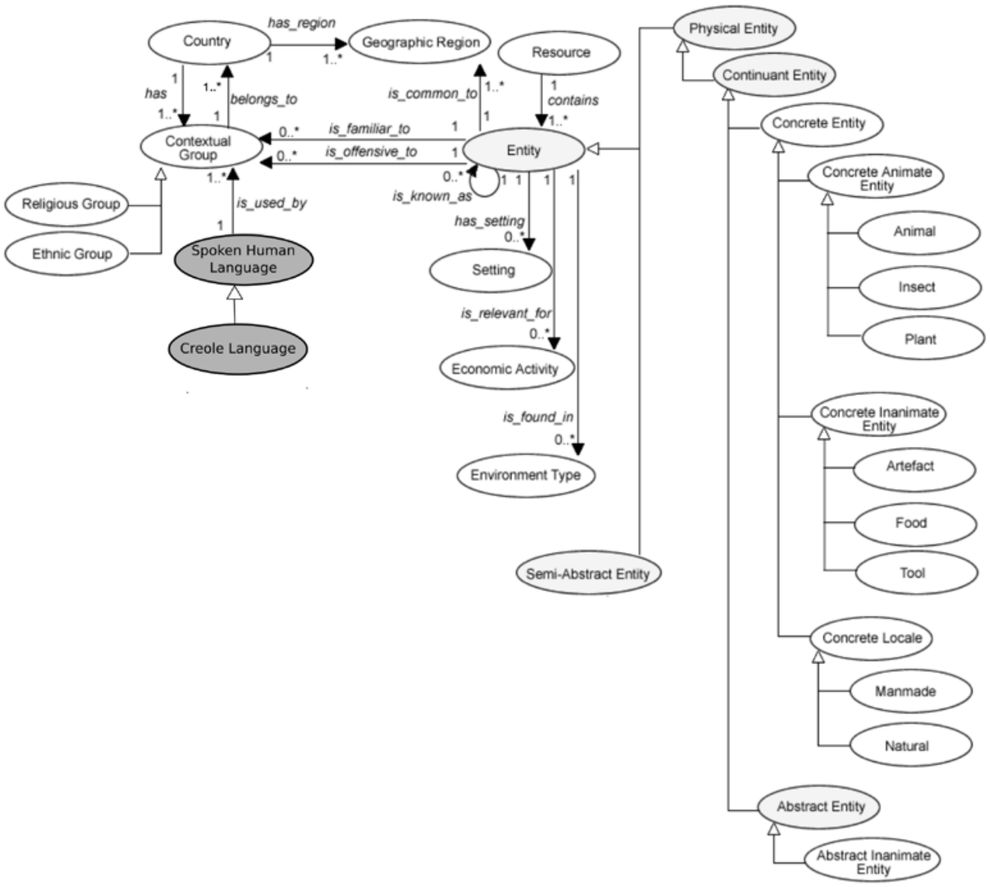


Fig. 2. The Contextual Element Resource Annotation Ontology

[1] http://www.ontologyportal.org/

3.3. Vocabulary Equivalence Lexicon Ontology (VELO)

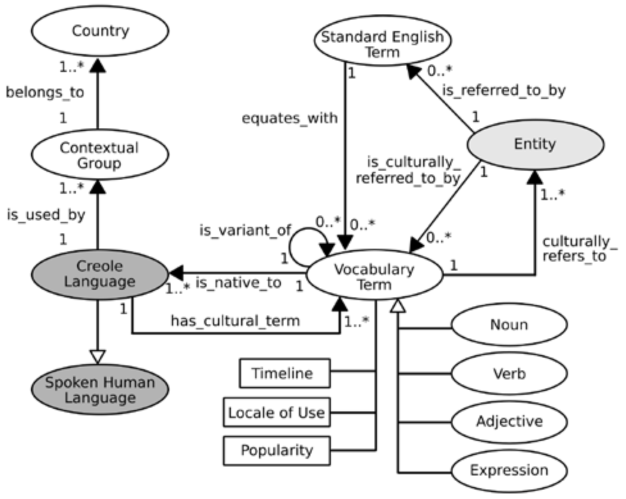


Fig. 3. The Vocabulary Equivalence Lexicon Ontology

The main concepts of VELO, the relationships between the concepts, and the attributes of the concepts are shown in Figure 3. VELO was designed to facilitate the mapping necessary for equating multiple vocabularies accurately. The ontology is based on the conceptual-linguistic approach described by [1], and adopts a similar structure to the ontologies in the DOSE platform [6] and the KYOTO project [21] by referencing upper-level concepts from SUMO and DOLCE. The intention behind VELO is to equate/map Standard English vocabulary to localised equivalents. It specifies the base concepts and relationships needed for achieving lexical equivalence across languages at the semantic level through the *Entity* concept. This can then be used for facilitating queries on communicative acts, language concepts, metaphors, and idioms that are culturally appropriate for a student using an ITS.

4. Deployment in Intelligent Learning Environments

4.1. Ontological Mapping and Merging

Ontological mapping and merging is necessary in order to combine the information distributed across the three ontologies described in the previous section. Figure 4 shows a partial snapshot of the important concepts in the ontological signature of the merged ontologies. Correspondence throughout the merging process is facilitated based on the use of the *Entity* concept in both VELO and CERA. Using the concept chain illustrated in Figure 4, it is possible to determine which contextual elements (referenced by *Entity* concepts) are suitable for a student based on familiarity through a student's affinity to one or more contextual groups in a society. Furthermore, the specific language terms that reference the concept can now be identified, leveraged and integrated into instructional events using rules.

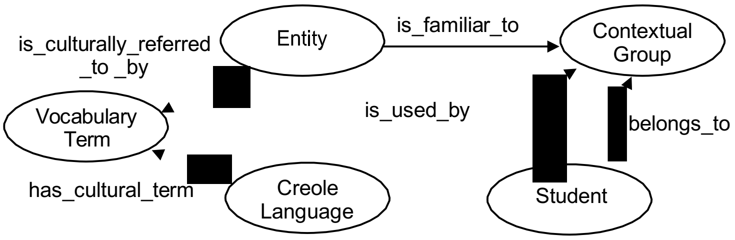


Fig. 4. Merged Partial Ontological Signature of the VELO, CERA and CSM Ontologies

To illustrate, consider two original sentences S1 and S2 which might be used in an ILE to respectively set the frame for a problem description, and give feedback to a student with a Trinidadian cultural context.

S1: Every week, John gives away free apples to the customer with the largest purchase.
S2: You did not answer the question correctly.

When S1 is provided as input to an ILE that uses the trio of ontologies, the resultant sentence S3 below would be produced for the student used in this example.

S3: Every week, John gives away free zabocas to the customer with the largest purchase.

In S3, the cultural reference to 'zabocas', would be matched conceptually under same semantic category through a shared higher level *Entity* concept as that of 'apple'. This cultural term would be used if a Trinidad English Creole vocabulary base is activated in VELO. Consequently, the general reference (apple) in S1 would be replaced with a more culturally-specific and culturally appropriate reference based on the student's cultural background as in S3 using rules. This demonstrates how the cultural semantic context of the educational material was changed while still preserving the learning context. When S2 is provided as input, there are several possible resultant sentences as shown in S4, S5 and S6 below.

S4: You did not answer the question correct.
S5: You eh answer the question correct.
S6: Yuh eh answer the question correct.
S7: Yuh eh answer d question correct.

In S4, the underlined words would be changed by grammatical rules loaded due to the activation of a Trinidad English Creole rule base since the student has a Trinidadian context. This gives an ILE the ability to produce appropriate localised variants of a source text when a particular level of formality is specified. For example, if formal variants are requested for S2, then only S4 would be generated. If very informal, colloquial variants are requested for S2, then S7 would be generated. It should be noted that the rules and ontologies facilitate different languages and cultural backgrounds. The design is not tied to a particular implementation as in this example. Therefore, if a student has a Jamaican context or a Singaporean context, the cultural references used would vary and therefore the output produced would vary.

4.2. Integration into Instructional Events

Instructional design models specify instructional events that take place during the learning process. A popular model often used in educational software was developed by Gagné [9] who identified nine instructional events. Based on the work of Branch [7], who linked culturally-aware instruction to these events, Table 1 was developed. It lists practical ways of using different types of contextualised content produced using the trio of ontologies for some of these types of instructional events.

Table 1. Using Contextualised Content for Instructional Events

Instructional Event	Contextualised Approach
Gaining the learner's attention	Integrate contextual elements, that are appropriate for the student, into instructional content as a form of stimulus change
Informing the learner of instructional objectives	Use a formal language variety that the student approves of and can relate to when stating instructional objectives
Presenting material to be learned	Use cultural references, scenarios, analogies in text, audiovisual or multimedia content
Providing learner guidance	Use a language variety that the student can relate to when giving instructional hints, directions or tips in order to provide meaningful context
Drawing out learner performance	Use familiar language expressions to encourage the learner to reflect using learning probes such as review quizzes
Providing informative feedback	Use familiar language expressions to phrase corrective feedback and inform the learner of the degree of answer correctness

For example, when providing informative feedback or drawing out learner performance for students who use a particular language variety in everyday life, the contextualised intensity of text-based sentences can be varied to create emotive feedback ranging from formal to informal, and also varying in the number of cultural references, metaphors and idioms used.Another example is the use of contextualised images when aiming to enhance retention and transfer or gain the student's attention. Images that depict contextual elements that the student is familiar with and which match the student's cultural background can be used to increase the relevance of the instructional content from a cultural perspective. A final example is the use of contextual elements in unexpected but instructionally and semantically appropriate places within text-based content. These elements when inserted in place of similar, semantically-relevant references in scenarios or questions descriptions can be used to gain a learner's attention or enhance the presentation of the learning material. The approach in the paper is currently suitable for an individual learner using an ILE. Collaborative learning challenges are more complex and require a different strategy for customising an ILE to deal with multiple learners with different cultural influences.

5. Conclusion and Future Work

The self-contained model of a traditional ITS is changing. In the past, the focus was on ensuring quality regarding what students learned. This has progressed to coaching to ensure that students learn effectively [20], and now the focus is on the kinds of students that are involved in learning from an ITS. If we can model students culturally, model their language, and model their cultural concepts, the focus would then be to communicate with them socially and conceptually in a culturally appropriate way. The next steps to consider are whether it is acceptable to communicate in culturally informed ways, and to determine when such communication is acceptable or not. The need to consider cultural ethics and privacy is more important now than ever. For example, students from some cultures may be reserved and having an outward display of (somewhat privately-used) cultural realism in an ITS can be frightening and startling. This might make users uncomfortable and suspicious and which could eventually affect successful usage and uptake of such an ITS in a practical way. The ontologies described aim to mitigate such effects and extend the current efforts to model cultural knowledge for intelligent learning environments. They are a first step in addressing the need for practical, reproducible approaches towards cultural contextualisation from conceptual, linguistic, and cultural perspectives.

References

1. Agnesund, M.: Representing Culture-Specific Knowledge In A Multilingual Ontology. In: Proc. International Joint Conference on Artificial Intelligence Workshop on Ontologies and Multilingual NLP, Nagoya, Japan, August 23-29, 1997. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.28.2928&rep=rep1&type=pdf (1997).
2. Blanchard, E.G.: Socio-Cultural Imbalances in AIED Research: Investigations, Implications and Opportunities. International Journal of Artificial Intelligence in Education 25(2), 204–228 (2015).
3. Blanchard, E.G., Mizoguchi, R.: An Introduction to the MAUOC, the More Advanced Upper Ontology of Culture, and its Interest for Designing Culturally-Aware Tutoring Systems. Research and Practice in Technology Enhanced Learning 9(1), 41-69 (2014).
4. Blanchard, E. G., Mizoguchi, R., Lajoie, S.P.: Structuring the Cultural Domain with an Upper Ontology of Culture. In: Blanchard, E.G., Allard, D. (eds.) The Handbook of Research on Culturally-Aware Information Technology: Perspectives and Models, pp. 179-212. IGI Global, Hershey, PA(2011).

5. Blumauer, A.: Linked Data - The Next 5 Years: From Hype to Action. Available online: https://www.linkedin.com/pulse/linked-data-next-5-years-from-hype-action-andreas-blumauer (2016).
6. Bonino, D., Corno, F., Farinetti, L., Ferrato, A.: Multilingual Semantic Elaboration in the DOSE Platform. In: Proc. ACM SAC, Special Track on Web Technologies and Applications, 14-17 March, 2004, pp.1642-1646. ACM, New York (2004)
7. Branch, R.M.: Educational Technology Frameworks that Facilitate Culturally Pluralistic Instruction. Educational Technology 37(2), 38-41 (1997).
8. Emani, C.K., Cullot, N., Nicolle, C.: Understandable Big Data: A survey. Computer Science Review 17, 70-81 (2015).
9. Gagné, R. M.: Instructional Technology Foundations. Lawrence Erlbaum Associates:,Hillsdale, New Jersey (1987).
10. Hofstede, G., Hofstede G-J., Minkov M.: Cultures and Organizations, Software of the Mind. Intercultural Cooperation and Its Importance for Survival. McGraw Hill, NY (2010).
11. ICT Facts and Figures 2017. Available online: https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2017.pdf
12. Jost, J. T., Hamilton, D.: Stereotypes in Our Culture. In: Dovidio, J.F., Glick, P., Rudman, L.A. (eds) On the Nature of Prejudice: Fifty years after Allport, pp. 208-224. Blackwell, Oxford (2005).
13. Kashima, Y., Gelfand, M.J.: A History of Culture in Psychology. In: Kruglanski, W.A., Stroebe W. (eds.) Handbook of the History of Social Psychology, pp. 499-520. Taylor & Francis, Hove, East Sussex (2012).
14. Mohammed, P., Mohan, P.: Breakthroughs and Challenges in Culturally-Aware Technology Enhanced Learning. In: Proc. Workshop on Culturally-aware Technology Enhanced Learning in conjunction with EC-TEL 2013, Paphos, Cyprus, 17 September, 2013. http://cats-ws.org/wp-content/uploads/2013/06/AIED2013-CATS Proc.pdf (2013).
15. Mohammed, P., Mohan, P.: Representing and reasoning about cultural contexts in intelligent learning environments. In: Eberle W., Boonthum-Denecke, C. (eds.) Proc. 27th International Florida AI Research Society (FLAIRS) Conference. May 21–23, 2014, Pensacola, USA, pp 314–319. AAAI Press, Palo Alto, CA. (2014).
16. Mohammed, P., Mohan, P.: Dynamic Cultural Contextualisation of Educational Content in Intelligent Learning Environments using ICON. International Journal of Artificial Intelligence in Education. 25(2), 249-270 (2015).
17. Sharifian, F.: On Cultural Conceptualisations. Journal of Cognition and Culture 3(3),187-207 (2003).
18. Shvaiko, P., Euzenat, J.: Ontology Matching: State of the Art and Future Challenges. IEE Trans. Knowledge and Data Engineering. 25(1), 58-176 (2013).
19. Nye, B.D.: Intelligent Tutoring Systems by and for the Developing World: A Review of Trends and Approaches for Educational Technology in a Global Context. International Journal of Artificial Intelligence in Education. 25(2), 177–203 (2015).
20. VanLehn, K.:The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist 46(4), 197-221 (2011).
21. Vossen, P.., Agirre, E., Bond, F., Bosma, W., Herold, A., Hicks, A., et al.: KYOTO: A Wiki for Establishing Semantic Interoperability for Knowledge Sharing Across Languages and Cultures. In: Blanchard, E.G., Allard, D.(eds.), The Handbook of Research on Culturally-Aware Information Technology: Perspectives and Models, pp. 265-294. IGI Global, Hershey, PA(2011).

Relevance of the cultural dimensions in affective-cognitive behavior during interaction with an intelligent tutoring system

N. Sofia Huerta-Pacheco[1], Genaro Rebolledo-Mendez[1], Victor Aguirre[2],
Sergio Hernandez-Gonzalez[1] and Erica Maria Lara-Muñoz[1,3]

[1] Faculty of Statistics and Informatics, Universidad Veracruzana, Mexico
[2] Statistics Department, ITAM, Mexico
[3] Instituto Tecnológico de Alvarado, Mexico
{nehuerta, grebolledo, sehernandez}@uv.mx,
aguirre@itam.mx, emlaram@itsav.edu.mx

Abstract.
Cultural Dimensions, as stipulated by different theoretical perspectives such as Hofstede's, are normally not considered to define student models. These cultural dimensions consist of traits that can be attributed to students and include both cognitive and affective characteristics. Some dimensions indicate students' ability to represent an effect in the affect which may be useful to predetermine affective models. This research project hypothesizes that students' cultural dimension may indicate affect tendency during the use of Intelligent Tutoring Systems (ITS). The methodology consisted of determining students' cultural dimensions, cognitive achievement, and analyzing affective responses (self-reported) when the student used the ITS on an individual way. The results suggested that there are affective behaviors associated to a Hofstede cultural dimension (Power distance index). The implications of these results are that some cultural characteristics may predict students' affective behaviors employing an ITS for mathematics. Additionally, affect models could be used to predefine affective-cognitive scaffolding.

Keywords: affective-cognitive states, cultural dimensions, intelligent tutoring systems, secondary education.

1.1 Introduction

The technological tools are current elements that contribute to the teaching- learning process of students at different educational levels, which are shown with contents of topics specialized in some areas.

These tools are designed so that users (students) have innovative elements, however, when referring to the adaptation of the tools to the user, there are several problems in the interaction, since they are not fully developed to adapt to the particular needs or characteristics of each user [1].

However, these reasons have not precluded several researches to identify some relevant characteristics that impact on learning with technology such as collaboration [2], cultural dimensions [3], learning styles [4], motivation [5, 6], affect [7–9] and among others. The aim of this study is to analyze whether students' cultural dimensions are related to both affect and knowledge during interaction with the intelligent tutoring system.

In this research, we focus on individual student factors used in all the interaction with an Intelligent Tutoring System (ITS) for mathematics when they acquire knowledge about variables (numerical and categorical) and the way they represent them. To do this, there are characteristics that are affected by the environment where the student works in a learning process, such is the case of cultural dimensions. Since students' cultural dimensions traits lies in that teaching instructed in the classrooms and the learning environment.

In the association of affection and cognition, particularly, there are several studies applied with technology [10–13], that allude that the affection presents predominant tendencies in the learning process (negative, neutral and positive) [8], which can be regulated for the student to acquire either greater or better knowledge.

On the other hand, the importance of culture in education shows contrasts that impact the cognitive process [14, 15]. Cultural dimensions are divided into five dimensions described by Hofstede, these dimensions alone represent influential factors in society as the Power distance, Uncertainty avoidance, Individuality, Masculinity and Long term orientation [3, 16].

In Mexico's basic education system, it is considered that an environment conducive to learning must indispensably contemplate the recognition of influential physical, affective and social factors in cognitive achievements in an individual and group manner [17], making relevant the study of the characteristics of the students, as well as their behaviors in the classroom.

Considering the above is done the following research question: What cultural dimensions are present and how these influences the acquisition of knowledge and the affect of students during the use of a ITS?

The research focuses on identifying associated cultural behaviors that give indication to be able to define the students' profiles, and thus provide elements considering their cultural and affective characteristics during the interaction with an intelligent tutoring system.

2 Methodology

This work was performed at the secondary school "Federal N. 2 Julio Zárate" in Xalapa, Veracruz, Mexico for four days. It was considered to be a simple random sampling (n=50 students) of five groups (N=110 students) in the first year on 2017 of secondary school with 62% of female and 38% of male with an age range of 12 to 14 years old.

The materials used consist of the intelligent tutoring system "Scooter tutor" [18, 19] in the non-reactive version (without Scooter agent), the two isomorphic tests of learning employed on similar experiments [18], the standardized questionnaires of cultural dimensions [16], the self-report of the affective states, and props. The evaluation was guided under the standards of the Belmont report [20].

Standardized learning tests are isomorphic measuring instruments designed to evaluate students' knowledge of the development of scatter plots before and after interaction with the intelligent tutoring system. To calculate the level of knowledge (test scores) of students, points are obtained in percentage by standard terms of evaluation defined by the system creator [18] and these tests measure the cognitive achievement in such a way as to identify the increase obtained by the students. Achievement is calculated with the following equation:

$$Cognitive\ Achievement = Score\ of\ Post\_test - Score\ of\ Pre\_test$$

The registration of affective self-reports is given through a booklet, which presents the five most relevant states in a learning situation with technology [8]. This is through the issuance of student judgments about their affective status at intervals of every 8 minutes during the two sessions of interaction with the ITS. The records of affective trials are composed of images with random faces (emoticons) referring to the states of boredom, frustration, confusion, concentration and the absence of affec- tion of the neutral state. The affective measure reported is given in terms of proportions of cases through interaction, and they are distributed in negative (boredom and frustration), neutral (absence of affection) and positive (confusion and concentration) tendencies.

Cultural dimensions test stated by Hofstede [3] employed in this research is obtained through an adaptation of the instrument of the 1994 version [16], this consists of 20 items with five to six categories of ordinal scale type Likert. In addition, each item is weighted in an equation per dimension providing a representative score of the level, either low (*Index<=33 points*), normal (*33 points>Index<66 points*), or high (*Index>=66 points*). These dimensions present different representations such as Power distance that is defined as the extent to which the less powerful members of community within a society expect and accept the power other person or Uncertainty avoidance is as the extent to which members of community within a society feel threatened by uncertain, unknown, ambiguous or unstructured situations. On the other hand, in Individualism a person is expected to take care of himself and his immediate family, just as Masculinity represents a society in which social roles of gender are clearly different and Long-term orientation represents a society that encourages future re- wards-oriented virtues, particularly adaptation, perseverance and savings.

It is important to mention that this test does not present an adequate validation and reliability [21], however, it is necessary to observe the internal structure by dimension and the biases in the answers.

The experimentation included the application of the tests and the interaction with the ITS. There were four experimentation stages during the mathematics class.

1. *Initial test*: This stage consisted of an explanation of the topic “Scatter plots” (10 minutes), the first learning test (20 minutes) and other questionnaires (20 minutes) in the classroom.

2. *Interaction I*: In this phase, the student first performed the interaction with the intelligent tutoring system for 40 minutes in the media classroom and self-reported affective states in interruptions during the lapse of 8 minutes.
3. *Interaction II*: In the same way that in the stage Interaction I, the student worked with the intelligent tutoring system for 40 minutes in the media classroom and self-reported affective states in interruptions during the lapse of 8 minutes.
4. *Final test*: The student was given the Post-test on a 20-minute period in the media classroom, as well as the cultural dimensions test (15 minutes) and participants were thanked for their participation in the research (5 minutes).

3 Result

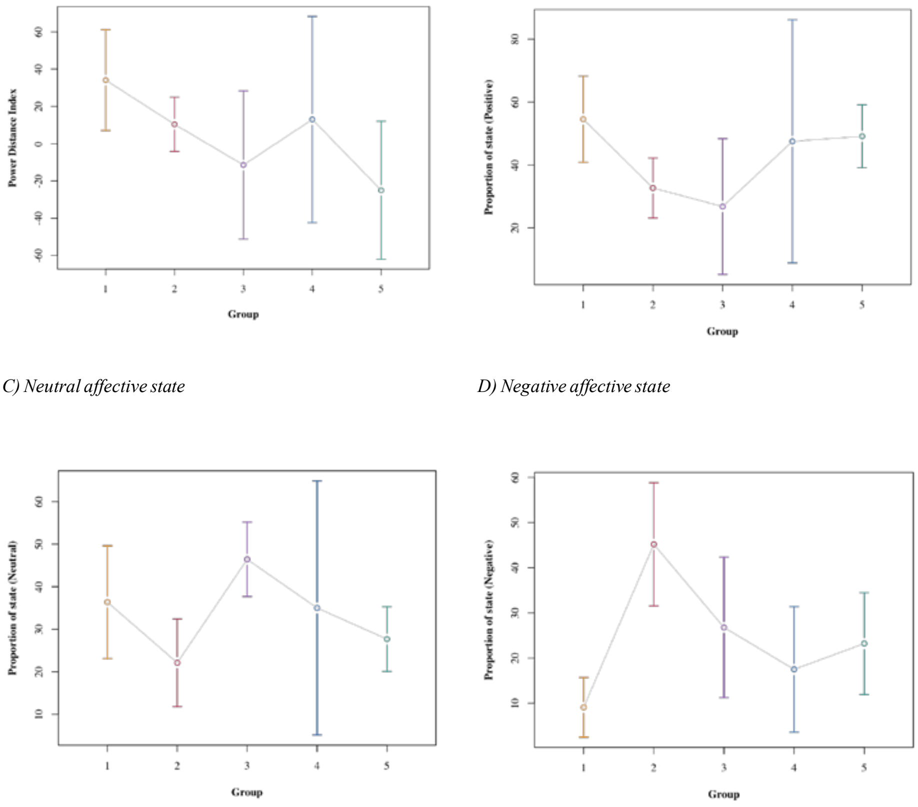
The preliminary findings in the interaction with the intelligent tutoring system present relevant characteristics to influence the affective-cognitive student behavior. It is significant to mention that the analyzed information did not assume the assumption of normality, the test score (pre-test and post-test) was measured in percentage points and worked with affective tendencies (negative, neutral and positive) and the results were assessed with nonparametric statistical techniques in R [22] and just considering the cases of positive achievement (*Cognitive Achievement* > 0).

The comparisons (pre-test) between the five groups, showed no significant differences (*K−W chi−squared=3.64, p-value=0.45*). However, all groups showed a high proportion (more than 60%) of neutral affective states during the initial time of interaction with the intelligent tutoring system. In addition, it was observed that all groups in the performance showed *42.75* average proportion score of the positive affective state and *25.75* average score of the negative states and differences by group in the proportion of affective tendencies.

On the other hand, it was observed that only one dimension showed the existence of significant difference *(p-value<0.05)* between the groups of the Power distance (PDI), showing that group 1 manifests a normal level *(mean=34.0, sd=40.30)* to differences of the other groups (see Figure 1-A) and a general average lower (*mean=2.9, sd=49.48*) than the all groups and much variation with respect to their average value. In addition, high levels (*Index>=66 points*) on average identified of Uncertainty avoidance (UAI), Individualism (IDV) and Masculinity (MAS) and normal average index in Long-term orientation (LTO). (see Table 1)

In the same way that significant differences were identified *(p-value<0.02*) between the pre-test and post-test and not in the post-test by group (*K−W chi−squared= 5.94, p-value=0.20*). Moreover, the post-test had a significant association (*r_s=0.323, p-value=0.02*) with the positive affective states, moreover the positive affect with Cultural dimension of the Power distance index (*r_s=0.326, p-value=0.02*).

Nevertheless, it showed a significant difference per group related to the proportion of positive affective states (*K−W chi−squared=10.74, p-value=0.02*), negative states (*K−W chi−squared=18.19, p-value=0.001*), neutral affective states (*K−W chi−squared=11.75, p-value=0.01*) and the Power distance index (*K−W chi−squared=9.07, p-value=0.04*), the results also presented that the some groups with the lowest index (*Index<=33 points*) for Power distances showed less representation in the positive trend of affective states and only the group 2 high proportion of negative trends. (see Figure 1)


A) Power distance index

B) Positive affective state

C) Neutral affective state

D) Negative affective state

Figure 1. Comparison by group and characteristics (affect and Power distance index)

Figure 2 shows the Principal Component Analysis [23] represent 61.01% varia- tions of the behavior of the affective states association with the Cultural dimension and Learnings scores (pre and post-test), this identifies and confirms that the positive affective trends (AE-Positive) are oriented to Power distance (PDI) and the post-test presents a high association with the pre-test as well as with the Power distance index and positive states. Finally, the negative tendencies (AE-Negative) do not present any significant association with the learning scores when only considering students with a cognitive achievement.

Table 1. Descriptive statistics (Cultural dimensions)

	Cultural dimensions				
Statistics	**PDI**	**UAI**	**IDV**	**MAS**	**LTO**
Number of Observations	50	50	50	50	50
Median	5	92.50	82.5	75.0	40.0
Mean	2.9	83.80	73.8	72.8	43.6
Standard Deviation (n-1)	49.48	71.20	63.18	87.99	22.38
Coefficient of Variation	1706.486	84.96	85.61	120.87	51.34

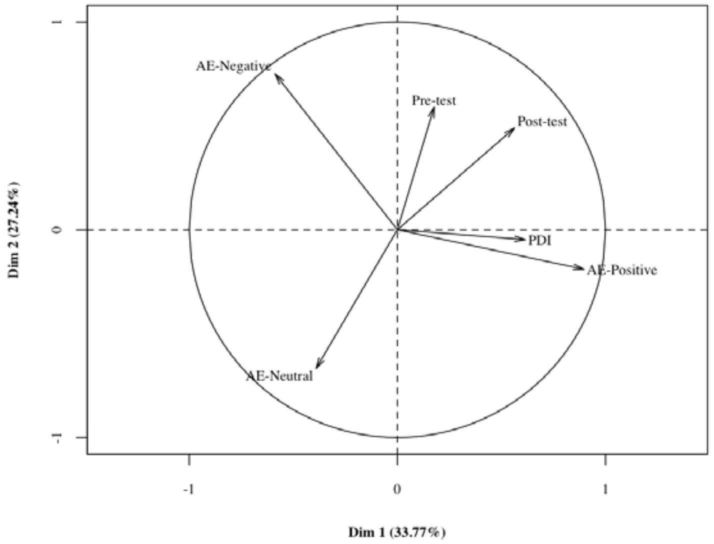


Figure 2. Representation of the characteristics in learning process

4 Discussion

This research project presents results suggesting different patterns of individual student' behavior, which were observed during the use of educational technology (ITS) for mathematics at the secondary level in Mexico. The exploration of independent characteristics (cultural dimensions, affect and cognitive achievement) is relevant because it allows understanding the student profile in a preliminary way during the learning process mediated with technology, contributing with information about the cultural criteria of the student who is likely to affect the academic environment of Mexican students.

The results suggest that there are significant associations between the cultural dimensions (Power distance index) and cognitive-affective states. This can be explained as the positive affective behavior of students may be closely associated to power distance in normal level to obtain higher score in the post-test.

In particular, considering this dimension will allow Mexican students to demonstrate positive states conducive to learning math issues by setting aside levels of traditional academic hierarchy.

However, it is important to mention that the affective measurement of students during the use of technology can be considered as an exploratory measure of the affection that the student presents according to his/her judgement, however, this requires specialized metrics [19] or to measure awareness and regulation [10] of the same over their states.

As a future work, it is proposed to evaluate other characteristics that affect the cognitive process in order to elicit a model of the user who is able to react to factors that are not conducive to learning. This model will allow creating a motor of inference that provides before the interaction of the students a profile to identify if these requires the use of a common intelligent tutor system or one with affective elements of regulation for to increase cognitive achievement and improve the interaction.

Acknowledgments. The first author acknowledges the financial support from the Mexican Council of Science and Technology (CONACYT, Scholarship 421460) and the third author gratefully acknowledges financial support from the Asociación Mexicana de la Cultura, A. C. Mexico.

References

1. Cabero J, Educativa JT (2001) Diseño y utilización de los medios en la enseñanza. Tecnol Educ 16–72
2. Ogan A, Walker E, Baker RSj, Rebolledo Mendez G, Jimenez Castro M, Laurentino T, De Carvalho A (2012) Collaboration in cognitive tutor use in Latin America: Field study and design recommendations. In: Proc. SIGCHI Conf. Hum. Factors Comput. Syst. pp 1381–1390
3. Hofstede G (2011) Dimensionalizing cultures: The Hofstede model in context. Online readings Psychol Cult 2:8
4. Kolb DA (1981) Learning styles and disciplinary differences. Mod Am Coll 1:232–255
5. Harter S (1978) Effectance motivation reconsidered. Toward a developmental model. Hum Dev 21:34–64
6. Rebolledo-Mendez G, Du Boulay B, Luckin R (2006) Motivating the learner: an empirical evaluation. In: Int. Conf. Intell. Tutoring Syst. pp 545–554
7. Porayska-Pomsta K, Mavrikis M, D'Mello S, Conati C, Baker RSj (2013) Knowledge elicitation methods for affect modelling in education. Int J Artif Intell Educ 22:107–140
8. Craig S, Graesser A, Sullins J, Gholson B (2004) Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. J Educ media 29:241–250
9. Baker RSj, D'Mello SK, Rodrigo MMT, Graesser AC (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. Int J Hum Comput Stud 68:223–241

10. Grawemeyer B, Mavrikis M, Holmes W, Gutierrez-Santos S (2015) Adapting feedback types according to students' affective states. In: Int. Conf. Artif. Intell. Educ. pp 586–590

11. D'Mello S, Lehman B, Pekrun R, Graesser A (2014) Confusion can be beneficial for learning. Learn Instr 29:153–170

12. Grawemeyer B, Wollenschlaeger A, Gutierrez S, Holmes W, Mavrikis M, Poulovassilis A (2017) Using graph-based modelling to explore changes in students' affective states during exploratory learning tasks.

13. Eagle M, Corbett A, Stamper J, McLaren BM, Baker R, Wagner A, MacLaren B, Mitchell A (2016) Predicting Individual Differences for Learner Modeling in Intelligent Tutors from Previous Learner Activities. In: Proc. 2016 Conf. User Model. Adapt. Pers. pp 55–63

14. Hofstede G (1986) Cultural differences in teaching and learning. Int J Intercult relations 10:301–320

15. Eldridge K, Cranston N (2009) Managing transnational education: does national culture really matter? J High Educ Policy Manag 31:67–79

16. Hofstede G (1998) Attitudes, values and organizational culture: Disentangling the concepts. Organ Stud 19:477–493

17. SEP (2017) Modelo educativo para la educación obligatoria. México

18. Baker RS (2005) Designing intelligent tutors that adapt to when students game the system. Carnegie Mellon University Pittsburgh

19. Rodrigo MMT, Baker RSj, Agapito J, Nabos J, Repalam MC, Reyes SS, San Pedro MOCZ (2012) The effects of an interactive software agent on student affective dynamics while using; an intelligent tutoring system. IEEE Trans Affect Comput 3:224–236

20. Department of Health E and W (2014) The Belmont Report. Ethical principles and guidelines for the protection of human subjects of research. J Am Coll Dent 81:4

21. Kruger T, Roodt G (2003) Hofstede's VSM-94 revisited: Is it reliable and valid? SA J Ind Psychol 29:75–82

22. R Development Core Team R (2011) R: A Language and Environment for Statistical Computing. R Found Stat Comput 1:409

23. Jolliffe IT (1986) Principal component analysis and factor analysis. In: Princ. Compon. Anal. Springer, pp 115–128

Cultural factors linked to collaborative learning in Intelligent Tutoring System in the Domain of Mathematics

Erica María Lara-Muñoz[1,2], Genaro Rebolledo-Méndez[1], José Rafael Rojano-Caceres[2], Nery Sofía Huerta-Pacheco[1]
[1] Facultad de Estadística e Informática, Universidad Veracruzana,
Av. Xalapa esquina Boulevard Manuel Ávila Camacho s/n. 91020 Xalapa, Veracruz, México
[2] Instituto Tecnológico Superior de Alvarado,
Escolleras Norte s/n, col. La Trocha, Alvarado, Veracruz, México
[1] {grebolledo, rrojano, nehuerta}@uv.mx, [2]emlaram@itsav.edu.mx

Abstract. The integration of technology into education requires a thorough analysis of the elements necessary to adapt it to the teaching-learning process, based on appropriate contextual analysis. This article presents the initial identification of elements or variables for the conceptualization of a collaborative model used in a mathematics Intelligent Tutoring System, deployed for secondary school students. Two exploratory studies were undertaken, the first to determine how students will be assigned to collaborative activities as to optimize the learning experience, and the second to identify the elements that influence collaboration and the extent to which collaboration is linked to cultural issues. The main contribution of this paper is to show the results of the second study, in which it was found that the association between collaborative and cultural elements, allow to improve the student's learning gains in collaborative activities use an Intelligent Tutoring System.

Keywords: Collaboration, Cultural Dimensions, Intelligent Tutoring System.

1 Introduction

Socials and cultural factors fundamentals to collaborative learning in technology mediated environment, allow that students improved their learning experiences and get greater benefits in it. To do this, the scales that Hofstede [1] suggest as cultural dimensions, and social elements as organization, participation, dialog, role and responsibly they offer the support to do it.

In several investigations [2], [3], [4], [5] it has been observed that students when interacting with educational technology have the opportunity to increase their level of learning, in addition, if the technology can be adapted within this process of learning, this will provide the necessary assistance that the student requires [6].

On the other hand, the changes of models and educational modalities, lead to certain aspects of migration or improvement in the teaching learning process, one of these aspects is the role of students, their become more dynamic entities in charge of the construction of his own knowledge [7]. Another aspect is the interaction of the student with classmates to carry out academic activities, this communicative and interactive process is given through collaboration where two or more people exchange opinions to create meanings. For this, there are adaptable and intelligent web-based education systems, called AIWBES [8], which adapt the user's preferences and knowledge, individually and in group, during interaction with this system. In this sense, social interactions that promote active and vicarious learning can also be carried out, where students can learn by directly doing exercises or observing activities that others do [9].

The relationship that some students may have with each other, allows each of them to include elements from different contexts, because although they live in similar environments, they may present different personalities, attitudes, knowledge and emotions to face similar situations, this difference is given for the culture that each one presents. Living in the family, at school, on the street, are what denote this difference in individual and collective behavior [10].

Unfortunately in Mexico it is a fact that the mathematics level is below the OECD average, results show that up to 57% of the students do not even reach the basic level of competences, that is, they cannot represent mathematically a Real-world situation, such as comparing the total distance between two alternative routes or converting prices to a different currency [11]. This is an alarming situation, due to this, the interest to include educational technology as a mathematics Intelligent Tutor System within the learning process in secondary school, but not only to include the tutor in this process, but also to adapt in the Intelligent Tutor System, collaborative and cultural elements that further promote student learning.

2 Collaboration in the educational process

Understand by collaboration to the knowledge construction process that originates in the social interrelation of people who share, compare and discuss ideas [12]. It is through this interactive process that the student builds his own knowledge [13].

Within the educational context, collaboration is an interactive form of learning where students must participate as equals, adding efforts, skills, knowledge, talents and competence that lead them to define a series of activities and tasks that allow them to reach their common goal.

By incorporating collaborative activities in the classroom, the teaching-learning process can be enriched, especially if the participation of students is more actively, generating in this way, the construction of their knowledge, fostering collaborative learning and improving the interpersonal relationships.

One of the important benefits of collaboration is the learning that can be obtained from this, when students participate in argumentation and negotiation activities, share and discuss ideas from each person's perspective and reach the consensus of the collaborative group [14]. Collaborative learning is a didactic technique that allows students to be guided in an educational environment, where they can interact with classmates and teachers, enriching the teaching-learning process to achieve their academic goals. In an environment of this type, students assume different roles, responsibilities, share experiences, knowledge and must be engaged by participating in joint processes, for the solution of specific activities in favor of their learning.

However, not all forms of grouping students to work collaboratively, leads to the best outcome [7]. Adequate group formation and structured interactions are important elements to increase the possibility of having a beneficent collaboration in a pair students [14].

As the formation of work groups is analyzed to obtain learning benefits, it should also be studied whether collaborative elements hat influence the learning process of students. One of the collaborative components used as part of this experiment to measure the collaboration of students was the Collaboration Test [15], which consists of 12 multiple-choice questions of nominal scale, from which information is obtained with relationship to five subscales of collaboration such as organization, participation, dialogue, role and responsibility. This test was applied with the goal to understand the kind of collaboration the participants think they had during the interaction with their teammate in the collaborative activity. Each of these subscales included in the test collects information on some of the questions as shown in table 1.

Table 1. Subscales in collaborative test.

Subscale	Question
Organization (S1)	Q1, Q6, Q8
Participation (S2)	Q3, Q4, Q5
Dialogue (S3)	Q2, Q3, Q5, Q9, Q12
Role (S4)	Q6, Q7, Q8, Q11
Responsibility (S5)	Q10, Q11

3 Cultural dimensions

The social behaviors observed in different countries are influenced mainly by thoughts and customs of the own culture [16]. Geert Hofstede is a research sociologist who explains the discrepancy between the behavior of different cultures, through a theory called cultural dimensions, this theory offers a panorama to examine how cultural values affect the behavior of people to act in a or another way.

The cultural dimensions of Hofstede are indicators that show the behavior of a complete society, not a single individual, however, this does not mean that one culture is better than another or has more value, but that the behavior of each is different from the other or not, according to the region [16], even within the same culture, there can be several subcultures which make up a global culture [17] within which can be observed different behaviors and opinions.

The first dimension to which Hofstede refers is the power distance index (PDI), here we can see how the members of a society, question or not, to the people who have the highest hierarchy, that is, in a society with great power distance, the members of a society do not question those who have higher levels, however a society with low power distance, each person has equal power between members of a group or community.

A second dimension is individualism (IDV) versus collectivism, in which it is observed if the members of a society are integrated in a group, or the link between one person or another is weak, that is, he prefers to make individual decisions and focuses only on the "me" and not on the "us".

Another dimension is masculinity (MAS) versus femininity, which refers to the way in which roles are distributed in society through gender. In a highly masculine society people are driven by competences and results, they are ambitious. Within society with low masculinity or femininity, people are more focused on building good relationships and ensuring a high quality of life for all.

The uncertainty avoidance index (UAI) refers to the way in which people feel in unfamiliar situations, in cultures with strong UAI, people avoid risks and unexpected situations since you are creating stress and anxiety. People with low UAI are more tolerant in unexpected situations, they are more relaxed and flexible.

People with long term orientation (LTO) encourage to be thrifty and to invest, respect traditions and fulfill social obligations such as respecting their elders and people of different ranks, on the contrary, those with short term orientation are encouraged to spend and want to make immediate profits, these people believe that the status between members is not important, unless they can get some benefit from them.

Although Hofstede's work has been done to know the influence of culture on the values that people have at work, and that their research gives an idea of what other cultures are like, and which factors are predominant in the organizational scope, its results have prevailed over time and its dimensions have been used even in the educational field, adapting the questionnaire to be applied to students [16].

The Hofstede cultural dimensions test consists of 20 questions, four questions for each of the five dimensions, the purpose of this test was to find some element that intervened positively in the results of the students.

4 Intelligent Tutoring System

The beginning of Intelligent Tutoring Systems gave rise to the moment when Artificial Intelligence (AI) was being worked on to imitate natural intelligence through the creation of machines that could achieve a human thought, these systems have been an important part in the area of IA in Education to create an environment of instruction that resembles a teacher in his teaching process.

These Intelligent Tutors Systems began to be developed with the purpose that knowledge could be imparted in some intelligent way to guide and assist a student in their learning process, so that they sought to emulate the behavior of a human tutor who could adapt to the behavior of the student, identifying the way in which this can solve a problem to provide the cognitive help required, when required and tailored to the student.

Intelligent Tutors Systems by their own nature were created to be used individually, however, it has been shown [18] that students in Mexico work collaboratively, even when it is an Intelligent Tutor, they get up from their places to ask questions to their classmates and complete their activities.

There is an Intelligent Tutor System for the area of mathematics called Scooter the Tutor [19], which teaches students to solve scatterplots and assists them with the necessary help and feedback so they can understand the subject and continue to solve exercises. This Intelligent Tutor System will be taken to include a collaborative model that helps secondary school student's work collaboratively in their math activities to benefit their results.

This Intelligent Scooter Tutor System is a desktop system tested on Windows 95 to Windows 8 operating systems, however, it is being migrated to a web system to be compatible with any browser and operating system, in order to students can use the system in the school, or remotely from your personal computer or mobile device.

5 Methodology

In the methodological process to find which elements or variables have an important degree of significance for the elaboration of a collaborative model, several tests were applied to a group of students, such as the collaboration test which identifies in five subscales (organization, participation, dialogue, role and responsibility) [15], the degree of collaboration of the students after carrying out a joint activity and the Hofstede cultural dimensions test adapted for educational situations that identifies the influence of the culture in students in the secondary school No. 2 "Julio Zarate" in Xalapa, Veracruz, México, in relation to the power distance index (PDI) towards their teachers, uncertainty avoidance index (UAI) in a collaborative activity, individualism (IDV) versus collectivism, masculinity (MAS) versus femininity and long term orientation (LTO).

5.1 Study units

The subjects involved in the development of this project were 116 morning hours students constituted in five school groups 1, 2, 3, 4 and 5 of the first grade (equivalent to seventh grade in the United States) of the General Secondary School No. 2 "Julio Zarate" located in the city of Xalapa, Veracruz, Mexico.

5.2 Procedure

The study was carried out in four days during the 50-minute math class in the media classroom, this is a computer lab used by teachers and secondary students, the classroom has capacity for 50 students at the same time and it consists of 34 computer equipment available with Windows operating system.

On the first day of interaction was the thematic induction, in this case scatter plot in a time of 10 minutes, later a standardized pre-test was done to know what the student's initial knowledge was, this test was done in a time of 20 minutes, a learning styles Kolb test [20] was applied in a time of 15 minutes, this test was applied because in the first study it was found that the best way to associate students in a collaborative activity is grouping them according to the same learning styles, this association allows students to obtain higher learning gains, than if students with different learning styles will join in the activity. The participation of the students on this day was individually. Once the learning style tests were taken, they were evaluated by the researcher for the conformation of the work couples of the following day.

For the second and third day, with the Intelligent Tutoring System, the interaction was done in a collaborative way by students pairs previously defined, this was done in a time of 40 minutes.

On the fourth day of interaction, the standardized test (post-test) was carried out in a time of 20 minutes, then the test of collaboration to answer it in 10 minutes and the last the test of cultural dimensions in 15 minutes. The collaboration test was applied in order to know the type of collaboration that existed between students. The cultural dimensions' test to know if any dimension affected or not, the performance of students during their collaborative activity.

The activities and execution times of this study can be seen in table 1.

Table 1. Activities and execution times of the exploratory study.

No.	Activity	Day	Execution time in minutes
1	Induction scatter plots	1	10
2	Pre-test		20
3	Learning styles test application		15
4	Work teams formation		---
5	Collaborative activity with the STI Scooter	2	40
6	Collaborative activity with the STI Scooter	3	40
7	Post-test	4	20
8	Collaboration test application		10
9	Cultural dimensions test application		15

6 Results

The tests carried out during the experimental scheme were, the pre-test to know the initial student's knowledge in the scatterplot topic, the test of learning styles, so that the students could be put together in pairs according to their same learning styles, the test of collaboration to know the type of collaboration (organization, participation, dialogue, role and responsibility) that existed during the activity, the test of cultural dimensions to know if any dimension affected or not, the student's performance during your collaborative activity. As for the analysis performed in the tests that were applied in the groups in the experiment, it was observed that there is no significant difference (*p-value=0.0866*) between the groups initially, presenting an equal knowledge in the pre-test, another aspect that was shown is that there is no an association between the learning styles and the groups evaluated (*p-value>0.05*), as well as the relationship between learning styles and the five sub-scales of collaboration measured during the interaction with the tutor. However, in the post-test it is identified that there is a significant difference between the groups (*p-value=0.02439*) as shown in Figure 1.

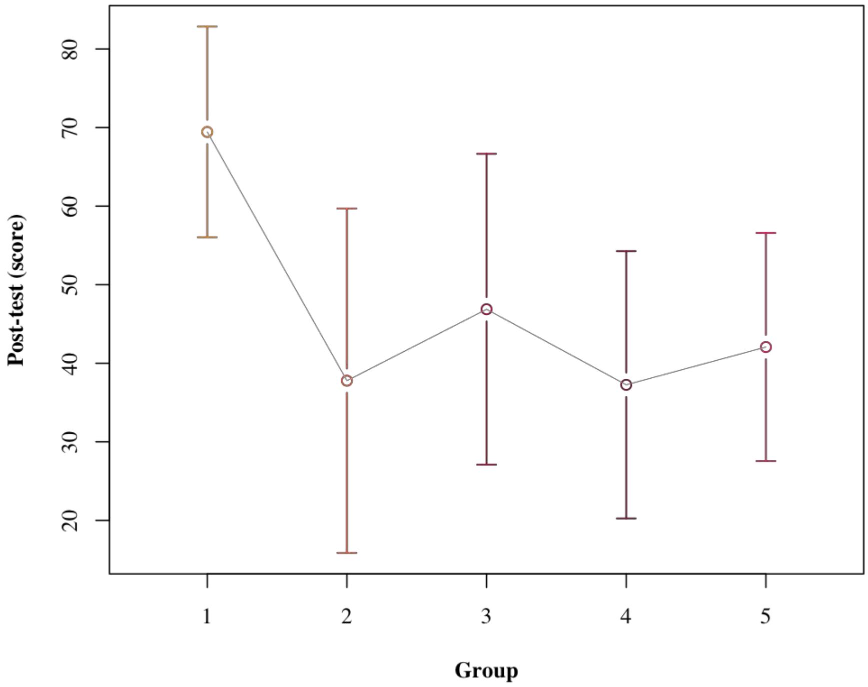


Fig. 1. Result of the post-test of groups.

In the analysis individually for each of the groups, it was found that the variables of both collaboration and cultural dimensions in some of its elements are related, that is, some behaviors are distinguished that do not occur naturally by themselves, but they are added with other characteristics, in this sense the collaboration is directly linked with characteristics of cultural dimensions or vice versa, this in benefit of the improvement of the result in the post-test of the students.

Of the five sub-scales, organization, participation, dialogue, role and responsibility evaluated in the collaboration test, and the five cultural dimensions defined in the Hofstede test, the power distance index 'PDI', uncertainty avoidance index 'UAI', individualism 'IDV' versus collectivism, masculinity 'MAS' versus femininity and long term orientation 'LTO' there was mostly an association between them in a particular way for each group.

In group 2 (G2) the relationship between UAI and Responsibility was observed with a value of *p-value=0.0498*, MAS with Participation (*p-value=0.0497*), as well as LTO with the same dimension of collaboration Participation (*p-value=0.0036*), in addition to MAS and role (*p-value=0.0024*). In group 3 (G3) the relationship between UAI and Organization was observed (*p-value=0.0307*). Group 4 (G4) showed relationship in UAI with Organization (*p-value=0.0102*), MAS and LTO with Responsibility with values of *p-value=0.0439* and *p-value=0.0001* respectively. On the other hand, group 5 (G5) only showed a relation of IDV with Conversation (*p-value=0.0054*). Group 1 (G1) did not present any relationship between cultural dimensions and collaboration sub-scales. You can see these results in table 2.

Table 2. Results of relationship of cultural dimensions and collaboration subscales.

	PDI	UAI	IDV	MAS	LTO
Organization		0.0307 (G3) 0.0102 (G4)			
Participation				0.0497 (G2)	0.0036 (G2)
Conversation			0.0054 (G5)		
Role				0.0024 (G2)	
Responsibility		0.0498 (G2)		0.0439 (G4)	0.0001 (G4)

Table 2 shows that the union of both elements, cultural dimensions and collaboration are present in the behavior of the groups, however, by themselves, they do not show any type of behavior, which indicates that both characteristics must be associated for obtaining better results.

With the results that are observed of the relationship between some cultural dimensions and some collaborative elements, the intelligent tutoring system to which the model going to include, should mediate this type of aspects. For example, if it is observed that the lack of responsibility is linked to the high student's uncertainty to work in a collaborative activity, then, we should include in the intelligent tutoring system, an element that explains more in detail, how to solve the exercise, with the goal to eradicate the student's uncertainty when they doing the activity. In this way, we would seek to eliminate or reduce the uncertainty so that the student is responsible in the development of their activity. Just as the system would be modified in this relationship, modifications would also be made for the other relationships between cultural dimensions and collaborative elements.

7 Conclusions and future work

It was observed that the group is a factor that affects the post-test, the learning style is an element that affects learning independently, that is, it is not linked to any cultural dimension or to any collaborative elements, and last, that the union of the collaborative and cultural elements must be associated to obtain better results.

As future works are the integration of variables for the formal definition of the collaborative model, considering the multiple linear regression approach to study the relationship between the variables of interest, to calculate the response variable through the estimation of the best linear predictor, in this case would be the post-test. Also the inclusion of it in a mathematics Intelligent Tutoring System and the evaluation of the model to check the predictions of it. All this will be done so that students can work collaboratively with an Intelligent Tutoring System to help them get better results in their math assessments.

An example of how it would be the inclusion of the model in the Intelligent Tutoring System is if the model predicts that the student would have a greater post-test if the student when doing a collaborative activity, will talk more with his classmate, then the Intelligent Tutoring System will have to include elements such as a forum, a chat, an editor, or any aspect that promote conversation in the collaborative activity. In this way, all the elements indicated by the collaborative model needed to improve the student's post-test would be added to the system.

Acknowledgments. To National Council of Science and Technology (CONACYT) by the doctoral scholarship No. 586959 granted to carry out PhD in computer science.

References

1. Hofstede, G. *Culture's consequences: Comparing values, behavior, Institutions and Organizations across Nations.* SAGE publications, 616 pages, (2003)
2. Walker, E.; Rummel, N.; Koedinger, K.: Integrating collaboration and intelligent tutoring data in evaluation of a reciprocal peer tutoring environment. *Research and Practice in Technology Enhanced Learning*, Vol. 4, No. 3, pp. 221–251. doi: 10.1142/S179320680900074X (2009)
3. Rodríguez, R.; Castillo, J.; Lira, A.: Diseño de un sistema tutor inteligente. *Apertura*, Vol. 5, No. 1, pp. 36–47 (2013)
4. Suebnukarn, S.; Haddawy, P.: COMET: A Collaborative Intelligent Tutoring System for Medical Problem-Based Learning. *IEEE Intelligent System*, Vol. 22, No. 4, pp. 14–21. doi: 10.1109/MIS.2007.66 (2004)
5. García, J.; González, S.; López, A.: Tutor inteligente para propuesta de investigación. *Conciencia tecnológica*, No. 47, pp. 43–48 (2014)
6. Lara, E.M.; Reyna, R.: Tecnología al alcance de los docentes. *Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, No. 10, pp. 1–18 (2013)
7. Lara, E.M; Rebolledo, G.; Rojano, J.R.: The influence of learning styles in collaborative activities. *Proceedings of the XVIII International Conference on Human Computer Interaction (Interacción '17)*, Article 33, 6 pages. doi:10.1145/3123818.3123843 (2017)
8. Brusilovsky, P; Peylo, C. Adaptive and intelligent web-based educational systems. *International Journal of Artificial Intelligence in Education (IJAIED)*, 13, 159–172 (2003)
9. Craig, S. D; Driscoll, D. M; Gholson, B. Constructing knowledge from dialog in an intelligent tutoring system: Interactive learning, vicarious learning, and pedagogical agents. *Journal of Educational Multimedia and Hypermedia*, 13(2), 163–183 (2004)

10. Martínez, B.: El aprendizaje de la cultura y la cultura de aprender. *Convergencia*, Vol. 15, No. 48, pp. 287–307 (2008)

11. OCDE.: PISA 2015 - Resultados. https://www.oecd.org/pisa/PISA-2015-Mexico-ESP.pdf (2005). Accedido el 8 de Enero de 2018

12. Maldonado, P. M.: El trabajo colaborativo en el aula universitaria. *Laurus*, Vol. 13, No. 23, pp. 263–278 (2007)

13. Vygotsky, L. (2000). El desarrollo de los procesos psicológicos superiores. Barcelona, España: Grijalbo.

14. Isotani, S.; Bourdeau, J.; Mizoguchi, R.; Chen, W.; Wasson, B.; Jovanovic, J.: Guest Editorial: Special Issue on Intelligent and Innovative Support Systems for CSCL. *IEEE Transactions on Learning Technologies*, Vol. 4, No. 1, pp. 1–4 (2011)

15. Huerta, N.: Aplicación de Métodos Biplot Clásicos, uso práctico en la educación con tecnología (reporte de pregrado). Universidad de Salamanca, Salamanca, España (2014)

16. Peñaloza, M.G.: Estudio cualitativo de la influencia de las dimensiones de Hofstede en el aprendizaje con objetos de aprendizaje para algoritmos (tesis de pregrado). Universidad Veracruzana, Xalapa de Enríquez, Veracruz (2011)

17. Zhao, R.N.; Bourdeau, J.: Building a cultural intelligence decision support system with soft-computing. International Journal on Advances in Intelligent Systems, Vol. 6 No. 1-2, pp. 136–150 (2013)

18. Ogan, A.; Walker, E.; Ryan, B.; Rebolledo, G.; Jiménez, M.; Laurentino, T.; Carvalho, A.: Collaboration in Cognitive Tutor Use in Latin America: Field Study and Design Recommendations. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*, pp. 1381–1390, doi:10.1145/2207676.2208597 (2012) 15.

19. Baker, R.: Designing Intelligent Tutors that adapt to when student game the system. *Doctoral thesis*, Carnegie Mellon University, Pittsburgh, U. S. (2005)

20. Kolb, D. *Learning styles and disciplinary differences*. San Francisco, California, Estados Unidos: Editorial W. Chickering (1981)

Learning analytics workshop: Building bridges between the Education and the Computing communities

Sébastien Béland [1], Michel C. Desmarais [2], and Nathalie Loye [1]

[1] Université de Montréal, [2] Polytechnique Montréal, Canada

Learning analytics workshop: Building bridges between the Education and the Computing communities

Sébastien Béland[1], Michel C. Desmarais[2], and Nathalie Loye[1]

[1]Université de Montréal, [2]Polytechnique Montreal, Canada

The Learning Analytics (LA) and Educational Data Mining (EDM) fields have generated a wealth of research over the last decade, including two yearly conferences and two scientific journals. However, these topics are relatively new in the field of educational science. This workshop brings together researchers and practitioners to share their perspective on how this research has impacted the education field.

Labarthe, Luengo, and Bouchet reports on the very topics that Educational Data Mining and Learning Analytics have addressed in the last decade. Through the analysis of papers from tens of conferences and journals, they reveal the main research trends of each field and show their similarities and differences.

Two other workshop papers describe practical applications of LA techniques over typical problems faced by educational practitioners. Xu, Chen, and Wu describe the results of a Neural Network approach to predict honor student grades from a wide diversity of factors, ranging from Internet usage to past grades. They show that the Neural network technique can achieve substantially better accuracy than more traditional linear regression methods, giving weight to the advantages of machine learning techniques over standard statistical techniques. Desmarais addresses the problem of selecting candidates for limited admission programs, when candidate sources have no common grading schemes. He shows that, using statistical distribution assumptions combined with an optimization technique and historical scores from the host institution, the proposed approach can improve the expected score of accepted students by about one third standard deviation.

The workshop also hosted demos of innovative Learning Analytics tools to help college administrators in their tasks of reporting performance indicators and provide insights to guide the creation, refinement and evolution of study programs, and presentations of educational games analytics.

A Performance Predictor for Honors Students Based On Elman Neural Network

Moke Xu
Shenyuan Honors College,
Beihang University
China
08753@buaa.edu.cn

Xuyang Chen
Shenyuan Honors College,
Beihang University
China
freeverc@gmail.com

Wenjun Wu
NLSDE Department of
Computer Science and
Engineering, Beihang University
China
wwj@nlsde.buaa.edu.cn

Abstract.
With the popularization of higher education, honors education has become an important work of research-oriented universities to cultivate excellent students. In order to evaluate the achievements of honors education and to make a guidance for honor educators, it is necessary to predict the performance of honors students effectively. This paper proposes a data-driven model to make predictions on students' performances based on an adjusted Elman Neural Network (Elman NN). Moreover, to be more significant, we made a comparison between Elman NN and some other methods. The result shows that our model performs much better. The performance predictor may provide a reference for honor educators in the professional choices and enable them to provide appropriate suggestions or motivations for those of the honors students who are at an early stage of learning risk or have a potential of an out-standing talent.

Keywords. Elman Neural Network, Data Mining, Predictive Model， Regression， Classification

1 INTRODUCTION

In recent years, honors programs of higher education have become available among the world famous universities and are widely recognized in American. Among the top 100 universities in the world, which have abundant education resources and small scales, more than 40% of the universities have honors programs providing honors students with challenging courses and high-level scientific research training opportunities. Following the success of their honors programs, top-notch universities in China also adopt honors programs to cultivate elite students.

Our experience of running the honors program in Beihang University for more than a decade reveals that there are two challenges for the student advisors. First of all, it's necessary to find out how to help honors students to choose majors which suit them best. Usually, the students have to choose their majors depending on their willingness and ability after the first school year. It is ideal to give students essential guidance in developing their interest and talent in the most suitable majors for them in the first year. But given the limited manpower of our honors program administers, we need a predictive tool to efficiently assess student ability and predict their future performance. Secondly, every year we have to manually identify the honors students as risk and give them counselling to overcome their academic difficulties and even adjust their negative timing habits in daily life. A powerful predictive model is also very important to help us in fulfilling this responsibility through necessary learning suggestions and teaching interventions.

In this paper, we adopt an Elman Neural Network as a modeling framework to implement the predictive model, which has been widely used in predictive problems in various fields. We redesign weighted context units in the hidden layer of Elman NN to reflect the latent interaction among honors students in the same year. Based on this improvement on the original Elman NN, we establish a predictive model to fit performance of honors students and verify the effectiveness of the model by the actual datasets. Experiments show that our model outperforms some other regular models.

The rest of the paper is organized as follows. Section 2 presents a summary of related work and a brief comparison to our model. Section 3 presents the dataset descriptions and processing details. Section 4 introduces our predictive model and related experimental results. Finally, Section 5 concludes the paper.

2 RELATED WORK

Prediction of student scores is an important research topic in the field of educational data mining. Many researchers have proposed predictive models based on a variety of machine learning techniques. Jie Xu, et al (2017) [1] developed a novel algorithm that enables progressive prediction of students' performance by adapting ensemble learning techniques and utilizing education-specific domain knowledge. It is proved that its prediction results are accurate enough compared to some other methods. Elbadrawy et al. (2016) [2], proposed a predictive model based on regression-based and matrix factorization–based methods to predict student performance. Dekker et al. (2009) [3], presented a case study to evaluate multiple drop-out prediction models.

All these previous efforts only focus on predicting future performance based on student current status and past academic performance without considering behavior features that are not directly related to their course study. They often rely upon Learning Management Systems on campus to collect study records as training datasets for developing their models. Such an approach has inherent limitation because it cannot capture students' daily activities that may have great impact on their study. Especially for the honors students at their first campus year, life style can bring negative influence on their study. To incorporate these factors into our predictive model, we decide to enrich the feature space of our model by introducing student daily activity features including consumption in campus cafeteria, Internet accessing at different time frames and library book-lending transactions. These data are collected from multiple e-campus service systems and assimilated into our training dataset. Given the temporal natural of honors student development and their daily activity data, we choose to adopt a simplified recurrent neural network, which was called Elman Neural Network [5], to build our predictive model.

3 DATA DESCRIPTION AND PROCESSING

3.1 Dataset Description

For honors students in Honors College, the design of honors project follows the principle of a solid foundation and gradual improvement. In the first year, students will learn basic subjects as a basis and preparation for further professional learning. In the following years, students will be major-oriented educated and learn more professional courses. As the knowledge basis of the first academic year is very important, we hope to predict the performance of the first school year in the first semester.

In this paper, we will establish a data-driven predictive model based on the data of students in grade 2015 and grade 2016, including their initial grades, learning and daily behaviors in the first semester, and we'll predict the performance of their core subjects and comprehensive scores. The input dataset contains 501 vectors, of which 205 are from students in Grade 2015 and 296 from students in Grade 2016. Every vector contains a 54-dimensional input vector and a 9-dimensional output vector.

After entering the University, many students will indulge in computer games resulting in reduced learning time. As students' internet access is a major factor affecting their academic performance, we collected students' internet accessing details including total length of Internet time, active periods, traffic, etc. And we organize Internet time, traffic data by month (X25-X54) and active periods by 6-hour periods (X21-X24). The college entrance examination scores represent the students' initial knowledge level and learning ability (X3-X7). The First midterm examination in college comes two months after enrolment, which indicates students' adaptability to university studies to some content. Moreover, we assume that students' monthly consumption, book-borrowing numbers and birth dates will also influence their final results. The initial CEE data, book-borrowing data, consumption and internet-accessing data can be collected easily through multiple e-campus service systems.

The 9-dimensional output data includes a 3-dimensional part of the comprehensive performance and a 6-dimensional part of performances in core courses. The consolidated performance part includes consolidated performance, the average grade of main courses and credit scores. For honors students, the consolidated performance is related to the latter two parameters by the Eq (1) as follows:

$$Y_1 = 0.6 \times \frac{Y_2}{Y_{2max}} + 0.4 \times \frac{Y_3}{Y_{3max}} \tag{1}$$

Y_2max means the maximum value of average performance of main courses for all students in the same grade. *Y_3max* means the maximum value of credit scores for all students in the same grade. The equation indicates the importance of core courses for honors students. The core subject grades section contains 6 elements, corresponding to their performances of the six core courses. These subjects are set especially for honors students in honors project, thus they can measure students' mathematical ability, experimental ability, programming ability, language ability properly, which are representative enough in measuring students' ability distributions.

The details of the input data and output data are shown in table 1. Xi means input data and Yi means output data.

Table 1. List of input data (Xi) and output data (Yi)

Symbol	Meaning	Symbol	Meaning
X1	Birth year	X25-X34	Internet-accessing time
X2	Birth month	X35-X44	Internet-downloading traffic
X3	Total score of CEE	X45-X54	Internet-uploading traffic
X4	Chinese perf. in CEE	Y1	Consolidated perf.
X5	Math perf. in CEE	Y2	Average grade of main courses
X6	Science perf. in CEE	Y3	Credit scores
X7	English perf. in CEE	Y4	Mathematics perf.
X8	Math perf. in FMEC	Y5	Basic Physics perf.
X9	Programming perf. in FMEC	Y6	General Chemistry perf.
X10	Number of books borrowed	Y7	Basic Life Sciences
X11-X20	Monthly consumption	Y8	Advanced Programming perf.
X21-X24	Internet total traffic by 6-hour periods	Y9	College English perf.

CEE: College Entrance Examination (perf. Means performance)
FMEC: The First Midterm Examination in College

3.2 Normalization

To reduce the amount of calculation and speed up the model training process, it is significant to normalize the input data before training. Min-Max Normalization, also known as dispersion standardization, is a linear transformation of the raw data so that the resulting values map between [0 - 1]. It is an effective normalization method.

The normalization equation is as follows. x* is the normalized value and x the initial value. Max and min means the maximum value and minimum value in all items.

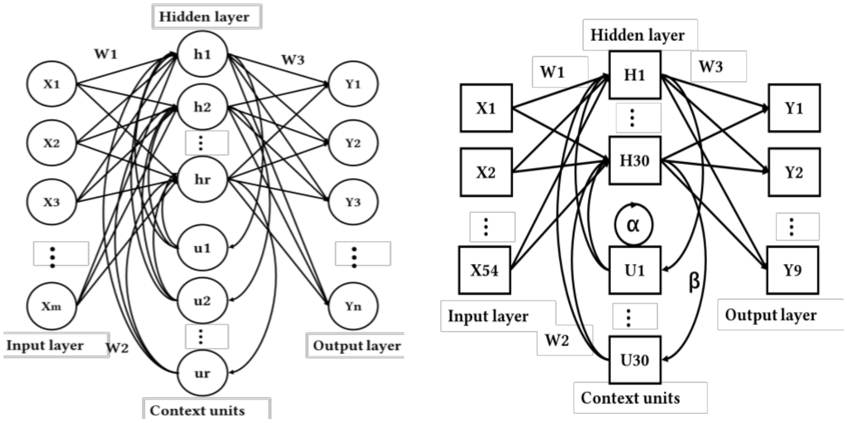
$$x^* = \frac{x - \min}{\max - \min} \qquad (2)$$

4 MODEL AND RESULTS

4.1 Principles of Elman NN and Our Adjustment

An Elman neural network is a three-layer network with the addition of a set of "context units" used to remember the output value of the hidden layer units, which can be considered as a delay operator. The hidden layer is connected to these context units fixed with a weight of one initially. At each training step, the input will propagate over the feed-forward part during which a learning rule is applied. The back-connections part is fixed and save a copy of the values of the hidden units in the context units. The saved values will propagate over the connections before the learning rule is applied. That means the network can take into account the internal relationship among input data.

Fig. 1. The structure of basic Elman Neural Network and our model.



For honors students, the performance of every student might be influenced not only by his or her initial scores and daily behaviors, but also by the behaviors of other students. That's why we choose an Elman neural network which can take the inter-influence factors into consideration. The training sequences of the input data in all training epochs are set randomly, making it more reasonable to take mutual influence into account. Suppose that there are m nodes for the input layer, n nodes for the output layer, and r nodes for the hidden layer. Thus there will be r context units. In our model, m=54, n=9, r=30. The structure of the initial Elman NN and our adjusted model are shown in Fig. 1(right).

In traditional Elman neural network, the weight from context units is to the input layers are set as ones. But in fact, the recurrent part will not play an important role as the parameters from the input layer for the hidden layer. The weight shall be adjusted to correspond well to the application in predicting performance. Also, to make it more significant, we assume that the influence from previous two steps should not be neglected. The equations are shown as follows:

$$\mathrm{H(k)} = \mathrm{f}(\ W^1\mathrm{X(k)} + W^2\mathrm{U(k)}) \tag{3}$$

$$\mathrm{U(k)} = \alpha\ \mathrm{H(k-1)} + \beta\ \mathrm{U(k-1))} \tag{4}$$

$$\mathrm{Y(k)} = W^3\mathrm{H(k)} \tag{5}$$

X(k) is the input value from the input layer. H(k)is the output value of the hidden layer. Y(k)is the output value of the output layer. f(x) is the activation function. It is always set as the sigmoid function. W_1 is the weight matrix between the input layer and the hidden layer W_2 is the weight matrix between the context units and the hidden layer. W_3 is the weight matrix between the hidden layer and the output layer. α is the feedback gain parameter for self-connection. β is the feedback gain parameter for the previous self-connection part.

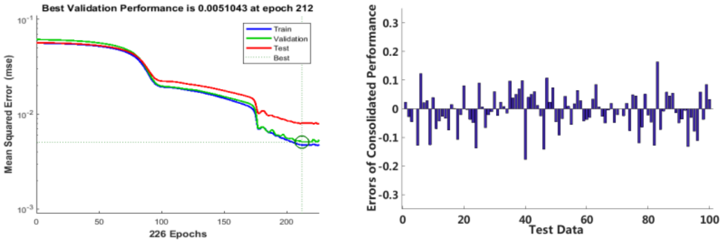
In our model, α and β should be set to a small value. Based on enough experiments, we found that the model performs well when we set α=β=0.05. When the values changes in a small range (0.02-0.2), the final results won't change a lot. That means the model is stable in such parameters.

4.2 Training and Tests

In training part, we set 401 train and validation data and 100 test data. As the predictive model is a regression problem actually, we set loss function (R(y,y*)) as mean square error (MSE) function, which always performs well in regression problems. The equation is as follows.

$$R(y, \hat{y}) = E_y(y - \hat{y})^2 \tag{6}$$

Fig. 2. The dynamic magnetization of training and the errors of consolidated performance



The process of model training and the results are recorded and shown in Fig. 3. The training function is chosen as gradient descent with momentum and adaptive learning rate backpropagation, which works well in Elman neural network according to a lot of experiments. We set learning rate as 0.05. There are 1000 training steps in each epoch. After 212 epochs, the model reached a stable point.

The output value are decimals ranging from 0 to 1 (representing scores ranging from 0 to 100). We calculated the errors of test data and present the errors of consolidated performance above. According the Figure 4, most errors of predicted results are no more than 0.1, which means our results are credible enough.

4.3 Comparisons

To be more significant, we made a comparison among Elman NN , BPNN (Backpropagation neural network), the most frequently used neural network model, and linear model. To ensure that the compared network is in the same size and scale, the BPNN is also arranged by three layers, including a 54-node input layer, a 30-node hidden layer, and a 9-node output layer. The training methods are all set similarly.

To evaluate the two methods properly, we calculated the confidence rate of both outputs based on the confidence interval of 10%.

$$\text{confidence rate (i)} = \frac{m_i}{n} \ , (i = 1,2, \dots ,9) \tag{7}$$

In this formula, n means number of items in test data, mi means number of credible items in test data. A tested item is treated as a credible one if:

$$\frac{|Y^*{}_i - Y_i|}{Y_i} < 10\% \ , \ (i = 1,2, \dots ,9) \tag{8}$$

Y* means the predicted result and Yi means the actual results, which is also the labeled value. The confidence rate in output data of both two methods is shown in Table 2. Obviously, the credible rate of our model is better. That means it is credible enough to predict student performance based on our model. Also, the prediction about average grade of core courses is the most accurate, which is also the most valuable parameter in measuring student learning ability.

Table 2. The comparison among our model, the BPNN model and linear model

Symbol	Meaning	CR of ENN	CR of BPNN	CR of LM
Y1	Consolidated performance	89%	86%	69%
Y2	Average grade of Core Courses	91%	88%	81%
Y3	Credit scores	82%	79%	73%
Y4	Mathematics performance	86%	84%	62%
Y5	Basic Physics performance	79%	72%	58%
Y6	General Chemistry performance	83%	85%	71%
Y7	Basic Life Sciences	88%	79%	65%
Y8	Advanced Programming perform	74%	66%	54%
Y9	College English performance	85%	69%	68%
Average Value		**84%**	**79%**	**67%**

The prediction confidence rate of advanced language programming performance is obviously lower than the others, for the uncertainty of the course. On the whole, most output items can be predicted accurately and we can trust the results at a low risk of making mistakes.

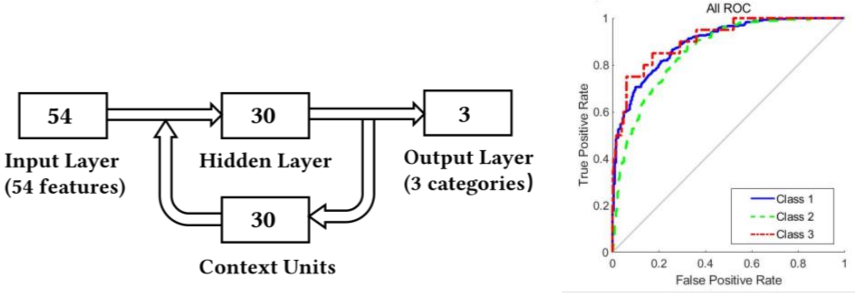
4.4 Classifications

In the classification model, we divide the honors students into three categories according to their consolidated performance, which respectively represent excellent, good and general level. The number of students in each category and the results are presented in Table 3. The structure of the model and the receiver operating characteristic curve (ROC curve) of test data are shown in Fig. 3.

Table 3. The number of students in each category and classification results

Symbol	Meaning	The whole number	The test data number	Correct Items	Correct Rate
C1	85-100	299	64	56	81.4%
C2	70-85	182	32	20	75.8%
C3	<70	20	4	3	75%

Fig. 3. The structure of classification model and the ROC curve



The correct rate of the student category prediction (shown in Table 4) is 77.4%. According to the ROC curve, the classification accuracy of the three categories is at the same level. It is obvious that we can also get good results through the classification model based on Elman NN.

4.5 Application of Models

As the prediction model can predict students' grades and categories at an early stage, risk students can be identified half a year in advance. In application, counselors will combine the results of classification and regression models. Firstly, they will collect data required and input the data into the model, which is an automated process wasting less time. After that, they can easily identify risk students based on the classification results. To know more details, they can consult the regression model to know the ability distribution details of those risk students according to the 9-dimension outputs. Finally, they shall offer some necessary suggestions.

In the early stage, counselors were unable to get sufficient information about students' learning status. Thus it is difficult to assess students’ performance manually. But the models offer predictions based on students’ data that are easy to get. The results with accuracy of 77.4% are valuable enough for honors educators to assess the learning level of every student. It is a convenient early-stage performance predicting tool on campus.

5 CONCLUSIONS

In summary, we have performed the process of the establishment of our predictive model and presented the results of prediction of students’ final performance and ability distribution based on data of 501 honors students. By adjusting the values of feedback gain parameters for self-connection in Elman neural network and training the network reasonably, the predictive model works better compared to BPNN and linear regression method. According to our experiments, the consolidated performance and average grade of core courses in output value can be predicted most accurately. It is convenient for honors program counselors to predict students’ performance and provide appropriate suggestions or motivations for different student categories in performance.

REFERENCES

[1] Xu, J., Han, Y., Marcu, D., van der Schaar M.: Progressive Prediction of Student Performance in College Programs. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–10 (2017).

[2] Elbadrawy A., Polyzou, A., Ren, Z., Sweeney, M., Karypis, G., & Rangwala, H.: Predicting Student Performance Using Personalized Analytics. Computer, 49(4), 61-69, (2016).

[3] G. W. Dekker M. Pechenizkiy J. M. Vleeshouwers: Predicting students drop out: A case study. Proc. Int. Conf. Educ. Data Mining pp. 41-50 (2009).

[4] AI-Radaideh, Q, AI-Shawakfa, E., AI-Najjar, M: Mining student data using decision trees. International Arab Conference on Information Technology, Yarmouk University, Jordan, (2006).

[5] Elman, J. L. Finding Structure in Time. Cognitive Science, 14: 179–211. doi:10.1207/s15516709cog1402_1.(1990).

[6] Liu H., Tian H., Liang X., Li Y.: Wind speed forecasting approach using secondary decomposition algorithm and Elman neural networks. https://doi.org/10.1016/j.apenergy.2015.08.014 (2015).

[7] Xing W., Guo R., Petakovic, E., Goggins S. Participation-based student final performance prediction model through interpretable Genetic Programming, Integrating learning analytics, educational data mining and theory, (2014).

[8] Ahmed A. B. E. D., Elaraby I. S.: Data Mining: A prediction for Student's Performance Using Classification Method. World Journal of Computer Application and Technology, vol. 2, no. 2, 43 – 47 (2014)

[9] Kukk V.: Course Implementation: Value-Added Mode. In: Uden L., Liberona D., Feldmann B. (eds) Learning Technology for Education in Cloud – The Changing Face of Education. LTEC 2016. Communications in Computer and Information Science, vol 620. Springer, Cham (2016)

[10] F. Okubo, T. Yamashita, A. Shimada, H. Ogata. A neural network approach for students' performance prediction Proceeding LAK '17 Proceedings of the Seventh International Learning Analytics & Knowledge Conference, 598-599, (2017).

[11] R. Alkhasawneh and R. H. Hargraves,: Developing a hybrid model to predict student first year retention in STEM disciplines using machine learning techniques, Journal of STEM Education: Innovations and Research, vol. 15, no. 3, 35-42, (2014).

Analyzing the relationships between learning analytics, educational data mining and AI for education

Hugues Labarthe[1,2], Vanda Luengo[2] and François Bouchet[2]

[1] Incubateur Académique, Rectorat de Créteil, 94700 Créteil, France
[2] Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6, 75005 Paris, France
hugues.labarthe@ac-creteil.fr, vanda.luengo@lip6.fr, francois.bouchet@lip6.fr

Abstract. Baker and Siemens have well explained the theoretical differences and similarities between the educational data mining (EDM) and learning analytics (LA) communities in their 2012 seminal paper, in which they also wished for bridging the gap between both communities. Moreover, since its creation as an independent conference in 2009, EDM has been evolving in parallel with the intelligent tutoring systems (ITS) / artificial intelligence for education (AIED) community. But what are the actual links that exist between these three communities in terms of members and research topics: to what extent do they overlap and work together? Are they getting closer from each other or drifting apart? Is each community specific to researchers with different backgrounds, modeling and analysis techniques? Those are some of the questions we investigate using a quantitative analysis led between 2007 and 2017 through: a social network analysis of the 3 communities, involving the 1822 scientists who participated in program committees and/or appeared as authors of the associated journals (IJAIED, JEDM and JLA); and a text analysis of abstracts of articles published in these journals. Results reveal the clear differences between these communities, their topics, practices and research methods.

Keywords: learning analytics, educational data mining, artificial intelligence in education, social network analysis, text analysis, communities

1. Introduction

At the beginning of the 2010s, two communities progressively structured themselves to study learning data: the *Society for Learning Analytics Research* (SoLAR) and the *International Educational Data Mining Society* (IEDMS). In the meantime, the International AIED Society, gathered around the encompassing “Artificial Intelligence for EDucation” (AIED) theme, also started to analyze more and more data coming for their systems (in particular, intelligent tutors). Thus, three research communities have been tackling similar issues, and there has now been enough history for a data-based approach (valued by all three communities) to examine what distinguishes them and what brings them together.

The theme of *Educational Data Mining* first appeared during the ITS (*Intelligent Tutor Systems*) conference in Montreal in 2000 [1]. But it is really in 2005, with the first workshop on EDM held in Pittsburgh in conjunction with the AAAI (*Association* *for the Advancement of Artificial Intelligence*) conference that the theme started to take off. Most of the research work presented at that time were led on data coming from ITS [2]. The first state of the art work was published in 2007 by Romero et Ventura [3], and was followed by the creation of the yearly EDM conference in 2008, and of its associated journal, the *Journal of Educational Data Mining* (JEDM), in 2009. In parallel, and independently, the *Society for Learning Analytics Research* (SoLAR) was founded in 2011 with its associated yearly conference, LAK (*Learning Analytics and Knowledge*), followed in 2014 by its own journal, the *Journal of Learning Analytics* (JLA). Finally, the AIED community has been structured for three decades around two alternating bi-yearly conferences, AIED (*Artificial Intelligence for Education*), which became yearly in 2017, and ITS (*Intelligent Tutoring Systems*), as well as a journal, IJAIED (*International Journal of Artificial Intelligence for Education*).

Very early on, the two new communities have acknowledged each other and the differences that exist between them, mainly in the background of its lead members (semantic web for LA, educational software for EDM), the analysis techniques they mostly use (social network analysis for LA, more machine learning for EDM), and their overall goal (empowering learners and teachers while leaving them in charge for LA, automated adaptation by the computer for EDM). Those key differences are well summarized in [4], in which the authors also call for joining the forces of the two communities to build upon each other's strengths. Although the interactions have been happening [5], both communities have also kept their respective identities [6], which have been established through publications to federate their respective domains [7–9].

Overall, a decade after the first EDM conference, and three after the first ITS, the three communities are thriving, and we can wonder about the relationships between each other and their respective impact on education. We decided to study three types of data: (1) the reviewers for the conferences associated to each community (AIED/ITS, EDM, LAK); (2) the authors of the papers published in the journals associated to each community (IJAIED, JEDM, JLA); (3) the abstracts of the papers published in the journals associated to each community. Using these datasets, we performed exploratory analyses of the overlap of the communities as well as of their individual specificities.

2. Data collection and cleaning

For each of the aforementioned datasets, we decided to consider a period of 11 years (2007-2017), which encompasses the whole existence of the EDM community. Although it may appear to give an emphasis to the data from that community, the LA community has published overall more intensively since its birth in 2011 (*cf.* table 1 further), and we therefore believe the 4 extra years are not affecting the validity of our results. Regarding the AIED community, although we had access to older data, we believed the changes in terms of popular scientific topics and approaches over time did not justify including it, and that it made more sense to use a similar period of 11 years.

The first dataset (reviewers) was collected mostly manually by extracting the list of reviewers' names included in the proceedings of each conference. We extracted the names from PDF version of the proceedings, selecting any name listed under the "Program Committee" and "Reviewers" sections, excluding others such as "Conference chairs" or "Organization committee". The choice of reviewers instead of authors was justified by the fact that many conferences authors may appear only once, and that authoring a single paper in a conference does not necessarily imply a tight relationship with the associated community. Conversely, being invited to review papers for a conference usually indicates a sustained link (including but not limited to authorship), more relevant for a community analysis like the one we wanted to perform.

The second (authors) and third (abstracts) datasets were extracted automatically using a webcrawler tool (Scrapy) specially configured to extract from each website the information relative to published papers (title, authors, abstract, keywords, volume, issue, year). For IJAIED, information was extracted from both the Springer and ijaied.org websites, but only the ijaied.org data was kept because the Springer data started in 2013 only. We excluded from these datasets articles explicitly identified as an editorial, including guest editorials for special sections in the case of JLA, to focus only on research papers. A tedious review of names, surnames and even positions resulted in creating a single table, reducing a list of 4026 names to 1505 individuals. The abstracts were analyzed using Python packages for text analysis and visualization.

Overall, when not counting twice authors and reviewers who published/reviewed more than once for a given journal/conference, we see in Table 1 that AIED remains logically the dominant community of the three, with 687 reviewers and 386 authors. In terms of reviewers, EDM and LA are very close from each other and are far less than half of the reviewers for AIED. However, in terms of journal authors, despite a later start, the LA community has published almost 2.5 times more articles than the EDM one, with almost twice more individual authors.

Table 1. Conferences, Journals, Authors and Reviewers between 2007 and 2017

Communities	Conferences	Conf. reviewers *	J. Issues	J. Articles**	J. Authors*
AIED	11	687	11	161	386
EDM	10	238	9	54	151
LA	7	233	4	132	267
Total *	28	990	33	349	748

* Double count free ** Editorials free

3. Conference reviewers community analysis

First, we focus on the conference reviewers' dataset to analyze the evolution of the reviewers' network among the three communities from 2007 to 2017. In a decade, the number of scientists reviewing for each year conferences' papers has increased by 103%, reaching 415 reviewers in 2017, showing the significant vitality of these research fields (*cf.* Table 2). Moreover, the total number of scientists involved in these 28 conferences has increased considerably from 204 to 990 (+385%), showing that the growth in yearly reviewers came from a community more than twice larger overall. Despite a small drop in the number of yearly reviewers from 2008 to 2010, the number of scientists involved in these reviews has never stopped increasing, with two peaks: +29% in 2008 for the first EDM conference and +36% in 2011 for the first LAK conference.

Table 2. The Continuous Enlargement of the Program Committees, from 2007 to 2017

Reviewers	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
Total number	204	131	136	124	244	265	280	293	266	314	415
Cumulated	204	263	306	330	449	535	623	681	761	848	990
Annual growth %		+29	+16	+8	+36	+19	+16	+9	+12	+11	+17

Due to its anteriority in the field, we could make the hypothesis that the AIED/ITS conferences provided most of the reviewers for the two other communities. To test this hypothesis, we examined the overlap of reviewers between LAK/EDM and the AIED/ITS conferences (*cf.* Table 3). Until 2014, the AIED community has recruited two thirds of the reviewers, with 88 % of them exclusively dedicated to its Program Committee. Then, it decreases to only half of the total, and 70-75% of exclusive reviewers. It is a sign not only of the growth of the LA/EDM communities, but also of the increased porosity with the older AIED community. As we can see in Table 3, the two new communities have been relying upon this first one, at least at their beginning. These communities have progressively grown from one fifth of the network together, to one third each, with LAK having the fastest growth. The proportion of cross-conferences' reviewers for more than one conference has remained constant overall, at around 8-12%, with two peaks to 15% in 2010, and to 14-19% in 2015-2017.

Table 3. Total numbers of reviewers by and between conference

Conf.	**Reviewers**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
All	Total #	204	131	136	124	244	265	280	293	266	314	415
AIED ITS	*% Total*	*100*	*89*	*93*	*81*	*72*	*72*	*66*	*68*	*46*	*45*	*47*
	% Exclusive	*100*	*91*	*90*	*81*	*88*	*88*	*88*	*88*	*69*	*70*	*75*
EDM	*% Total*		*20*	*16*	*34*	*23*	*21*	*23*	*21*	*35*	*38*	*32*
	% Exclusive		*58*	*41*	*55*	*62*	*61*	*58*	*61*	*55*	*68*	*64*
LAK	*% Total*					*16*	*18*	*24*	*23*	*40*	*38*	*37*
	% Exclusive					*78*	*79*	*69*	*76*	*75*	*78*	*79*
Cross conf.	*Total number*		11	13	19	24	26	33	29	50	43	60
	% of Line 1		*8*	*10*	*15*	*10*	*10*	*12*	*10*	*19*	*14*	*15*

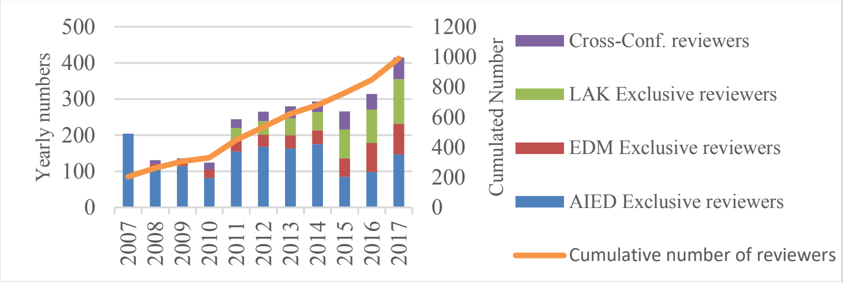
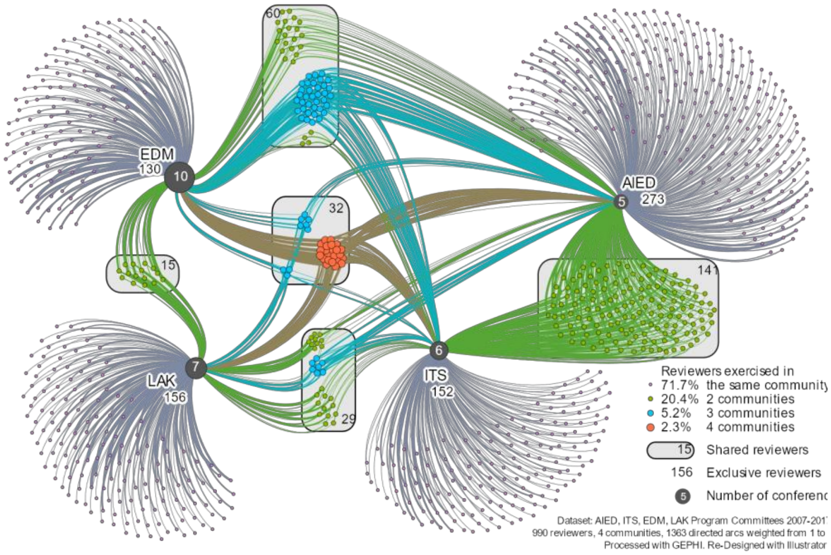


Figure 1. Evolution and distribution of the community of the reviewers for the conferences

Figure 1 illustrates the continuous growth of the overall reviewer community from 200 to almost 1000 in a decade, dominated by AIED during the first 8 years. From 2015 onwards, the number of cross-conferences reviewers has been growing too, which raises the question of knowing which communities overlap. But how many of them stayed in their original community and how many have been reviewing for more than a single conference? Overall, the 990 unique reviewers identified have been mentioned a little bit over 3000 times. Despite an average number of 3 conferences reviewed for each reviewer, 71% of them have appeared only in one community (711 nodes with outdegree=1). Figure 2 shows the number of reviewers who have been reviewers outside of their original community.



Dataset: AIED, ITS, EDM, LAK Program Committees 2007-201
990 reviewers, 4 communities, 1363 directed arcs weighted from 1 to
Processed with GEPHI. Re-Designed with Illustrator

Figure 2. The community of reviewers for each conference

AIED/ITS conferences have been sharing a quarter of all their reviewers (141 out of 425): it could come from the fact that those conferences have been alternating over the period considered (odd years for AIED and even years for ITS) – although we see that both of them also have their own subset of reviewers. But beyond this particular case, the number of persons who really belong to two or more communities remains limited: only 13.7% of the reviewers (136 individuals) cross-reviewed between, at least, two of the following communities: AIED/ITS (considered as a single one), EDM and LAK. As illustrated by Figure 2, the common core of the three communities consists of 32 reviewers. The most surprising result was to see how the LAK community was the least related to the others, when compared with the bonds between EDM, ITS and AIED. The reviewers common to each pair of community, as well as to the three communities are in Table A in Appendix, and in Table 4 for a synthesis.

Table 4. Percentage of Shared Reviewers on All Reviewers for each pair of conferences

AIED-ITS	EDM-ITS	EDM-AIED	LAK-ITS	EDM-LAK	LAK-AIED
25	20	18	14	14	10

4. Journal authors community analysis

Using the second dataset, we considered the papers published in the communities' respective journals (IJAIED, JEDM and JLA). From 2007 to 2017, there are 996 signatures corresponding to 748 unique authors of 349 articles. 80% of these unique authors signed 1 paper; 14% signed 2, and 6% signed at least 3 of them. Overall, the low number of authors of more than one paper limits this analysis, but we performed the same cross-reference analysis as in the previous section for reviewers. It reveals that a dozen of authors published in each pair of journals (*cf.* Table B in appendix), and 8 central authors published in the three of them.

5. Textual analysis of journal abstracts

Scientific communities are centered around the scientists that are part of them, but also around some common themes. To identify the themes that are characteristics of each community, we have tried to identify the keywords characteristics of the papers published in the journal of each community, using the third dataset.

First, we performed a cleaning of the abstracts using Python Natural Language Toolkit (NLTK) to perform the usual first step (tokenization, lemmatization and stop words removal). Then we used the word_cloud package to identify visually if some keywords were appearing more in some abstracts than others (*cf.* Figure 3). All communities are obviously very centered on "student", "learning" and "usages". The LA and EDM communities also share the focus on data, which is missing from the AIED community.

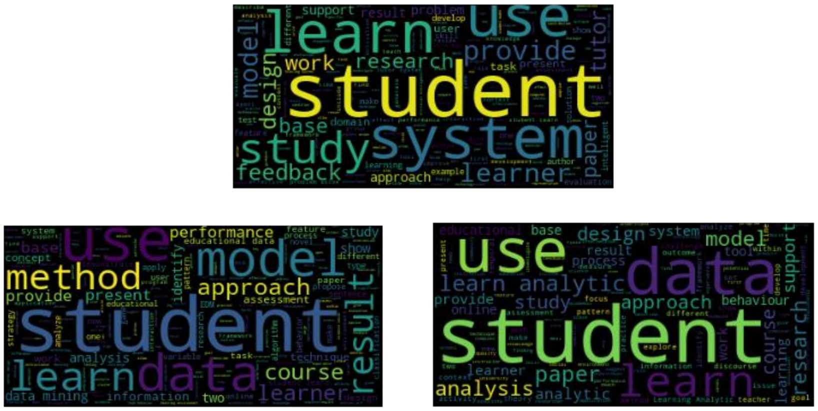


Figure 3. Word clouds for IJAIED (top), JEDM (left) and JLA (right) abstracts

However, more than the similarities between the communities, we are interested in what distinguish them from one another. To identify the keywords representative from each community, we extracted from the compilation of the abstracts of each journal the associated keywords using the Rapid Automatic Keywords Extraction (RAKE) algorithm. To avoid the fact that it may overrepresent keywords cited many times by the same article, we kept only the keywords that appeared in at least 20% of the abstracts from each journal. We obtained a set of 110 keywords appearing in at least 29 abstracts from IJAIED, 79 keywords appearing in at least 10 abstracts from JEDM, and 80 keywords appearing in at least 26 abstracts from JLA. Then we extracted (a) the keywords from JEDM not appearing in JLA, (b) the keywords from JLA not appearing in JEDM, (c) the keywords from IJAIED not appearing in JLA nor JEDM. They are summarized in Table 5. Overall, we see that the EDM community remains very anchored in a discovery approach (investigate, evidence, assess, understand, experiment…) when the LA community is more in the practice (support, inform, development, act, teach…). Although the particular techniques used in the papers do not appear with this analysis, the focus of EDM community on a more mathematical approach (features, log, class…) is visible, when compared to LA which focuses on "text", "chi square" and "ratings". As for the AIED community, its roots in tutor systems to provide feedback while modeling skills and knowledge from the student is also clearly visible.

Table 5. Keywords specific to each community based on abstracts

Journals	Keywords
JEDM but not JLA	large, propose, technique, behavior, group, compare, ability, educational data mining, improve, demonstrate, ask, investigate, evidence, problem, make, assessment, new, cover, concept, information, analyze, log, discover, apply, assess, finding, feature, class, relate, understand, collect, experiment, task, search, state, type

JLA but not JEDM	support, focus, inform, call, analytics, development, learn analytics, high, time, n, explore, chi, rater, ever, learning, age, tool, LA, go, use, act, put, analytic, text, teach, different, pre, end, lea, two, pose, relation
IJAIED only	skill, tutor, instruct, evaluation, domain, interaction, interact, era, test, line, train, know, add, view, ten, well, AI, way, feed, effective, p, prove, low, computer, ratio, art, mode, solve, evaluate, tutor system, feedback, e tutor, effect, q, knowledge, par, help, stem, late, differ, port, adapt, instruction, come

6. Conclusion

Through an analysis of the social networks of the conference reviewers and journal authors from the AIED, EDM and LA community, we have shown that Siemens and Baker's call has been heard, as more and more scientists are at the frontiers between the communities with 139 shared reviewers and 48 shared authors. The research themes however remain clearly distinct, as shown by the keywords analysis of the journal abstracts, with an emphasis on agents and tutors for AIED, automation and prediction for EDM, and visualization for LA. However, these are the different pieces of the same puzzle: enhancing learning experience through technology.

This work presents some limits: we focused on 3 important communities, but which do not represent the whole field of educational technology – extending this approach to other communities such as the "user modeling" one, or more local communities (EC-TEL in Europe) would provide a larger overview of the domain. We could also include conference authors and abstracts in our analysis, to see if more diversity of themes can be identified that way. The lack of information regarding authors' faculties for reviewers as well as for many authors did not allow us to confirm the fact that LAK is closer to education than the other communities. Finally, we have not considered the temporal aspects of the network evolution over the decade, but only the final outcome. Nonetheless, we hope that this work will contribute in structuring the communities, and encourage more scientists to follow the trend towards more interactions between them.

References

1. Gauthier, G., Frasson, C., VanLehn, K. eds: Intelligent Tutoring Systems: 5th International Conference, ITS 2000, Montreal, Canada, June 19-23, 2000 Proceedings. Springer-Verlag, Berlin Heidelberg (2000).
2. Koedinger, K.R., Corbett, A.T.: Cognitive tutors : technology bringing learning science to the classroom. In: Sawyer, R.K. (ed.) The Cambridge Handbook of the Learning Sciences. pp. 61–77. Cambridge University Press (2006).
3. Romero, C., Ventura, S.: Educational data mining: A survey from 1995 to 2005. Expert Syst. Appl. 33, 135–146 (2007).
4. Siemens, G., Baker, R.S.J. d.: Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. In: Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge. pp. 252–254. ACM, New York, NY, USA (2012).

5. Baker, R.S.J. d., Siemens, G.: Educational Data Mining and Learning Analytics. In: Sawyer, R.K. (ed.) Cambridge Handbook of the Learning Sciences. pp. 253–274. Cambridge University Press, New York, NY (2014).
6. Balacheff, N., Lund, K.: Multidisciplinarity vs. Multivocality, the Case of "Learning Analytics." In: Proceedings of the Third International Conference on Learning Analytics and Knowledge. pp. 5–13. ACM, New York, NY, USA (2013).
7. Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J. d. eds: Handbook of educational data mining. Taylor & Francis Group, Boca Raton (2011).
8. Sawyer, R.K. ed: The Cambridge handbook of the learning sciences. Cambridge University Press, New York, NY (2014).
9. Gasevic, D., Dawson, S., Mirriahi, N., Long, P.D.: Learning Analytics – A Growing Field and Community Engagement. J. Learn. Anal. 2, 1–6 (2015).

Appendix

Table A. Name of reviewers for more than one conference in 2007-2017

Communities	Reviewers
AIED – EDM: 60 shared	Agnihotri L., Aïmeur E., Aleven V., Arroyo I., Barnes T., Beck J., Biswas G., Bosch N., Boticario J. G., Champaign J., Chi M., Conati C., Cox R., Crossley S., D'Mello S., Dragon T., Dufresne A., Feng M., Forbes-Riley K., Fossati D., Goldin I., González-Brenes J., Grafsgaard J. F., Heiner C., Hicks A., Hsiao S. I-H., Hutt S., Isotani S., Keshtkar F., Kim J., Koedinger K. R., Lallé S., Larranaga M., Litman D., Liu R., Lynch C., MacLellan C., Martin B., Matsuda N., Mavrikis M., Mojarad S., Mostafavi B., Mostow J., Muldner K., Olney A., Pavlik P., Porayska-Pomsta K., Rau M. A., Ritter S., Rodrigo Ma. M. T., Rus V., San Pedro M. O. Z., Santos O. C., Shaw E., Stewart A., Wang Y., Weibelzahl S., Williams J. J., Zapata-Rivera D.
AIED – LAK: 29 shared	Allen L. K., Brooks C., Brusilovsky P., Carmichael T., Daniel B., Dascalu M., Dessus P., Dillenbourg P., Dimitrova V., Fujita N., Greer J., Hatala M., Henze N., Herder E., Hoppe H. U., Kirschner P., Lindstaedt S., Maillet K., Martinez-Maldonado R., Ogata H., Reffay C., Roll I., Sampson D., Schmidt A., Sergis S., Suthers D., Teplovs C., Zervas P., Zouaq A.
EDM – LAK: 15 shared	Alexandron G., Conde M. A., Drachsler H., Gobert J., Klamma R., Lang C., Merceron A., Monroy C., Pardo A., Pechenizkiy M., Romero C., Siemens G., Verbert K., Wolpers M., Worsley M.
AIED – EDM – LAK: 32 shared	Azevedo R., Baker R.S.J.D, Blink M., Bouchet F., Boyer K. E., Desmarais M., Eagle M., Fancsali S., Gasevic D., Graesser A. C., Heffernan N. T., Jovanovic J., Kay J., Lester J., Luengo V., Mazza R., McCalla G., McLaren B. M., Mitrovic T., Nkambou R., Paquette L., Pardos Z., Pelánek R., Pinkwart N., Reimann P., Penstein-Rosé C., Sahebi S., Snow E. L., Stamper J., Trausan-Matu S., Yacef K., Yudelson M.

Table B. Name of authors for more than one journal in 2007-2017

Journals	Authors
IJAIED & JEDM: 15 shared	Azevedo R., Boyer K. E., Chung G. K.W.K., Conati C., D'Mello S., Goldin I., Harley J. M., Koedinger K. R., Lester J., Luckin R., Miller L. D., Nugent G., Person N., Samal A., Soh L.-K.
IJAIED & JLA: 14 shared	Blair K. P., Chin D. B., Cutumisu M., Gowda S. M, Heffernan N. T, Hoppe H. U., Kay J., Linn M. C., Paquette L., Pardos Z., Rau M. A., San Pedro M. O. Z., Schwartz D. L., Segedy J. R.
JEDM & JLA: 11 shared	Bannert M., Blikstein P., Cai Z., Crossley S., Kinnebrew J. S., Kitto K., Recker M., Schneider B., Sonnenberg C., Winne P. H., Yacef K.
All: 8 shared	Allen L. K, Baker R.S.J.D, Biswas G., Graesser A. C., McNamara D. S., Pelánek R., Penstein-Rosé C., Snow E. L

An enrolment admission strategy based on data analytics

Michel C. Desmarais

Polytechnique Montreal, Canada michel.desmarais@polymtl.ca
http://www.professeurs.polymtl.ca/michel.desmarais/

Abstract. Every university program that has a limited capacity of enrolment faces the task of selecting the candidates that have the best chance of success. We introduce a selection strategy based on data analytics that only requires a ranking of candidates from different sources to determine a number of candidates to select from each source. The strategy relies on the distribution of student marks and on historical data of each source. It consists in determining a minimal threshold mark which, in turn, is used to determine proportions of students to admit from each source. The strategy ensures a maximum success rate under certain assumptions.

Keywords: Student Enrolment · Learning Analytics · Candidate Selection

1 Introduction

A case for the use of Learning Analytics in educational institutions can be made for the objective of selecting the candidates that have the best chance of success at a given university program. In the words of [8], we can consider the selection process as a standard machine learning prediction task:

> "Admission is to a great extent a prediction task, where admissions committees aim at estimating a candidate's chance of future study success. For these kinds of tasks, Meehl (1954) provided strong evidence for the superiority of the statistical approach over the clinical one. Since then, a plethora of studies has challenged this result but none contradicted Meehl's conclusion (Kahneman, 2011)."

While the candidate selection problem is trivial if the decision is based on a single criterion, such as the result of an admission test score (GPA, for eg., [1]; or GRE), or on any single score by which a candidate can be ranked, such score is not always available. Often, the decision must rely on a set of scores that are not comparable.

The typical situation is that an admission decision is based on the ranking of students within a given cohort and for a given institution. The choice is simple for the students from the same institution, but not for the students from different institutions. One solution is to ask candidate students to take an admission exam, but this is unpractical for students that apply from abroad or from distant locations. Moreover, the admission test may not be highly reliable [6].

Other solutions, often considered more reliable, are to to revert to interviews and personal statements [2]. But not only are their reliability questioned [4, 5, 7], these approaches also incur issues of time and efforts, which can be critical for large cohorts.

We introduce a means to decide on student admission based on historical data of the host institution itself. Given the information on student marks and their origin, one approach consists in determining the proportion of students from a given origin that are above a given score. The approach relies on computing the expected mean score of a proportion of students above a given score for a given origin. And the key to the approach is that the scores of all students are on the same scale, namely the institution's own grades.

The strategy is first described below, followed by a short demonstration of the impact it has compared to a simpler solution.

2 Historical data cutoff admission treshold (HDCT)

We will refer to the proposed approach as the Historical data cutoff admission treshold (HDCT). To illustrate its basic principle, consider Figure 1. It shows a distribution of student scores on a Z-scale that follows a Normal distribution ($\mathcal{N}(0,1)$) along with the proportion of students above the score, which corresponds to one minus the *cumulative distribution function* (labeled "cummul. admiss."). The dotted line indicates that the score of 0 corresponds to a proportion of 50% of students are above that score. We can also see from the "cummul. admiss." curve that at score $Z = 0.5$ we have about 80% of students above that score.

This graph is the basis of the HDCT admission process. The general principle is to determine the proportion of students to retain based on a common minimal score, obtained from the institution's historical data. Given that it is reasonable to assume that all student scores are on the same scale, namely the institution's historical scores, they are comparable even though the students may have different origins. And the key is not to rest the decision on a score obtained from the origin institution, but on historical data from the host institution. This approach incurs that the institution keeps track of which origin institution the student comes from and, as we discuss later, of the ratio of admitted students over the number of candidates.

To illustrate the general approach based on the above introduction, figure 2 shows the case where we have students from three different origins, source *a*, *b*, and c. The mean, standard deviation (s.d.), and the relative proportion of

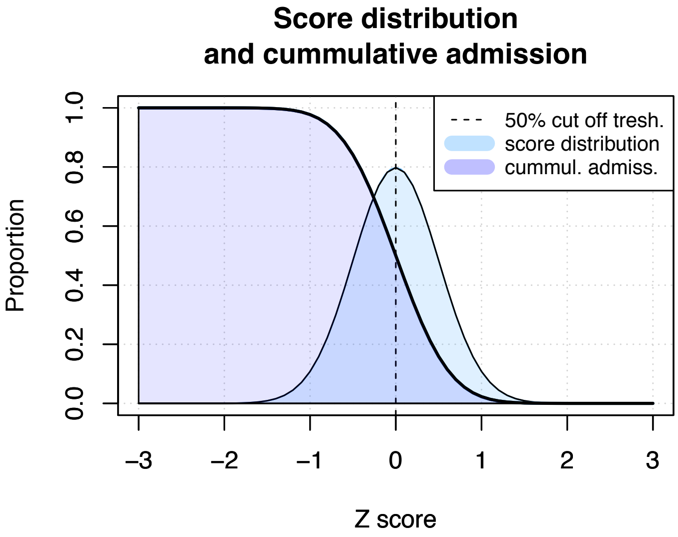


Fig. 1. Relation between the students score distribution on a Z-scale and the cumulative proportion of students admitted.

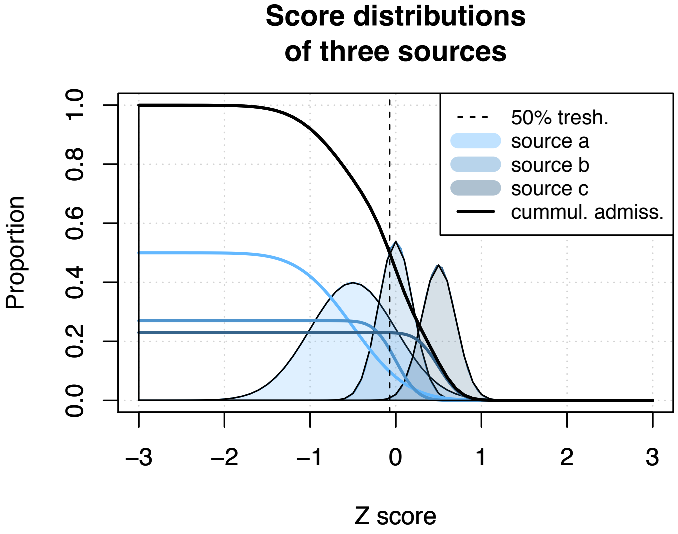


Fig. 2. Score distributions of three sources with different means and sandard deviations. The cumulative admission curve is shown for each source $(1 - cdf)$. They correspond to the three colored lines. The global cumulative admission is the black curve and corresponds to the sum of all three cumulative admission curves.

candidates from each source is shown below, along with the proportion admitted at the 50% cut off threshold.

Source	*Mean*	*S.d.*	*Proportion*	*Admitted@50%*
a	−0.5	0.5	50 %	20 %
b	0	0.2	27 %	64 %
c	0.5	0.2	23 %	99.8 %

We can see from the cumulative distribution curves that source a (mean=−0.5), source-b (mean=0), and source c (mean=0.5) respectively represent 50%, 27%, 23% of all students applicants. Because the variance of the distributions is not equal (0.5, 0.2, 0.2) and they also have uneven proportions, the cut off threshold to admit 50% of students is not at $Z = 0$, but instead around $Z = 0.07$. This threshold is shown as the dotted line in Figure 2: the score where the global cumulative distribution curve reaches 50% of all students, which in turns corresponds to 20% of source a, 64% of source b, and almost all of source c.

The implication of this graph is that if we had, for example, 1000 candidates and we wanted to admit only 500 of them, then only about 200 source a would be admitted, because it has had on average 0.5 standard deviation below the mean in the historical data. Whereas based on a policy of admitting the same ratio for all sources, we would then admit 250 of them for source a. Divergence from a uniform admittance ratio is even more stringent for the other two sources: almost all students from source c would be admitted because they historically scored 0.5 standard deviation above average and have a lower standard deviation, and most of source b would also be accepted.

The Z-score corresponding to the proportion of students we wish to admit from the total applicants is calculated based on an optimization function that can be defined as:

$$\arg\min_{Z} = \sum_{s} \left((prop_{s \in \mathrm{source}} \cdot pnorm(Z, \overline{sco}_s, sd_s)) - prop.admitted \right)^2$$

where:

- $prop_{s \in \mathrm{source}}$ is the proportion of applicants from a given source, s,
- *pnorm* is the cumulative distribution function (for the Normal distribution) that takes as arguments:
  - Z: the Z-score to optimize (threshold),
  - $\overline{sco}_s$: the mean historical score of the given source, and
  - sd_s: its standard deviation;
- *prop.admitted*: the proportion of students we wish to admit to meet the limited admission capacity.

2.1 Smoothing factor

In some cases, the number of students from a give source may be small, or even nonexistent if it represents a new source. To avoid extreme values of mean and standard deviations that result from small samples, a smoothing factor should be used. Assuming we have N_s students from source s, a smoothing factor α can be used to bring the mean of the score with the following smoothing formula:

$$\hat{x}_{is} = \frac{\sum_i^{N_s} x_i + \alpha\overline{x}}{N_s + \alpha}$$

where $\hat{x}_{is}$ is the smoothed value that should replace the value of the mean and $\overline{x}$ is the general mean of all students. A reasonable value is to have $\alpha = 5$, although the choice is rather arbitrary. A similar smoothing should be applied to the standard deviation based on historical data.

3 Impact example

To assess the impact of the admission strategy over a simpler one, we run a simulation and compare the difference in the expected scores of students admitted with each strategy.

The simpler strategy is to accept an equal proportion of students from each source.

Let us take the numbers from Figure 2 to run a simulation and assume we admit 1000 students. The expected average score from a given source corresponds to the number of students at a given score ($freq$(sco)), proportionally represented by the source's density of the distribution, times the score. This is repeated for each source and divided by the number of students (N) :

$$\mathrm{E}(\overline{\mathrm{sco}}) = \frac{\sum_{s \in \mathrm{source}} \sum_{\mathrm{sco} \in s} freq(\mathrm{sco}) \times \mathrm{sco}}{N}$$

The numbers that correspond to each strategy for each source are reported in the following table:

	Equal proportion		*HDCT*	
Source	N	E($\overline{sco}$)	N	E($\overline{sco}$)
a	250	−0.11	98	0.21
b	135	0.16	173	0.12
c	115	0.66	229	0.50
global	500	0.14	500	0.31

The major difference between the equal proportion and the proposed HDCT approach is that much fewer candidates are accepted from source *a* for the benefit of greater numbers from sources *b* and *c*. The effect is that the expected scores from source *a* increases while it decreases for sources *b* and *c*, but the overall effect is an increase in the expected score of 0.17 ($0.31 - 0.14$).

4 Conclusion

This paper describes a strategy to select the proportion of candidates to admit in order to maximize the expected success rate of the students to a given program. The strategy is based on historical data from the host institution. The advantage of the approach is that it does not require a standardized score across students from different institutions, which is most of the time unavailable unless the candidates are subject to an admission test. Considering that candidates can come from remote location and that running an admission test can involve considerable time and effort, this is a major advantage.

However, the approach has its limitation, the first of which is to have historical data from the different institutions the candidates come from. Often, the sample can be small and a correction in the form of a smoothing factor is proposed to alleviate this issue.

Another limitation is that, as described in this paper, it assumes the distribution of scores is Gaussian. Now, this limitation is not inherent to the general approach. Non Gaussian, or even arbitrary distributions could be handled, but the computations would need to be adapted to the actual distribution.

Finally, another issue is that the distributions have to reflect the scores of the origin institution, which must be derived from the historical data of the accepted candidates in the host institution. As presented in this paper, we assume the historical data is a faithful representation of that distribution, but if the selection is based on a small proportion of applicants, this assumption would be false. Here again, this is not a limitation of the approach itself, and computational adjustments would have to take this factor into account. The adjustment will rely on information about the ratio of admitted students per institution.

To close the loop on the question of how Learning Analytics can bring value to education, we use the admission problem that every institution faced with the need to select candidates from disparate source is confronted with. The candidate selection approach uses a strategy that relies on statistics and optimization techniques. It is an objective, effective, and efficient means to achieve the goal of selection the candidates that have the best chances of success.

References

1. Didier, T., Kreiter, C.D., Buri, R., Solow, C.: Investigating the utility of a GPA institutional adjustment index. Advances in health sciences education **11**(2), 145–153 (2006)
2. Eva, K.W., Reiter, H.I., Rosenfeld, J., Trinh, K., Wood, T.J., Norman, G.R.: Association between a medical school admission process using the multiple mini-interview and national licensing examination scores. Jama **308**(21), 2233–2240 (2012)
3. Kahneman, D.: Thinking, fast and slow. Macmillan (2011)
4. Meehl, P.E.: Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. (1954)
5. Murphy, S.C., Klieger, D.M., Borneman, M.J., Kuncel, N.R.: The predictive power of personal statements in admissions: A meta-analysis and cautionary tale. College and University **84**(4), 83 (2009)
6. Salvatori, P.: Reliability and validity of admissions tools used to select students for the health professions. Advances in Health Sciences Education **6**(2), 159–175 (2001)
7. Siu, E., Reiter, H.I.: Overview: what's worked and what hasn't as a guide towards predictive admissions tool development. Advances in Health Sciences Education **14**(5), 759 (2009)
8. Zimmermann, J., von Davier, A., Heinimann, H.R.: Adaptive admissions process for effective and fair graduate admission. International Journal of Educational Management **31**(4), 540–558 (2017). https://doi.org/10.1108/IJEM-06-2015-0080, `https://doi.org/10.1108/IJEM-06-2015-0080`

Exploring Opportunities for Caring Assessments

Diego Zapata-Rivera [1] and Julita Vassileva [2]

[1] Educational Testing Service, [2] University of Saskatchewan

Preface

The notion of intelligent systems that “care” about students is at the center of ITS research [1-3]. A variety of adaptive learning systems that “care” have been developed in the past [4, 5]. These systems make use of student/user models to adapt their interactions to a particular student (e.g., amount and type of feedback, content sequencing, scaffolding, and access to visualization tools and other materials). Student model variables include cognitive abilities, metacognitive skills, affective states, and other variables such as personality traits, learner styles, social skills, and perceptual skills [5].

Caring assessment systems are defined as systems that provide students with a positive assessment experience while improving the quality of evidence collected about the student’s knowledge, skills and abilities (KSAs) [6]. Taking a test is typically a stressful situation, and many people underperform due the stress. Caring assessment systems take into account assessment information from both traditional and non-traditional sources (e.g., student emotions, prior knowledge, and opportunities to learn) to create situations that students find engaging, and to collect valid and reliable evidence of students’ KSAs.

Taking a test is not just a passive mechanism for assessing how much people know. It actually helps people learn, and it works better than a number of other studying techniques [7]. Caring formative assessment can be done by a computer system or by peer-learners. Learners testing each other in a friendly, collegial and constructive way, can be an engaging and effective form of collaborative learning and preparation for assessment that also helps establish peer-mentorship relationships among learners. Developing systems or approaches (e.g. games) that support learners test each other in a friendly, collegial and constructive way, is a new and promising direction of research.

This workshop provides a great opportunity for ITS and assessment researchers to share information about the potential of applying ITS techniques and approaches in the development of a new generation of caring assessments. Examples of ITS technologies that have been successfully used for assessment purposes include automatic scoring of essays and short responses [8]. The use of dialogue systems for assessment is being explored [9, 10]. This workshop is a timely and relevant event for the ITS and assessment communities. New assessments for skills such as problem-solving, collaboration, and scientific inquiry include the use of highly interactive simulations and collaboration with artificial agents. Advances in ITSs will play an important role in the development of the next generation of assessment systems.

Eight recognized members of the research community were invited to serve as members of the program committee. Each member reviewed up to two submissions. The program committee members are: Ivon Arroyo, *Worcester Polytechnic Institute*; Ricardo Conejo, *University of Malaga*; Vania Dimitrova, *University of Leeds*; Sidney D’Mello, *University of Colorado Boulder*; Art Graesser, *University of Memphis*; G. Tanner Jackson, *Educational Testing Service*; Irvin R. Katz, *Educational Testing Service*; and Steve Ritter, *Carnegie Learning*.

Seven papers were submitted and all of them were accepted for presentation at the workshop. Each paper received feedback from at least two reviewers. The accepted papers include: *When Should an Adaptive Assessment Care?* (Blair Lehman, Jesse R. Sparks, and Diego Zapata-Rivera); *Incorporating Emotional Intelligence into*

Assessment Systems (Han-Hui Por and Aoife Cahill); *Diagnostic Assessment of Adults' Reading Deficiencies in an Intelligent Tutoring System* (Genghu Shi, Anne M. Lippert, Andrew J. Hampton, Su Chen, Ying Fang, and Arthur C. Graesser); *Tower of Questions: Gamified Testing to Engage Students in Peer Evaluation* (Nafisul Islam Kiron and Julita Vassileva); *Exploring Gritty Students' Behavior in an Intelligent Tutoring System* (Erik Erickson, Ivon Arroyo, Beverly Woolf), *Disengagement Detection Within an Intelligent Tutoring System* (Su Chen, Anne Lippert, Genghu Shi,Ying Fang, and Arthur C. Graesser); and *Assessments That Care About Student Learning* (Stephen E. Fancsali and Steven Ritter).

These papers offer different perspectives and current research toward the goal of making "caring" assessments part of the educational milieu.

The workshop included a thought-provoking discussion section that covered topics such as:

- The need for educating the public on the characteristics of different types of assessments and their appropriate use.
- Alternate criteria for adaptive testing that not only take into account the difficulty and the sequencing of questions but also other aspects of the student and the learning context and way of interaction.
- Assessments that provide additional feedback/guidance on content related issues and testing strategies (e.g., time management warnings).
- Using student model information from formative learning environments to inform the assessment systems.
- Possible approaches for integrating emotion data into assessment.
- Strategies for engaging students in peer assessment gaming activities.
- Exploring connections with other research areas (e.g., persuasive technologies).
- Evaluating the effects of additional features on test reliability, validity, and fairness.

We thank the authors for submitting relevant papers to the topic of the workshop, the program committee members for their time reviewing and providing constructive feedback to the authors, and the ITS workshop organizers, Nathalie Guin and Amruth Kumar, for providing us with this great opportunity to convene and address this topic.

Best regards,

Diego Zapata-Rivera and Julita Vassileva

References

1. Self, J.A. 1999. The distinctive characteristics of intelligent tutoring systems research: ITSs care, precisely, *International Journal of Artificial Intelligence in Education*, 10, 350–364
2. DuBoulay, B., Avramides, K., Luckin, R., Martinuz-Miron, E., Rebolledo Mendez, G., & Carr, A. 2010. Towards systems that care: a conceptual framework based on Motivation, Metacognition and Affect. *International Journal of Artificial Intelligence in Education*, 20, 197–229

3. Kay, J., & McCalla, G. 2003. The careful double vision of self. *International Journal of Artificial Intelligence and Education*, 13, 1–18
4. Brusilovsky, P., & Milan, E. 2007. User models for adaptive hypermedia and adaptive educational systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web. Methods and strategies of web personalization*. LNCS 4321, Berlin Heidelberg: Springer-Verlag. 3–53
5. Shute, V. J., & Zapata-Rivera, D. 2012. Adaptive educational systems. In P. Durlach (Ed.), *Adaptive technologies for training and education*. New York, NY: Cambridge University Press. 7–27
6. Zapata-Rivera, D. 2017. Toward Caring Assessment Systems. *In Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17)*, ACM, New York, NY, USA, 97–100. DOI: https://doi.org/10.1145/3099023.3099106
7. Karpicke, J., & Blunt, J. R. 2011. Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping. Science 20 Jan 2011: 1199327 DOI: 10.1126/science.1199327
http://science.sciencemag.org/content/early/2011/01/19/science.1199327
8. Shermis, M.D., & Burstein, J. 2013. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge Chapman & Hall.
9. Zapata-Rivera, D., Jackson, T., Liu, L., Bertling, M., Vezzu, M., and Katz, I. R. 2014. Science Inquiry Skills using Trialogues. *12th International conference on Intelligence Tutoring Systems*. 625–626.
10. Graesser, A.C., Dowell, N., & Clewley, D. 2017. Assessing Collaborative Problem Solving Through Conversational Agents. In: von Davier A., Zhu M., Kyllonen P. (eds) *Innovative Assessment of Collaboration. Methodology of Educational Measurement and Assessment*. Springer, Cham. 65–80

When Should an Adaptive Assessment Care?

Blair Lehman, Jesse R. Sparks, and Diego Zapata-Rivera

Educational Testing Service, Princeton NJ 08541, USA
[blehman, jsparks, dzapata]@ets.org

Abstract. Assessments can be a challenging experience for students. Students often have to consider more than just the knowledge being assessed, such as how to manage emotions that can impede performance (e.g., anxiety). But what if assessments cared about students and allowed them to just focus on the content of the assessment? In the present paper, we propose three time points at which assessments could care about students and discuss recent research that supports this model of assessments that care. The three time points include before, during, and after the assessment. Before students begin the assessment, the assessment format and design features can be adapted to the student; during the assessment adaptive support can be provided; and after the assessment students can be provided with personalized feedback. Adaptations would be made based on student characteristics (e.g., interest, self-efficacy) and behaviors during the assessment (e.g., emotions, response patterns). Ultimately, these adaptations at each time point would provide an individualized assessment experience for students, which could promote engagement and increase the quality of evidence collected about students' knowledge, skills, and abilities.

Keywords: Emotions, student characteristics, non-traditional assessments.

1 Introduction

Test taking has long been identified as an emotional experience for students (see Zeidner [1] for a review). Initially, research focused on test anxiety and its negative impact on performance in high-stakes assessments [1]. More recent research has also investigated the impact of students' low motivation or disengagement on performance in low-stakes assessments [2]. In both of these cases, the target emotion hinders students from performing to the best of their abilities on the assessment. Thus, students have an unpleasant experience and the assessment is not a valid measure of students' knowledge, skills, and abilities for those intending to use the scores.

Control-Value Theory [3], however, proposes a variety of positive (and negative) emotions that students are likely to experience during assessments and suggests that the positive emotions are beneficial for performance. Empirical research on overall student emotions for traditional assessments (e.g., multiple-choice items) has supported this proposed relationship [4]. Recent research on students' moment-to-moment emotions during a non-traditional assessment (e.g., conversation- and game-based assessments) has also shown that students experience both positive and negative emotions and that engagement, specifically, is beneficial for performance [5].

The majority of research on emotions during assessments has focused on documenting the range of emotions that students experience. However, this information can also be leveraged to provide emotion-sensitive support to students. Emotion-sensitive support has been integrated into several intelligent tutoring systems (ITS; see [6] for a review). This type of support has been found to particularly benefit students that were struggling with the learning activity (e.g., [7], [8]). The result of integrating this type of support for students during assessments would be assessments that care [9]. These so-called "caring" assessments, which consider students' experience while completing the assessment, can benefit the student and improve assessment validity. Students can have a more positive experience while completing the assessment and the assessment can be used to gather more valid evidence of the students' knowledge, skills, and abilities because the student is more engaged with the task.

This type of on-demand emotion-sensitive support has only been explored in one computer-based assessment [10], although it has been more thoroughly investigated in the ASSISTments program [11] that blends tutoring and assessment (e.g., [12]) and in educational activities that aim to improve learning, as mentioned previously. The effort-monitoring computer-based assessment developed by Wise et al. [10] provided reminders to students when careless responding was detected and was successful at getting students to respond in a more effortful manner. It is also important to note that caring assessments should not be limited to responding only to student emotions; for example, research on ITSs has shown that behaviors such as gaming the system [13] and student characteristics such as domain-relevant interest and prior knowledge [14] impact students' experiences and learning outcomes.

In the present paper, we propose a model of caring assessments that includes three time points at which assessments can adapt: before, during, and after the assessment. The adaptive support provided by ITSs is usually limited to the time *during* the learning activity. In the context of assessment, we would like to propose expanding beyond the assessment activities themselves to include front-end selection (*before*) of both format and design features as well as end of assessment feedback (*after*). The adaptations would be based on student characteristics and behaviors observed during the assessment. This larger characterization of the assessment process is supported by recent research showing that students experience a variety of emotions during assessment preparation (i.e., studying), assessment completion, and review of assessment performance feedback [15]. Next, we will discuss our proposed model for assessments that care.

2 Assessments that Care

We propose that caring can be integrated into assessments at three time points (before, during, and after the assessment) through various types of adaptations based on student characteristics and behaviors. Student characteristics can include more general personality traits as well as beliefs and perceptions about a specific domain. Student behaviors are dependent upon the attributes and content of the assessment and include the actions students take within the environment. These actions can be used to infer cognitive, emotional, and motivational states. The adaptations that can be made based on these inferences are often dependent on decisions made at previous time points. For example, on-demand adaptations to respond to student disengagement (e.g., providing motivational statements) will be constrained based on the previously selected assessment format (e.g., conversation-based assessment vs. traditional assessment). Thus, it is important to consider how adaptations build upon each other to create a more engaging assessment experience. Next, we discuss each time point and research that suggests that these adaptations are advantageous for students.

2.1 Time 1: Before the Assessment

The first time point is before the student begins the assessment. At Time 1 there are two types of adaptations that can occur: adaptations to the assessment format or to the design features of the assessment. Both types of adaptation would require information about student characteristics prior to administration of the assessment to create a student profile that would be used for adaptation decisions. Thus, before the assessment can be administered, information would need to be collected from students. This could potentially be problematic as the collection of additional information could either increase the total time for a test administration or require a separate administration session. Next, we will discuss each adaptation separately.

The main decision for assessment format is whether to have students complete a traditional assessment, a non-traditional assessment, or an assessment that has both types of items. Non-traditional assessments have been developed, in part, because they are hypothesized to provide a more engaging experience for students. This more engaging experience is proposed to then result in students performing to the best of their abilities. Recent research on game-based assessments (GBA) has shown that student performance is typically positively correlated with a more positive experience (e.g., [16]). However, there have been very few efforts directly comparing performance and experience between different assessment formats that assess the same knowledge and skills.

One exception comes from research on GBAs that assess argumentation skills. Lehman, Jackson, and Forsyth [17] compared student performance and experience on a traditional assessment and a GBA. The findings revealed that students who performed better on one assessment format than the other reported different emotional experiences. Specifically, students that performed better on the GBA compared to the traditional assessment reported more positive experiences during the GBA than those who performed worse on the GBA. However, this work did not explore the student characteristics that could be predictive of which assessment format afforded students the opportunity to perform to the best of their ability and have a positive experience. Knowledge of the relevant student characteristics would be critical to enable effective *a priori* assignment of students to a particular assessment format.

After the assessment format has been selected, the next opportunity for adaptation is what version of the assessment to administer to the student. By version, we mean that there is more than one option for the design of the tasks within the same assessment format assessing the same knowledge and skills. These different versions may involve varying more superficial aspects of the environment (e.g., surface features or presentation mode) to accommodations for students with disabilities. It is likely that non-traditional assessments will afford more opportunities for a variety of versions as they often include more elements to the environment such as agents who can have different characteristics or assume different roles. These design features can then be adapted to meet the students' needs.

Sparks, Zapata-Rivera, Lehman, James, and Steinberg [18] have begun investigating the use of different assessment versions in the context of a conversation-based assessment (CBA) that assesses science inquiry skills. Four versions of the CBA were developed that varied the knowledge level of a virtual peer agent (high, low) and how questions were framed (comparison, agreement). The findings revealed that overall the type of assessment evidence that could be collected varied for each version of the CBA and that the CBA version interacted with student characteristics (urban vs. rural school, prior knowledge). These findings suggest that some students could benefit more from different combinations of assessment design features rather than presenting all students with the same version of the assessment. It is also important to note that both types of adaptations before the assessment will require careful evaluation of the validity and equating of different assessment formats and versions to ensure comparability of scores across assessments (discussed further below).

2.2 Time 2: During the Assessment

The second time point at which adaptations can be employed to care about students is during the assessment. This type of adaptation is similar to the type of support that students receive from ITSs designed for learning. Specifically, there would be two layers of adaptation that encompass the inner and outer loop that dynamically select reactions to students' immediate actions (e.g., type of feedback) (inner loop) and adaptively select the next task for students to complete (outer loop) [19]. These adaptations can also include supports that address students' cognitive, emotional, and motivational states. Regardless of the type of support, these are all deployed based on an underlying student model that tracks students' knowledge and other states (e.g., gaming the system [13]) based on their behaviors in the environment.

Although a student model that includes information about students' cognitive, emotional, and motivational states has been incorporated into ASSISTments (e.g., [11], [12]), pure assessments (i.e., where learning is not an explicit goal) have generally utilized a less well-developed student model. Typically, computer adaptive assessments only include cognitive states (i.e., response quality as an indicator of knowledge level). One exception comes from the previously mentioned Wise et al. [10] study in which adaptive motivational support was successfully provided when student effort was monitored through response times. Recent research on student emotions during CBAs has revealed instances in which emotion-sensitive support could be beneficial for students [5]. For example, high intensity frustration was found to be persistent, grow in frequency over time, and be negatively related to performance. This finding suggests that a more complex student model that includes cognitive, emotional, and motivational states could benefit students during assessments.

2.3 Time 3: After the Assessment

The third time point at which assessments can care is when students receive feedback about the quality of their performance on the assessment. We have included performance feedback as part of the assessment process because its perceived utility is important for assessment validity [20]. Specifically, if students receive feedback that is difficult to understand, vague about how to make improvements, or demotivating, then the assessment is not effective as a tool for improving students' knowledge as students will be less likely to engage in productive learning behaviors after the assessment. It is important to note that feedback could potentially occur during the assessment as well. Given that feedback during the assessment is not always appropriate or desirable, we have chosen to only focus on feedback provided after the assessment. However, Time 2 could be expanded to incorporate the use of feedback, particularly in the case of formative assessments where such feedback may be more appropriate.

Score reports are often used to provide information about performance after an assessment, and the majority of score reporting research has focused on how to clearly display information such as measurement error [21]. However, some researchers have proposed that score reports should differ by audience (e.g., students vs. teachers) [20] and should be increasingly interactive [22]. We propose to go a step further when taking the audience into consideration. Specifically, we would like score reports to be individually tailored to each student. The individualized score reports would utilize the student model (student characteristics and behaviors) from the assessment to provide contextualized information about the quality of performance and practical next steps to improve performance [23]. Importantly, this report would need to be presented in a way that is meaningful to students and motivates them to engage in the strategies to improve future performance. We view the presentation of this tailored report as particularly important because if students do not view the report as useful, or are unwilling to adapt their future behaviors based on the report, the accuracy of the score report itself becomes less important.

3 Conclusion

We have proposed three time points at which adaption could be incorporated into assessment development to create "caring assessments." The three time points we proposed include *before* students begin the assessment (assessment format and design features), *during* the assessment (on-demand support), and *after* the assessment (personalized feedback). We have expanded the opportunities for caring beyond the assessment itself to encompass adaptations based on student characteristics outside of the assessment and the presentation of feedback after an assessment has been completed. However, it may be necessary to also include support for assessment preparation (i.e., studying) to create a complete caring assessment package [15].

Systems that provide adaptive support based on students' behaviors and even students' emotions during an educational activity are nothing new. There have been a variety of ITSs that detect, track, and respond to student emotions (see [6] for a review). However, this type of adaptivity has rarely been employed in educational activities that have assessment as the primary or only goal. There are two potential reasons for not including this type of adaptation in assessments. First, any type of adaptation will create a different assessment experience for students, which can make it more difficult for students' performance on the assessment to be equated. As mentioned previously, asking students to complete different assessments (formats and/or design features) requires that all of the assessments be equated to ensure that performance outcomes are comparable across the assessments. Equating is already part of assessment development when different forms of the same assessment are created [24], but the type of dynamic support that would be provided in an assessment that cares would likely further complicate the equating process.

The second reason that adaptive support has been employed more frequently in learning than in assessment activities has to do with the type of support that can be provided. An adaptive system that has the goal of facilitating student learning can provide a variety of support that gradually leads students towards the correct answer, or even provides the correct answer when students are struggling. This type of support is not likely to be useful when the goal of the system is to accurately assess students' current level of understanding. However, this does not mean that other types of adaptive support could not be utilized. For example, *Affect-Sensitive AutoTutor* [7] employs an intervention that targets both students' attributions and motivation. When students are found to be bored, confused, or frustrated, the tutor agent states that the students' current negative emotion was due to either the nature of the material (e.g., "This material is really challenging") or to the tutor (e.g., "I probably didn't explain the information very well") (attribution), followed by a statement encouraging the student to persist with the learning session (motivation). A similar approach could be adopted in assessments when students become disengaged; however, research is needed to determine the most effective approaches based on the student and the context.

We have presented some initial evidence that supports our proposed model of caring assessments. However, the evidence that we have presented is limited, in many cases to one study or context, and is only correlational. Thus, there are two critical next steps for future research in the development of caring assessments. First, the student characteristics that are most relevant to each time point of adaptation need to be identified. Second, the model needs to be tested for effectiveness of adaptations at each of the individual time points and for the overall model. It is important that we understand not only how adaptations at each time point impact students' performance and experience, but also how the adaptations interact across time points to impact the assessment. These caring assessments are hypothesized to provide three advantages:

(1) students will be more engaged and more likely to perform to the best of their ability, which in turn (2) will allow the assessment to collect more valid evidence of students' knowledge, skills, and abilities, and (3) students' more positive assessment experience may lead to more positive feelings in general about the domain and help to build students' self-efficacy. In other words, caring assessments will benefit a wide range of stakeholders who are involved in the assessment process.

References

1. Zeidner, M.: Test anxiety: The state of the art. Plenum Press, New York (1998).
2. Wise, S., Smith, L.: The validity of assessment when students' don't give good effort. In: Brown, G., Harris, L. (eds.) Handbook of Human and Social Conditions in Assessment, pp. 204-220. Routledge, New York (2016).
3. Pekrun, R.: The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. Educational Psychology Review, 18, 315-341 (2006).
4. Pekrun, R., Goetz, T., Frenzel, A., Barchfield, P., Perry, R.: Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). Contemporary Educational Psychology, 36, 36-48 (2011).
5. Lehman, B., Zapata-Rivera, D.: Intensity is as important as frequency for emotions during test taking. Contemporary Educational Psychology (in preparation).
6. Sottilaire, R., Graesser, A., Hu, X., Goldberg, B. (eds.): Design recommendations for intelligent tutoring systems, Vol. 2, Instructional Management. U.S. Army Research Laboratory, Orlando (2014).
7. D'Mello, S., Lehman, B., Graesser A.: A motivationally supportive affect-sensitive AutoTutor. In: Calvo, R., D'Mello, S. (eds.) New perspectives on affect and learning technologies, pp. 113-126. Springer, New York (2011).
8. Forbes-Riley, K., Litman, D.: Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. Speech Communication, 53, 1115-1136 (2011).
9. Zapata-Rivera, D.: Toward caring assessment systems. In: Tkalcic, M., Thakker, D., Germanakos, P., Yacef, K., Paris, C., Santos, O. (eds.) Adjunct Proceedings of User Modeling, Adaptation and Personalization Conference, pp. 97-100. ACM, New York (2017).
10. Wise, S., Bhola, D., Yang, S.-T.: Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. Educational Measurement: Issues and Practice, 25, 21-30 (2006).
11. Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N., Koedinger, K., Junker, B., et al.: The Assistment Project: Blending assessment and assisting. In: Looi, C., McCalla, G., Bredeweg, B., Breuker, J. (eds.) Proceedings of the Artificial Intelligence in Education Conference, pp. 555-562. ISO Press, Amsterdam (2005).
12. Pardos, Z., Baker, R., San Pedro, M., Gowda, S., Gowda, S.: Affective states and state tests: Investigating how affect and engagement during the school year predict end of year learning outcomes. Journal of Learning Analytics, 1, 107-128 (2014).
13. Baker, R., Corbett, A., Roll, I., Koedinger, K.: Developing a generalizable detector of when students game the system. User Modeling & User Adapted Interaction, 18, 287-314 (2008).

14. Lehman, B., D'Mello, S., Graesser, A.: Who benefits from confusion during learning? An individual differences cluster analysis. In: Yacef, K., Lane, C., Mostow, J., Pavlik, P., (eds.) Proceedings of the Artificial Intelligence in Education Conference, pp. 51-60. Springer-Verlag, Berlin/Heidelberg (2013).
15. Peterson, E., Brown, G., Jun, M.: Achievement emotions in higher education: A diary study exploring emotions across an assessment event. Contemporary Educational Psychology, 42, 82-96 (2015).
16. Jackson, G., Lehman, B., Forsyth, C., Grace, L.: Game-based assessments: Investigating relations between skill assessment, game performance, and user experience. Computers in Human Behavior (in review).
17. Lehman, B., Jackson, G., Forsyth, C.: A (mis)match analysis: Examining the alignment between test-taker performance in conventional and game-based assessments. Journal of Applied Testing Technology (in preparation).
18. Sparks, J.R., Zapata-Rivera, D., Lehman, B., James, K., Steinberg, J.: Simulated dialogues with virtual agents: Effects of agent features in conversation-based assessments. In: Proceedings of Artificial Intelligence in Education Conference (2018).
19. VanLehn, K.: The behavior of tutoring systems. International Journal of Artificial Intelligence in Education, 16, 227-265 (2006).
20. Zapata-Rivera, D., Katz, I.: Keeping your audience in mind: Applying audience analysis to the design of interactive score reports. Assessment in Education: Principles, Policy, and Practice, 21, 442-463 (2014).
21. Zwick, R., Zapata-Rivera, D., Hegarty, M.: Comparing graphical and verbal representations of measurement error in test score reports. Educational Assessment, 19, 116-138 (2014).
22. Zapata-Rivera, D., Underwood, J., Bauer, M.: Advanced reporting systems in assessment environments. In: Kay, J., Lum, A., Zapata-Rivera, D. (eds.) Proceedings of Learner Modeling for Reflection Workshop at the Artificial Intelligence in Education Conference, pp. 23-31 (2005).
23. Zapata-Rivera, D.: Adaptive score reports. In: Masthoff, J., Mobasher, B., Desmarais, M., Nkambou, R. (eds.) Proceedings of the User Modeling, Adaptation, and Personalization Conference, pp. 340-345. Springer, Berlin/Heidelberg (2012).
24. Kolen, M., Brennan, R.: Test equating, scaling, and linking: Methods and practices (2nd Ed.). Springer, New York (2004).

Incorporating Emotional Intelligence into Assessment Systems

Han-Hui Por and Aoife Cahill

Educational Testing Service, Princeton, NJ 08541, USA
{HPOR, ACAHILL}@ets.org

Abstract. This paper proposes developing emotionally intelligent assessments to increase score validity and reliability. We summarize research on three sources of data (process data, response data, and visual and sensory data) that identify students' needs during test taking and highlight the challenges in developing caring adaptive assessments. We conclude that because of its interdisciplinary nature, the development of caring assessments requires closer collaborations among researchers from diverse fields.

Keywords: Caring assessments, Adaptive testing, Emotions.

1 Introduction

Assessments can be stressful. Under-performance due to test anxiety can undermine the validity of a test score by failing to recognize the true performance of the student. Caring assessments [37] are systems developed to respond to students' needs by adjusting content sequencing, moderating the amount and type of feedback, adding visualization aids, etc. In addition to students' ability, caring assessment systems take into account additional information from both traditional and non-traditional sources (e.g., student emotions, prior knowledge and opportunities to learn). Caring assessments go beyond traditional assessments by providing the encouragement and resources that students might need.

One way to identify a student's learning needs is via their emotions during test-taking, such as the affective-sensitive version of AutoTutor [9], the uncertainty (i.e., confusion) adaptive version of ITSpoke (UNC-ITSpoke) [12], and emotion-sensitive versions of Cognitive Tutors and ASSISTments [1]. These systems have the "emotional intelligence" to recognize the emotions and needs of their learners and use the information to help the learners achieve their learning goals.

In this paper, we argue that a fair and valid caring assessment should take multiple interdisciplinary sources of information into account. We give an overview of current state-of-the-art capabilities -- both technological and psychometric -- that are relevant for designing and developing caring assessments. We outline some of the many challenges faced and suggest some areas for future work, particularly focusing on interdisciplinary collaborations. While we are particularly interested in the role of caring assessments in the context of summative assessments, parts of our discussion will also refer to elements of caring assessments that will also be helpful for formative assessments.

2 Adaptive Testing

In caring assessments, students see different sets of questions and aids that are tailored to their ability and needs. In traditional computerized adaptive testing (CAT), pre-calibrated test items or testlets (sets of items, as in multi stage adaptive testing) are presented to students based on the quality of responses to previous questions. Given the response (correct/incorrect), an estimate of the student's ability is updated before further items are selected and presented. As a result of adaptive administration, different students experience different items. By successively fielding items selected to provide the maximum information about a student's ability, the maximum information about the student's ability is collected with each question, resulting in a shorter exam. Compared to a traditional static exam, a CAT exam is an assessment that tailors itself to each student's ability. Unfortunately, an algorithm that assigns items based on ability alone may not necessarily support a positive testing experience [27]. However, it may be possible to leverage psychometric approaches to incorporate additional information -- beyond ability -- into item selection in CAT.

2.1 Item Selection using Item Response Theory (IRT) and Bayesian Networks Models

IRT [19] is the current dominant statistical model used in CAT to assign students comparable, meaningful scores, even though students see sets of items tailored to their individual needs. Within the IRT framework, the estimated ability of a student is expected to be the same regardless of the items. The usefulness of IRT to select hints based on students' ability was demonstrated in FOSS (Full Option Science System), an ITS that was part of the NSF-funded Principled Assessment Designs for Inquiry (PADI) assessment [29].

To take into account both ability as well as the emotional needs of students, parameters of items in caring assessments can be estimated using models that integrate person- and variable-centered information such as the mixed-measurement item response theory (MM-IRT) [23, 24]. Such models focus on unobserved characteristics by identifying latent classes of individuals who respond to items in unexpected but distinct ways [13, 40]. Indeed, there is a recognition that observed scores may not correspond with unseen differences in how individuals respond to tests, such as differences in test strategies [23] or reactions to testing procedures [3].

Bayesian graph models from the Bayesian Network framework is another promising approach. These models allow the modeling of relevant conditional probabilities to update the likelihood of an event in the network. A major advantage is that the assessment of item mastery can be defined using multiple latent traits. In particular, it has been shown that the POKS (Partial Order Knowledge Structures) Bayesian modeling approach is computationally simpler and can outperform a 2-parameter IRT model in some instances [7]. The Andes Tutor [5] and Hydrive [22] both incorporate Bayesian network models to select and score items.

3 Research on Emotions and Affective Computing

Caring assessments have to take into account students' emotional states (e.g., anxiety, frustration) during test-taking. Using self-reported measures, Lehman and Zapata-Rivera [18] identified the emotions that occurred when students completed conversation-based assessments (CBAs). They found that students experienced similar emotions across two studies and concluded that boredom, confusion, curiosity, delight, engagement/flow, frustration, happiness/enjoyment, hope, and pride are the prevalent emotions in CBAs. In an adaptive assessment, the use of self-reported measures disrupts test flow and other predictive indicators are necessary to identify emotions accurately. In addition to correct/ incorrect responses from the typical item types such as multiple choices items, we can enhance our prediction of students' emotions using three sources of data: process data, response data, and visual and sensory data.

Process Data such as response time, keystrokes, mouse clicks

Response Data such as responses to multiple choices or constructed response items, text or verbal feedback from students

Visual and Sensory Data such as eye tracking, heart rates, postures, facial expressions

The data and model complexity used in an assessment to predict emotions depends on the assessment objectives. Assessments used to identify areas of students' strengths, and weaknesses require detailed information about each student to provide very specific and individualized support. On the other hand, assessments conducted by teachers to inform teaching and learning direction require aggregated information of the group, and less precise information about individual students.

The amount of information also depends on the nature of the assistance the assessment aims to provide. A formative assessment is likely to provide more assistance in the form of additional visual aids, redirecting students' focus to the correct cues or paraphrasing questions when necessary. A summative assessment for determining proficiency levels can still benefit from collecting some amount of information to promptly identify students experiencing technical difficulties during the assessment. In the case of multi-year assessments, information can also be collected to aid in the development of exams in subsequent years. In the next section, we summarize research that has been done with each of the three information sources.

3.1 Process Data

Process data, such as typing speed, response time, keystrokes, mouse clicks and action sequences in problem solving tasks, trace students' progress through an assessment. Most process data can be collected in the background with minimal incremental costs and is unobtrusive to students taking the exam. In general, they fall into three categories: what a student does, in what order, and how long it takes to do it. For instance, an analysis of the patterns and pauses in students' typing in a NAEP writing test showed that students who used the delete key more often, as a measure of their attempts to edit, had higher scores than students who did not delete as much [33]. The findings suggest that the latter group could benefit from encouragement to edit. Wise et al. [34] found that monitoring learners' response time and displaying warning messages when learners exhibited rapid guessing behavior improved scores and score validity, as indicated by the higher correlation between the test score and the learners' GPA and SAT scores. Other studies using timing data explored students' test taking behavior [15,16].

3.2 Response Data

Natural language processing (NLP) techniques are widely used to automatically measure the quality of constructed (free) responses in educational assessments. NLP is used in automated scoring engines to assess students' level of comprehension or writing proficiency and subsequently drive the feedback that students receive. Beigman-Klebanov et al. [2] show that by using NLP techniques it is possible to automatically predict a student's *utility value* -- a measure of how well the student can relate what they are writing about to themselves or other people -- from the student's writing. Flor et al. [10] show that it is possible to automatically categorize the dialogue acts (including expressing frustration) in a collaborative problem-solving framework using NLP techniques. NLP techniques are used on both written and spoken text. Spoken data can also provide a rich amount of information on both speaker emotions as well as their thought process (disfluencies, pause structure, etc.). Studies have also examined the use of low-level linguistic features to predict student emotions during human and computer tutoring sessions [17, 8]. Future research can focus on using complex linguistic analysis to learn more sophisticated relationships between the content of students' responses and their emotions in real time assessments.

3.3 Visual and Sensory Data

Visual and sensory data can also be captured to provide information on the students' progress or emotional state. The interest in sensory information stems from the findings that increased heart-rate and perspiration often precede our actual awareness of emotions, and studies have shown that heart rate and respiratory frequency can distinguish between neutral (relaxed), positive (joy) and negative (anger) emotions [31, 36]. However, while pulse rate monitors can be small, most devices would likely be obstructive when taking tests.

Although advances in facial recognition technology have vastly improved in recent years, identifying emotions accurately in real time is still a challenging task. Facial expressions are an integral part of emotions, but can also exist independently of emotions [25] and vice-versa. More recent developments suggest that new facial recognition algorithms have had some success with extracting features to classify students' emotions [35] in real time.

Eye tracking also allows us to pinpoint sections of the items that students are focused on. Studies on the usefulness of eye tracking data have provided preliminary evidence that they provide insights into how students respond to test items and solve problems [21, 28, 30]. A study that used eye tracking devices found that students with a history of performing poorly on reading tests did better when they had to write a summary of a reading passage before answering multiple-choice questions on the content [32]. The eye-tracking data showed that those students spent more time reading the initial text, and less time referencing the passage, suggesting that the students had built a mental memory model of the text. The advantage was stronger in students weaker in reading.

4 Challenges of Developing Caring Assessments

The development of caring assessments presents numerous challenges and research directions. Above all, the development of caring adaptive assessments requires closer interdisciplinary collaboration. Current research on the use of process data, NLP data, and visual/sensory data is largely focused on how these features correlate with either students' performance [21, 28, 30, 33, 34] or human raters in the field of automated scoring [11, 14, 20]. On the other hand, data from multiple sources would allow us to build accurate large-scale models of behavior from which we could then generalize students' behavior [4, 6, 38] and adapt to their needs. One area of future research is to focus on the predictive value of data from multiple sources in predicting students' emotions, and the impact of responding with aids on students' learning.

Other challenges include the costs and benefits of caring assessments over traditional ones. For caring assessments to be adopted as an industry standard, it will be necessary to demonstrate the effectiveness of the caring components both in terms of improving the student experience as well as contributing to overall test reliability and validity. As the approach to caring assessments is different in different educational context (e.g., summative vs formative), additional work is needed to define the elements that make up caring assessments so that the elements and combination of elements can be studied for their effectiveness.

In addition, the widespread adoption of caring assessments will be dependent on technology. Established assessments such as the GRE, TOEFL, LSAT depends on the capabilities of their testing centers. Therefore, if a caring assessment requires high-resolution cameras, all test centers would need to provide that hardware. In large-scale international assessments with test centers in all corners of the world, this is no small challenge.

4.1 Psychometrics Challenges in Caring Assessments

The psychometrics of caring assessments also presents some challenges. The current challenge is to adapt assessments based on students' ability and needs. While adaptive testing is not new, we need further research to establish if available models can accommodate multi modal, individual, and item level characteristics.

Scoring Complex Data Sequences

Another significant challenge for interactive assessments that respond to students' needs is that students can choose to take a large combination of actions. Should a student be rewarded with more points for taking fewer steps to get to the correct response? Recent research in psychometrics suggests that incorporating process data in assessments is tenable. A transition network using weighted directed networks can capture activity sequences, with nodes representing actions and directed links connecting two actions only if the first action is followed by the second action in the sequence [39]. As for scoring, Shu et al. [26] proposed a Markov-IRT model to characterize and capture the unique features of students' individual response process during a problem-solving activity in scenario-based tasks by laying out the model structure, its assumptions, the parameter estimation and parameter space. The Markov-IRT model allows test developers to determine the mapping of specific combinations to scoring rubrics.

Implications for Summative Assessments

Psychometric research can also contribute to scoring issues, particularly for high stakes summative assessments, where assigning valid and reliable scores that reflect students' skill mastery is a critical component. These assessments involve further issues such as score discrimination between students, in that students who score higher have better mastery than students with lower scores, and score comparability across cohorts of students who take different versions of the assessments.

Further, standard concerns in testing that are typically of lesser importance in learning assessments will surface. Issues such as fairness in testing, item overexposure, the establishing of cut scores, scaling and equating of scores, reporting and use of scores have been extensively studied and will also need to be adapted for a caring assessment.

5 Conclusion

We posit that caring assessments have a place in both formative and summative assessments. To get there, we will require that researchers from diverse backgrounds, such as computer science, engineering, natural language processing, learning, and psychometrics, work closely together to make sure that any new caring assessment is as valid and reliable as possible.

References

1. Baker, R.S., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z., Salvi, A., Aleven, V., Kusbit, G.W., Ocumpaugh, J., Rossi, L.: Towards sensor-free affect detection in cognitive tutor algebra. International Educational Data Mining Society (2012)
2. Beigman Klebanov, B., Burstein, J., Harackiewicz, J., Priniski, S., Mulholland, M.: Enhancing stem motivation through personal and communal values: Nlp for assessment of utility value in student writing. In: Proceedings of the 11th Workshop on Innovative Use of NLP

for Building Educational Applications. pp. 199-205. Association for Computational Linguistics, San Diego, CA (June 2016)

3. Bolt, D.M., Cohen, A.S., Wollack, J.A.: Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. Journal of Educational Measurement **39**(4), 331–348 (2002)
4. Calvo, R.A., D'Mello, S.: Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Transactions on affective computing **1**(1), 18-37 (2010)
5. Conati, C., Gertner, A., Vanlehn, K.: Using bayesian networks to manage uncertainty in student modeling. User modeling and user-adapted interaction 12(4), 371–417 (2002)
6. von Davier, A.A.: Computational psychometrics in support of collaborative educational assessments. Journal of Educational Measurement **54**(1), 3–11 (2017)
7. Desmarais, M.C., Pu, X.: A bayesian student model without hidden nodes and its comparison with item response theory. International Journal of Artificial Intelligence in Education **15**(4), 291–323 (2005)
8. D'Mello, S.K., Dowell, N., Graesser, A.C.: Cohesion relationships in tutorial dialogue as predictors of affective states. In: AIED. pp. 9–16 (2009)
9. DMello, S.K., Lehman, B., Graesser, A.: A motivationally supportive affect-sensitive autotutor. In: New perspectives on affect and learning technologies, pp. 113–126. Springer (2011)
10. Flor, M., Yoon, S.Y., Hao, J., Liu, L., von Davier, A.: Automated classification of collaborative problem solving interactions in simulated science tasks. In: Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. pp. 31–41. Association for Computational Linguistics, San Diego, CA (June 2016)
11. Flor, M., Yoon, S.Y., Hao, J., Liu, L., von Davier, A.: Automated classification of collaborative problem solving interactions in simulated science tasks. In: Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. pp. 31–41. Association for Computational Linguistics, San Diego, CA (June 2016)
12. Forbes-Riley, K., Litman, D.: Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. Speech Communication 53(9-10), 1115–1136 (2011)
13. Hernández, A., Drasgow, F., González-Romá, V.: Investigating the functioning of a middle category by means of a mixed-measurement model. Journal of Applied Psychology **89**(4), 687 (2004)
14. Jeon, J.H., Yoon, S.Y.: Acoustic feature-based non-scorable response detection for an automated speaking proficiency assessment. In: Thirteenth Annual Conference of the International Speech Communication Association (2012)
15. Lee, Y.H., Haberman, S.J.: Investigating test-taking behaviors using timing and process data. International Journal of Testing 16(3), 240–267 (2016)
16. Lee, Y.H., Jia, Y.: Using response time to investigate students' test-taking behaviors in a naep computer-based study. Large-scale Assessments in Education 2(1), 8 (2014)
17. Lehman, B., DMello, S.K.: Predicting student affect through textual features during expert tutoring sessions. Presented at the annual meeting of the Society for Text and Discourse (2010)
18. Lehman, B., Zapata-Riveria, D.: Student Emotions in Conversation-Based Assessments. IEEE Transactions on Learning Technologies (In Print)
19. Lord, F.: Application of item response theory to practical testing problems. Hillsdale, NJ, Lawrence Erlbaum Ass (1980)

20. Madnani, N., Cahill, A., Riordan, B.: Automatically scoring tests of proficiency in music instruction. In: Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. pp. 217–222 (2016)
21. Mayer, R.E.: Unique contributions of eye-tracking research to the study of learning with graphics. Learning and instruction 20(2), 167–171 (2010)
22. Mislevy, R.J., Gitomer, D.H.: The role of probability-based inference in an intelligent tutoring system. ETS Research Report Series 1995(2) (1995)
23. Mislevy, R.J., Verhelst,N.: Modeling item responses when different subjects employ different solution strategies. ETS Research Report Series 1987(2) (1987)
24. Rost, J.: A logistic mixture distribution model for polychotomous item responses. British Journal of Mathematical and Statistical Psychology 44(1), 75–92 (1991)
25. Rozin, P., Cohen, A.B.: High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial ex- pressions of americans. Emotion 3(1), 68 (2003)
26. Shu, Z., Bergner, Y., Zhu, M., Hao, J., von Davier, A.A.: An Item Response Theory Analysis of Problem-Solving Processes in Scenario-Based Tasks. Psychological Test and Assessment Modeling 59(1), 109 (2017)
27. Shute, V.J., Hansen, E.G., Almond, R.G.: You can't fatten a hog by weighing it–or can you? evaluating an assessment for learning system called aced. International Journal of Artificial Intelligence in Education 18(4), 289–316 (2008)
28. Tai, R.H., Loehr, J.F., Brigham, F.J.: An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessments. International journal of research & method in education 29(2), 185–208 (2006)
29. Timms, M.J.: Using item response theory (IRT) to select hints in an ITS. Frontiers in Artificial Intelligence and Applications 158, 213 (2007)
30. Tsai, M.J., Hou, H.T., Lai, M.L., Liu, W.Y., Yang, F.Y.: Visual attention for solving multiple-choice science problem: An eye-tracking analysis. Computers & Education 58(1), 375–385 (2012)
31. Valderas, M.T., Bolea, J., Laguna, P., Vallverdú, M., Bailón, R.: Human emotion recognition using heart rate variability analysis with spectral bands based on respiration. In: Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE. pp. 6134–6137. IEEE (2015)
32. Wang, Z., Sabatini, J., OReilly, T., Feng, G.: How individual differences interact with task demands in text processing. Scientific Studies of Reading 21(2), 165–178 (2017)
33. White, S., Kim, Y.Y., Chen, J., Liu, F.: Performance of Fourth-Grade Students in the 2012 NAEP Computer-Based Writing Pilot Assessment: Scores, Text Length, and Use of Editing Tools. Working Paper Series. NCES 2015-119. National Center for Education Statistics (2015)
34. Wise, S.L., Bhola, D.S., Yang, S.T.: Taking the Time to Improve the Validity of Low-Stakes Tests: The Effort-Monitoring CBT. Educational Measurement: Issues and Practice 25(2), 21–30 (2006)
35. Yang, D., Alsadoon, A., Prasad, P., Singh, A., Elchouemi, A.: An Emotion Recognition Model Based on Facial Recognition in Virtual Learning Environment. Procedia Computer Science 125, 2–10 (2018)
36. Yu, S.N., Chen, S.F.: Emotion state identification based on heart rate variability and genetic algorithm. In: Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE. pp. 538–541. IEEE (2015)
37. Zapata-Rivera, D.: Toward caring assessment systems. In: Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization. pp. 97–100. ACM (2017)

38. Zapata-Rivera, D., Liu, L., Chen, L., Hao, J., von Davier, A.A.: Assessing science inquiry skills in an immersive, conversation-based scenario. In: Big data and learning analytics in higher education, pp. 237–252. Springer (2017)
39. Zhu, M., Shu, Z., Davier, A.A.: Using Networks to Visualize and Analyze Process Data for Educational Assessment. Journal of Educational Measurement 53(2), 190–211 (2016)
40. Zickar, M.J., Gibby, R.E., Robie, C.: Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. Organizational Research Methods 7(2), 168–190 (2004)

Diagnostic Assessment of Adults' Reading Deficiencies in an Intelligent Tutoring System

Genghu Shi[1, 2], Anne M. Lippert[1, 2], Andrew J. Hampton[1, 2], Su Chen[1, 2], Ying Fang[1, 2], and Arthur C. Graesser[1, 2]
[1] University of Memphis, Memphis TN 38111, USA
[2] Institute for Intelligent Systems
gshi@memphis.edu

Abstract. In this paper, we investigate whether a version of AutoTutor that teaches comprehension strategies can be used to diagnose reading deficiencies in adults with low literacy. We hypothesized that the speed and accuracy with which participants answered questions during the AutoTutor conversation could be diagnostic of their mastery of reading comprehension components: *words*, the explicit *textbase*, the *situation model*, and *rhetorical structure*. We used linear mixed effect models to compare the accuracy and response times of 52 low literacy adults who worked on 29 AutoTutor lessons during a four-month intervention period. Our results show that adults' response accuracy for questions addressing more basic reading components (e.g., meaning of words) was higher than for those pertaining to deeper discourse levels. In contrast, question response time did not vary significantly among the theoretical levels. A correlation analysis between theoretical levels and performance (accuracy and time) supported this trend. These results affirm that adults with low literacy tend to have more proficiency for basic reading levels than for deeper discourse levels. In addition, the results of exact binomial test showed that hints or prompts were effective in scaffolding learning reading. Furthermore, we describe how response accuracy on the four comprehension components can provide a more nuanced diagnosis of reading problems than a single overall performance score. More fine-grained diagnoses can assist both educators wanting more detailed insight into learner difficulties, and ITS developers looking to improve the personalization and adaptivity of learning environments.

Keywords: CSAL AutoTutor, Reading strategies, Comprehension framework.

1 Introduction

One in six adults in the United States has low levels of literacy skills [1]. Low literacy has a negative impact on the social health and economic stability of entire countries as well as the personal well-being of its citizens [1, 2]. Adult literacy educational programs are often funded by government or non-profit organizations, but unfortunately these programs generally do not reach the level that can accommodate all adults in need. Moreover, it is difficult to teach comprehension strategies at deeper levels because few teachers and tutors in literacy centers are trained to cover these levels of reading difficulty. Intelligent tutoring systems can help close this gap and provide the necessary, deeper training. An intelligent tutoring system that can differentially diagnose reading deficits constitutes an important first step in adaptively remediating individuals' deficits. In this study, we explore the assessment capabilities of a version of a web-based intelligent tutoring system, AutoTutor [4, 7], specifically created for adults with low literacy. In particular, we use AutoTutor to classify the reading comprehension deficiencies of adults within the Graesser and McNamara [3] multilevel theoretical framework of reading comprehension.

AutoTutor for CSAL

The version of AutoTutor we developed was part of an intervention led by the Center for the Study of Adult Literacy (CSAL) [4, 7], and helps improve reading comprehension in low literacy adults. The system has two computer agents (one tutor and one peer student) that hold conversations with the human learners and with each other, called trialogues [4, 5]. Trialogues illustrate comprehension strategies to adult learners, help them apply these strategies, and give them feedback when assessing their performance, all in natural language. CSAL AutoTutor has 35 lessons that focus on distinct theoretical levels of reading comprehension [6, 7]. For each lesson, the system starts out assigning words or texts at a medium level of difficulty and AutoTutor asks 8-12 questions about the words or text, all embedded in an overarching conversation. Struggling readers tend to have even more pronounced difficulties in writing, so most of their responses are entered by clicking response options on the interface. Learner response accuracy on the medium level questions determines whether AutoTutor assigns new words or texts at a hard or easy (above or below some performance threshold) level [8]. When answers do not include all component parts of a good answer, the learner receives hints or prompts, providing another chance to pick an answer from the remaining two choices with somewhat more guidance.

CSAL AutoTutor was designed to "care" about the particular motivations, metacognitions and emotions of struggling adult readers. The caring aspect of CSAL AutoTutor is critical because most adults participating in literacy programs do so voluntarily, and if the instruction is not adult-oriented, engaging, and pertinent to adult daily life, they will stop attending. Thus, in addition to allowing easy access, individualized self-paced instruction, and intuitive design for low literacy adult learners, AutoTutor was designed to optimize engagement. First, lessons were carefully scripted to contain texts that have practical value to the adult (such as rental agreements, job applications, recipes, health information) or are expected to interest adults. Second, texts are adaptively selected by AutoTutor to be at a reading level that the student can handle (not too hard or too easy), so that the student does not become frustrated or bored. Third, trialogues were written to boost the self-esteem of the adult learner who may feel embarrassment or shame over his or her skill level. Both agents express positive encouraging messages when the adult is not performing well, and sometimes stage game-like competitions between the adult and a peer agent (with the adult always winning, thereby enhancing self-esteem). These caring functionalities of AutoTutor help create situations that users find engaging and welcoming and simultaneously allow the system to assess learner ability.

1.1 The Multilevel Framework of Comprehension

The Graesser and McNamara [3] framework identifies six theoretical levels: *words*, *syntax*, the *explicit textbase*, the *referential situation model,* the *discourse genre and rhetorical structure,* and *the pragmatic communication level* (between speaker and listener, or writer and reader). Because AutoTutor for CSAL includes only one lesson for syntax and none for pragmatic communication, we did not include these levels in our study. Of the levels we included, *word* represents the lower-level basic reading components that include morphology, word decoding, and vocabulary. The *textbase* consists of meaning of the explicit ideas in sentences and texts. The *referential situation model* (sometimes called the mental model) represents the subject matter that the texts are describing. *Genre and rhetorical structure* focuses on the type of discourse and its composition, such as narrative, persuasive, and informational genres, and also the subcategories of these genres. The last three theoretical levels (all except *word*) represent deeper discourse levels.

We hypothesize that the accuracy and time on questions in AutoTutor will be diagnostic of adult learners' mastery of comprehension components. By comparing the accuracy and time on questions of four theoretical levels [3], we can better pinpoint where adult learners' strengths and weaknesses in reading comprehension lie. Such results can provide a more nuanced diagnosis of reading problems than a single overall performance score and ultimately help improve the adaptivity of an ITS like AutoTutor. We also hypothesize that adult learners who do not answer correctly on the first attempt, and receive guidance through hints or prompts for the second attempt will perform better than chance on these questions. These results will provide insight into AutoTutor's effectiveness in helping adult learners with reading comprehension.

2 Method

2.1 Participants

The participants were 52 adults recruited from CSAL literacy classes in Metro-Atlanta (*n* = 20) and Metro-Toronto (*n* = 32). They worked on 29 lessons during a four-month intervention. Each lesson took 20 to 50 minutes to complete. Their ages ranged from 16–69 years (Mean = 40, SD = 14.97). Most of the participants were female (73.1%). All participants read at 3.0–7.9 grade levels, and 30% reported that they were either diagnosed as learning disabled or attended special education classes in their childhood.

2.2 Measures and Data Collection

Only the adults' initial responses (1 as correct, 0 as incorrect) of medium level questions in each of the 29 lessons contributed to the diagnostic analysis. This ensured a balanced design, as all participants were assigned the medium level texts, but not all participants subsequently received the easy or difficult texts. In addition, the medium level questions produce higher level discrimination. We used only the initial (as opposed to second) attempts to questions because we felt these would best reveal adults' actual mastery of the theoretical levels of comprehension. For these medium-level observations, we collected the accuracy (1 or 0) and the time to produce an answer (in seconds). Time was measured from the onset of the question to the onset of the participant's answer.

To assess the effectiveness of the hints or prompts, we collected accuracy (1 or 0) of the second attempt to all questions which were answered incorrectly on the first attempt by learners. Second attempts involved all difficulty levels (medium, easy, and hard).

We calculated accuracy and time measures for 29 lessons. Most of the lessons focus on more than one theoretical level (at most three) but have varying degrees of relevance within a lesson. For example, the lesson "Compare and Contrast" addresses mainly the *rhetorical structure* level, but also includes material involving the *textbase* and *situation model* levels. Thus, we included a relevance score for each of the four theoretical levels for each lesson. The most relevant theoretical level on a lesson received a score of 1.00, with scores of 0.67 and 0.33 assigned to the second and third order, respectively. The fourth theoretical level received a 0.00 and was thus nullified for that lesson.

2.3 Data Analysis

From each set of participant log files, we extracted time and accuracy data for the 29 lessons. We found that the distribution of response time per question was positively skewed. To alleviate the bias brought by potential outliers, we truncated the data by replacing observations falling outside three standard deviation above the mean with the corresponding value at three z-score units beyond the mean.

We first performed a descriptive analysis of the data by exploring the means and standard deviations of accuracy and time on questions of the four theoretical levels. Next we used mixed effect modeling [9], where item (question) was the unit of analysis, to test for differences in time and accuracy among the four theoretical levels. To account for the variability in participants, lessons, and questions, these components were included in the linear mixed effect models as random intercepts. We also added by-participant random slopes on different theoretical levels and random intercepts of the interaction between lesson and item for the nesting relationships. Follow-up correlational analyses were performed on the continuous measures of theoretical levels, as well as on the accuracy and time for the 29 lessons. In addition, we conducted an exact binomial test on the accuracy of second attempts to see if the proportion of correct responses is greater than chance (50%).

3 Results

Figures 1 and 2 show the means of accuracy and time on questions separately as a function of four theoretical levels. Here we see accuracy is highest and answer times are shortest for the *word* level (reference level in the analysis) compared to the three discourse levels (*textbase, situation model,* and *rhetorical structure*).

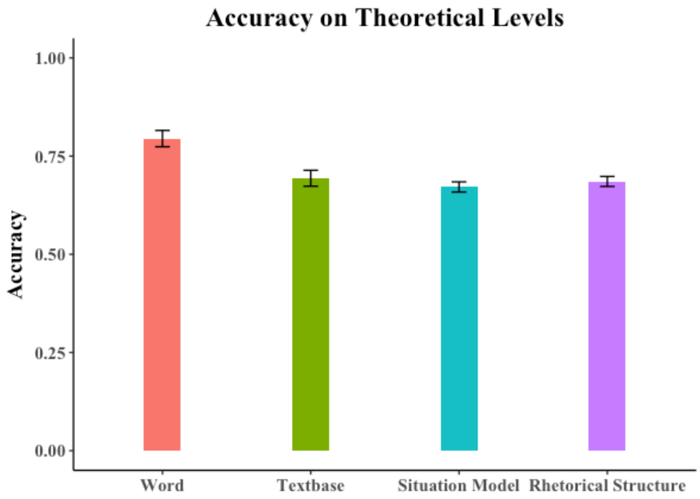


Figure 1. Adults' means accuracies (scale 0–1) on four theoretical levels, with error bars.

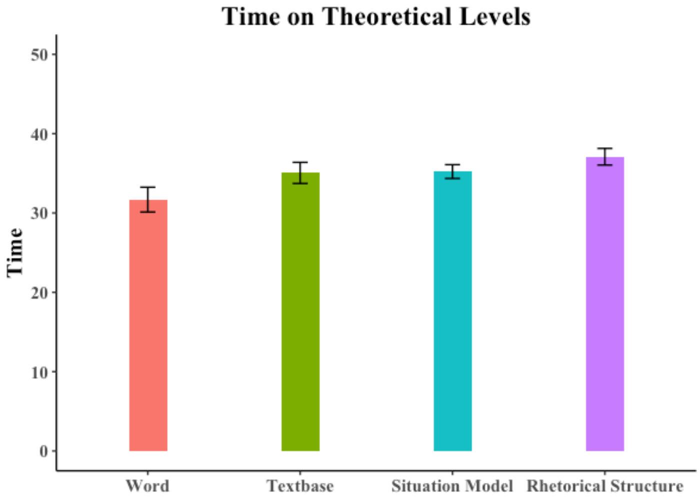


Figure 2. Adults' mean times (in seconds) to answer questions on four theoretical levels, with error bars.

Results from our logistic mixed effect model of response accuracy showed a significant difference ($\chi^2(3) = 8.34$, $p = 0.040$) in accuracy among the four theoretical levels.

Table 1 presents the output of the model. We can see that the estimated odds ratio (Estimated Odds) of *word* level is significantly higher than each of the three discourse levels (*textbase, situation model, rhetorical structure*). A post-hoc analysis with pairwise comparison showed that there was no significant difference among the three discourse levels. In contrast, results of our linear mixed effect model of suggested that time not significantly vary among theoretical levels. $F(3,25.8) = 0.058$, $p = 0.981$.

Table 1. Output of Mixed effect models on Performance and Time

		Word	Textbase	Situation Model	Rhetorical Structure
	No. of Items	1455	1981	5049	5071
Accuracy	Model Parameter	1.66	-0.588	-0.763	-0.584
	p Value	--	0.058	0.004	0.028
	Estimated Odds	1.66	1.07	0.894	1.07
Time	Model Parameter	34.3	2.23	2.84	3.15
	p Value	--	0.804	0.716	0.694
	Predicted Time	34.3	36.5	37.1	37.7

Our correlational analysis showed a significant positive correlation between mean accuracies on 29 lessons and word level ($r = .386$, $p < .05$), but this correlation did not extend to any of the discourse levels. The times showed no significant correlations among theoretical levels. The pattern of correlations reinforced the results of mixed effect models of accuracy and time. In addition, the *word* level had a significant negative correlation with each of the three discourse levels (*textbase, situation model, rhetorical structure*, with *r* values of -0.365, -0.485, and -0.567, respectively).

The results of exact binomial test with 712 correct responses out of 1044 questions showed that the proportion of correct responses was significantly greater than chance (one tail *p*-value = 0.00).

4 Discussion and Conclusion

We performed mixed effect models and correlation analysis to see if there were differences among adult learners' accuracy and response times to questions in each of the four theoretical levels. As expected, the results indicated that adult learners' performance on *word* level was higher than the three discourse levels, and correlational analysis reinforced this trend. One reason for adult learners' higher performance for *word* level items is that *word* items tend to focus on individual words or single sentences. This type of stimulus is less taxing on working memory compared to items that address deeper discourse levels, which are more time-consuming, strategic, and taxing on cognitive resources.

In a previous study [6], learning gains within the four theoretical levels were tracked by considering performance on all items (medium, easy, and hard). Results revealed learning occurred for lessons involving *rhetorical structure*, but not on other theoretical levels. This implies that learning gains may be affected by the particular time frame (i.e., within lessons versus across lessons) used for assessment, the difficulty of the words and texts, and the specific theoretical levels being used. Future work is needed to further clarify these issues.

With respect to response time, we found no difference between theoretical levels, despite a trend in the data that suggested learners were slower to respond as theoretical level increased. Part of the explanation for this apparent discrepancy may be due to the modest sample size ($N = 52$), which did not provide adequate power to detect all differences. Another reason may be disengagement—the data may have been muddied by adult learners who became bored or distracted. Identifying chunks of disengagement and either removing or controlling for these periods in our analysis may reveal relevant response time variability.

The results of exact binomial test indicated that hints and prompts significantly increased a learner's probability of correctly answering a question that he or she had previously answered incorrectly. This led us to the conclusion that the trialogues in AutoTutor did help learners.

In summary, we showed how AutoTutor can be used to assess reading ability in low literacy adults and how AutoTutor trialogues scaffold learning of reading comprehension skills. By assessing comprehension within a multi-level theoretical framework, we attempted to provide a more nuanced diagnosis of adults' reading abilities than a single overall performance score. Future research could focus on designing comprehension tests for each of the theoretical levels of the multilevel comprehension framework. The results of these tests could be used to establish target population norms for each of the six components of comprehension. Knowing the range of abilities of the target adult population could help designers develop more adaptive intelligent tutoring systems for adult literacy and provide customized learning content to low literacy adults.

Acknowledgements

This research was supported by the National Center of Education Research (NCER) in the Institute of Education Sciences (IES) (R305C120001) and the National Science Foundation Data Infrastructure Building Blocks program under Grant No. (ACI-1443068).

References

1. OECD (2013) OECD Skills Studies Time for the U.S. to Reskill? What the Survey of Adult Skills Says: What the Survey of Adult Skills Says. OECD Publishing
2. Vernon, J. A., Trujillo, A., Rosenbaum, S. J., & DeBuono, B. (2007). Low health literacy: Implications for national health policy.
3. Graesser AC, McNamara DS (2011) Computational analyses of multilevel discourse comprehension. Topics in Cognitive Science, 3 (2), 371-398

4. Graesser AC, Li H, Forsyth C (2014) Learning by Communicating in Natural Language with Conversational Agents. Curr Dir Psychol Sci 23:374–380
5. McNamara DS, O'Reilly TP, Best RM, Ozuru Y (2006) Improving Adolescent Students' Reading Comprehension with Istart. Journal of Educational Computing Research 34:147–171
6. Shi, G., Pavlik Jr., P., & Graesser, A.C. (2017). Using an additive factor model and performance factor analysis to assess learning gains in a tutoring system to help adults with reading difficulties. In X. Hu, T. Barnes, A. Hershkovitz, L. Paquette (Eds), Proceedings of the 10th International Conference on Educational Data Mining (pp.376-377). Wuhan, China: EDM Society.
7. Graesser, A.C., Cai, Z., Baer, W., Olney, A.M., Hu, X., Reed, M., & Greenberg, D. (2017). Reading Comprehension Lessons in AutoTutor for the Center for the Study of Adult Literacy. In S. Crossley and D. S. McNamara (Eds.), Adaptive Educational Technologies for Literacy Instruction (pp. 288─294). New York: Routledge.
8. Graesser, A.C., Feng, S., & Cai, Z. (2017). Two technologies to help adults with reading difficulties improve their comprehension. In E. Segers and P. Van den Broek (Eds.), Developmental perspectives in written language and literacy. In honor of Ludo Verhoeven (pp. 295-313). John Benjamin Publishing Company.
9. Bates D, Mächler M, Bolker B, Walker S (2015) Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67:1–48

Tower of Questions: Gamified Testing to Engage Students in Peer Evaluation

Nafisul Kiron, Julita Vassileva
Computer Science Department
University of Saskatchewan, Canada
ni.kiron@usask.ca, julita.vassileva@usask.ca

Abstract. In recent years, the use of gamification in various software application areas is commonly used with success. Gamification is a technique of using game rules, designs and mechanics in non-game applications. Educational testing is an area that can benefit from this technique. It can help motivate, engage and encourage learners to participate in problem-solving and testing. In this research in progress, we propose a gamified peer-testing system called "The Tower of Questions", in the form of a web-based tower defense game. Tower defense games are a subgenre of strategy games commonly found in computer, mobile, and console-based platforms. Our game is a question and answer game that the students will play with each other. Towers will be created and given to the students each time they ask questions. The students will then attack other students' towers by answering those questions. This will continue until all the towers have either been defended or conquered. We believe this testing system will engage students in testing each other constructively and challengingly.

Keywords: Gamification, Peer-evaluation, Game-based testing

1 Introduction

The active participation of students in designing class test questions can make a more engaging learning environment [1]. Students who can keep pace with what the instructor is teaching in the class have a better understanding of the course materials and do better in the tests. Similarly, students, who can envision the questions the instructor may ask in the tests have a better knowledge about the key lessons of the course. Both types of students possess the ability to process their course studies thoroughly and come up with good class test questions. Course instructors sometimes engage students in making test questions and distributing these to the class to test each other [2]. The idea is to let the students submit the questions they find challenging. This creates scope for discussions and further learning of the course materials among the students and acts as a revision technique.

Having students actively participate in class activities can benefit the entire class. One way of achieving this is by gamifying the activities. Gamification means applying the elements of games, for example, rules, aesthetics, rewards etc. to non-game applications [3]. Gamification can make activities interesting by allowing students to compete, giving them status, achievements, self-expression, rewards etc. [4][5]. The course instructor can design the activities by adding game mechanics to the tasks by using a goal-oriented system. The gameplay experience of the players can be enhanced by providing them with long and short-term goals [6]. That is, by making the activities playful and rewarding, while all the points received from those activities will accumulate to something greater like securing a position in the leaderboard.

By using gamification in testing, students can be motivated to participate in a competitive learning environment. Letting students construct their own questions from the information they receive from the course gives them the freedom in designing questions from their own conceptual structures [1]. With the rules and rewards offered by the system, the students will be able to make plans on how they would want their games to end [6]. Having students test each other playfully will make them focus on the game-goals instead of thinking it as a class test. This will also create scope for discussions, learning the details, and preparation for future tests.

2 Literature Review

A number of studies on gamification in education have shown that it can engage people in various activities by imposing game rules, game aesthetics, rewards or a combination of all of these [5]–[9].There are several successful educational and scientific services and applications utilizing gamification, for example, Khan Academy, treehouse, foldit, galaxy zoo etc. Most of the research done so far emphasized on engagement, data collection, behavioral outcomes and performance improvements by using game elements [6][9][10].

Yu et al. [1] introduced a web-based question posing system called QPPA (Question-Posing and Peer Assessment). This system allows students to construct, assess, review and practice answering questions [1]. There is a ranking list that shows the statistics of the students' performance. In another study, Yu [2] used multiple peer-assessment mode to increase question generation by students. The interaction between question authors and peers has been facilitated by a web-based system by allowing them to explain and negotiate each other's' feedbacks. Both studies focused on student-engagement and participation in generating questions.

In their study of a gamified assessment system, Kocadere and Çağlar [7] used the game dynamics, mechanics and components defined by Werbach and Hunter [11]. In their study, they found that 9 out of 11 participants preferred gamified assessment. They used game components such as question unlocking, points, leaderboard etc. in their assessment system to make it enjoyable and motivational. Attali et al. [12] studied the effect of points as a means of feedback in a gamified assessment. They performed the studies on adult and middle school participants. In their assessment, they considered the accuracy and speed of answering questions for awarding points for solving mathematical problems. They found that rewarding points might influence the efforts of the participants.

There are mixed results from studies regarding the effect of gamification on intrinsic motivation [12]. One research found that using game-like features to reward students for their performance might not be very effective in the long run and the effect of changing the incentives in the short run was inconclusive [12]. In another study where students found the peer-assessment system favorable, it was found that the sources of motivation might have come from a mixture of multiple factors like a sense of achievement, security, altruism, "challenging one's own and other's existing knowledge" etc. [1]. Kocadere and Çağlar [7] discussed both positive and negative aspects of gamified assessment grouped by themes (enjoyment, flow, motivation, learning, low anxiety, leaderboard and content unlocking). These studies were mostly focused on engagement for learning and assessment. Depending on the use of gamification and implementation of game-like motivational affordances the outcomes will vary from study to study [10]. Therefore, the use of gamification in testing for engaging students in a tower defense type peer-testing game is a solution we think is worth exploring.

3 Proposed System

3.1 Tower Defense Games

Our proposed system uses some game dynamics and mechanics from tower defense games, which are a subgenre of strategy games [13][14]. There are many variations and versions of this genre, but the basic rule is the same. In a tower defense game, players defend their towers from enemy attacks. Enemies attack the tower to conquer it. In our game, the questions asked by the players will create virtual towers, and by answering the questions other players will attack it. At the beginning of the game, players ask questions to create towers, for each question asked one tower will be created. In regular tower defense games, the tower has a health-bar that shows how many attacks it can receive before breaking down. In our case, the tower can be conquered by attacking it with the correct answer. During the gameplay, the players get gems for creating new towers and by conquering other player's towers. The leaderboard will be based on the number of gems the players earned throughout the gameplay. There is a time limit for attacking the towers after which those towers will be considered safe and cannot be attacked. However, the closer the deadline is the more damage the towers will take and the amount of reward gems will be increased accordingly.

3.2 Design and Method

"The Tower of Questions" will be a web-based game. The game mechanics are similar to that of a real tower defense game. The players and enemies are students from the class. The players use the game to post questions based on their course topics. Each student can ask multiple questions from the available question types. The question can be true-false, MCQ (Multiple choice questions) or in short answer form. Each question posted in the game will act like a tower. The other students in the class will try to attack that tower by answering it.

At the beginning of the game, the course instructor will set the number of gems available for the game. Players can earn gems by asking questions and by answering other players' questions. Throughout the entire gameplay, the number of gems earned by the players will not exceed the amount of gems set by the instructor at the beginning of the game. Each question asked by the players will deduct a fixed amount of gems from the main reserve. The main reserve is the place where all the un-earned gems are stored. If there are no gems left in the reserve, no further questions can be asked by the players. However, the instructor may increase or decrease the amount of total gems in the reserve and let the game proceed or end.

Once a student asks a question, a tower will be created virtually in the game for that question. The player who asked the question is the Lord of that tower. Then the tower will be made visible to other players. That tower will then be available for a fixed period of time to other players to attack it by answering the question. Each player can attack each tower of other players only once. During the time of the attack, other players cannot attack it. The attack consists by the attacker submitting an answer to the question, the answer is shown to the Lord of that tower for review. When reviewing the answer, the Lord will mark it correct, incorrect or partially correct. This concludes the attack. Until the Lord has marked the answer, no other players can attempt an attack on the tower by answering it. After the Lord has marked an answer fully or partially correct, that question and answer will be made publicly visible and cannot be attacked again. If it is correct, the attacker will receive a portion of the gems awarded for the creation of the tower, otherwise, the Lord keeps the gems. Each player can give partial marks up to three times, after which they have to award full marks. That means that only 3 different attackers can give partially correct answers, after that every next attempt would be marked as either "right" or "wrong". If the answer was wrong, the tower will be open for attacks again. However, if the question was not successfully answered within the fixed period, the Lord will have to answer it and then it will be made visible to public and the tower will be considered safe from all attacks. In this case, the Lord keeps the gems earned by creating the tower.

The players will continue to add new towers by asking questions and attack other player's towers by answering until all the questions have been answered, all the available gems have been earned or all the question deadlines have been reached. After that, the course instructor monitors the status of the game and the current leaderboard. During the entire gameplay, the game is moderated by assistants assigned by the course instructor. Players can report low quality or spam questions and unfairness in marking during the gameplay. The moderators will keep a watch on reported issues and keep the gameplay stable. For example, they review the leaderboard and especially the top achievers – were their questions well formulated? Did they actually have an answer? After a human review, the final leaderboard is posted. Fig. 1 illustrates a flowchart of the processes.

An example walkthrough of the game is as follows:

- A player called Lord-X posts 15 unique questions of several types in the game, thus 15 towers are created. For each question, Lord-X receives 10 gems. So, for the 15 questions, Lord-X earns 150 gems.
- The players can post as many questions as they want while there are enough gems available in the reserve. For example, there are 500 gems left in the reserve and Lord-X asked 15 questions in the game. So, Lord-X will earn 150 gems and the remaining gems in the reserve will be 350. Similarly, player Lord-Y and Lord-Z ask 15 and 19 questions respectively. Therefore, they earn 150 and 190 gems respectively and the number of remaining gems in the reserve becomes 10.
- Now that the players have some towers, they start attacking each other. Lord-X successfully attacks Lord-Y's tower by answering a true-false question. So, he receives 6 out of the 10 gems from that tower. The remaining 4 gems are for Lord-Y to keep for his contribution in building that tower. The distribution of gems for true-false and MCQ type questions are the same.

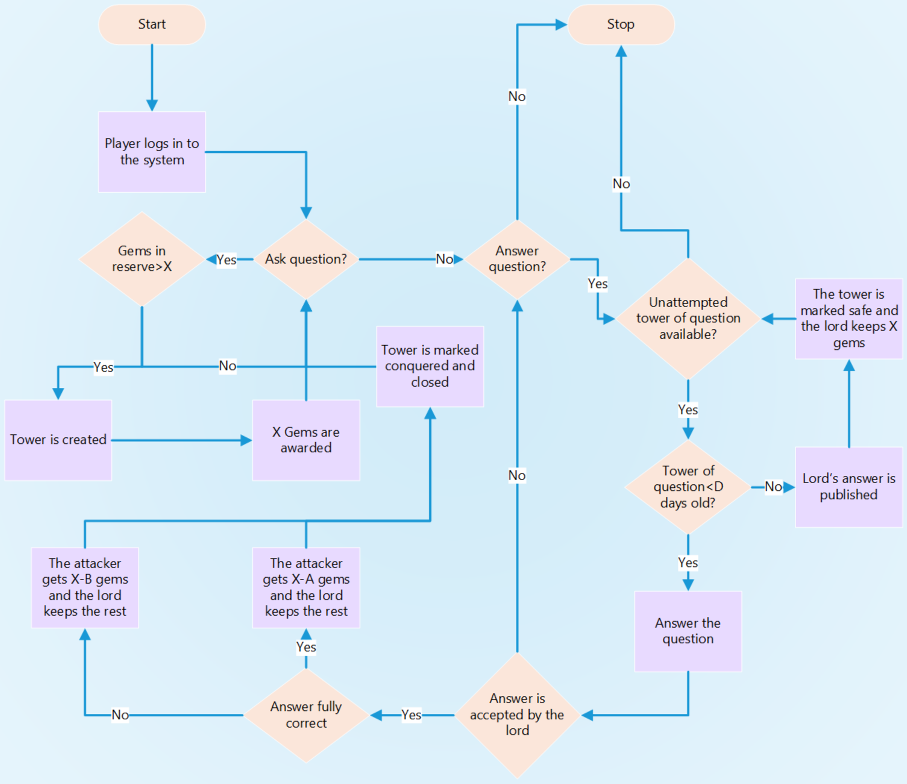

Start
Stop
Player logs in to the system
No
No
Gems in reserve>X
Yes
Ask question?
No
Answer question?
Yes
Unattempted tower of question available?
The tower is marked safe and the lord keeps X gems
Yes
No
Tower is marked conquered and closed
Tower is created
X Gems are awarded
No
Yes
Tower of question<D days old?
No
Lord's answer is published
Yes
The attacker gets X-B gems and the lord keeps the rest
The attacker gets X-A gems and the lord keeps the rest
Answer the question
Yes
No
Answer fully correct
Yes
Answer is accepted by the lord

Fig. 1. Flowchart of the game. The variables A, B, X and D are set by the course instructor.

- Then, Lord-Y attacks Lord-Z's tower with a short answer. Lord-Z found the answer to be partially correct. So, he marks it partially correct and publishes the Lord's and attacker's answer in the system. Lord-Y receives 4 and Lord-Z keeps the remaining

6 gems for that tower. The system allows the players to partially award gems to the attacker up to the number of times allowed by the teacher. For example, the system allows each player to award partial gems up to 3 times. So, the player can award partial gems for 3 towers that have partially correct answers. After that, they will have to give full marks.

- The players keep on attacking each other's tower and in the end, only a few towers are left. The reserve still has 10 gems from the original 500. Lord-X, Lord-Y, and Lord-Z each hold 146, 164 and 180 gems respectively. Lord-Y decides to ask another question, and this uses up the remaining 10 gems from the reserve. Lord-Y's new score is 174 gems. Lord-Z tries to post another question but fails because there are no gems left in the reserve.
- With what is left, they attack each other for the last stand and eventually must stop. Each of them defends a few towers completely because no one has answered those questions. So, the Lord's answer for those questions is published publicly. Since these towers are untouched, the Lords of these towers keeps the full 10 gems for each untouched tower.
- Finally, before the course instructor publishes the final leaderboard a final check for all reported question and answers are reviewed by the moderators. The moderators are actively monitoring the game throughout the entire session. They investigate situations which the players report as unfair.

4 Conclusion and Future Work

In this paper, we propose a gamified web-based game to motivate students to participate in a question and answer posing system, masked as a tower defense game. The game will allow students to test each other using the asynchronous web-based system. By rewarding gems for submitting test questions and defending the towers, we are expecting students to post high-quality questions to better defend their position in the game. The game supports learning since is requires students to think of good questions about the material themselves, which is an important aspect of active learning; it does in a playful context, and allows students to test each other in a game that, we believe, will motivate them to learn the material better and perform well in real exams.

Our future plan is to test the system with students at the University of Saskatchewan. We will evaluate the interactions within the system by counting the number of banked gems among the students and the number of questions and answers. We will evaluate separately the quality of the questions generated and the good questions will become part of the test-item bank for the class (a useful byproduct of the game). Finally, a post gameplay survey will be presented to the students to learn their level of satisfaction in using the system. We will measure the student engagement through their participation and satisfaction. We will also measure student achievement through the scores in the game, counting both the scores earned by creating questions and by answering them. We will attempt to correlate these scores with those obtained at mid-term and final exams and we expect to find positive significant positive correlations.

References

[1] F. Y. Yu, Y. H. Liu, and T. W. Chan, “A web-based learning system for question-posing and peer assessment,” *Innov. Educ. Teach. Int.*, vol. 42, no. 4, pp. 337–348, 2005.

[2] F. Y. Yu, “Multiple peer-assessment modes to augment online student question-generation processes,” *Comput. Educ.*, vol. 56, no. 2, pp. 484–494, 2011.

[3] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, “From game design elements to gamefulness,” *Proc. 15th Int. Acad. MindTrek Conf. Envisioning Futur. Media Environ. - MindTrek ’11*, no. September, p. 9, 2011.

[4] Bunchball, “Gamification 101 : An Introduction to the Use of Game Dynamics to Influence Behavior,” no. October, 2010.

[5] C. C. I. Muntean, “Raising engagement in e-learning through gamification,” *6th Int. Conf. Virtual Learn. ICVL 2011*, no. 1, pp. 323–329, 2011.

[6] F. F. Nah, V. R. Telaprolu, S. Rallapalli, and P. R. Venkata, “Gamification of Education Using Computer Games,” vol. 8018, no. July 2013, 2013.

[7] S. A. Kocadere and Ş. Çağlar, “The design and implementation of a gamified assessment,” *J. E-Learning Knowl. Soc.*, vol. 11, no. 3, pp. 85–99, 2015.

[8] B. B. Morrison and B. DiSalvo, “Khan academy gamifies computer science,” *Proc. 45th ACM Tech. Symp. Comput. Sci. Educ. - SIGCSE ’14*, pp. 39–44, 2014.

[9] P. Łupkowski and P. Wietrzycka, “Gamification for Question Processing Research – the QuestGen Game,” vol. 1, no. 7.

[10] J. Hamari, J. Koivisto, and H. Sarsa, “Does gamification work? - A literature review of empirical studies on gamification,” *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, pp. 3025–3034, 2014.

[11] K. Werbach and D. Hunter, *For the Win: How Game Thinking Can Revolutionize Your Business*. Wharton Digital Press, 2012.

[12] Y. Attali and M. Arieli-Attali, “Gamification in assessment: Do points affect test performance?,” *Comput. Educ.*, vol. 83, pp. 57–63, 2015.

[13] “Tower Defense games on Kongregate.” [Online]. Available: https://www.kongregate.com/tower-defense-games. [Accessed: 21-Mar-2018].

[14] P. Avery, J. Togelius, E. Alistar, and R. P. Van Leeuwen, “Computational intelligence and tower defence games,” *2011 IEEE Congr. Evol. Comput. CEC 2011*, pp. 1084–1091, 2011.

Exploring Gritty Students' Behavior in an Intelligent Tutoring System

Erik Erickson[1], Ivon Arroyo[1], Beverly Woolf[2]

[1] Worcester Polytechnic Institute, Worcester MA 01609, USA
[2] University of Massachusetts Amherst, Amherst MA, 01003, USA
eerickson@wpi.edu

Abstract. This research focuses on determining whether a student's GRIT impacts their behavior within an intelligent tutoring system, towards developing better student models and feature sets that can help a tutor predict student behavior and determining whether computer tutors might foster improvements in students' grit, perseverance and recovery from failure. We use rare Association Rule Mining to explore how students' grit may be associated with students' behaviors within MathSpring, an intelligent tutoring system, as a first step.

Keywords: Grit, Perseverance, Student Models, Association Rule Mining.

1 Introduction

Studies have shown that grit is more predictive of life's outcomes compared to the "Big Five" personality model, which is a group of broad personality dimensions (e.g. conscientiousness, extraversion, agreeableness, and neuroticism [15]), but unlike IQ, the previous gold-standard predictor for life outcomes, grit may not be a static quality but one that can be developed [12]. Grit has become ubiquitous in the lexicon of public schools across America [20]. Educators are looking for answers to some lingering questions: "*Can students increase their grittiness?*" and "*How do students go about doing so?*". Gritty individuals can maintain high determination and motivation for a long time despite battling with 'failure and adversity'. Students can increase their grittiness through classroom activities [20]. Educators are interested in fostering growth in children, and would be interested in fostering grit in their students.

Our research focuses on how a student's grit and perseverance might impact behavioral patterns in a tutoring system, towards understanding how digital tutors might foster gritty-like behaviors, and in turn, grit assessments.

We move research on grit forward as a tool to refine student models in intelligent tutoring systems, by answering the following questions:

RQ#1. *Can we predict if a student is gritty or not by looking his/her behaviors? Here, grit is a target to predict, or a consequence.*

RQ#2. *Does the grit of a student influence student behavior inside a tutor? In which way(s)? Here grit is a cause or antecedent*

2 Method

Grit has typically been assessed using Duckworth's instrument of the Grit Scale [13], asking students to report on twelve Likert-scale questions. Some examples of questions are, "*I often set a goal but later choose to pursue a different one*" and "*Setbacks don't discourage me.*"

Our testbed is MathSpring, an intelligent tutoring system (ITS) that personalizes problems by assessing students' knowledge as well as effort and affect as they engage in mathematics practice online [5-7]. Students used MathSpring during class time over several days, as part of their regular mathematics class, and solved many math problems, while the system captured detailed event-level and problem-level information on their performance. These students also filled out a grit scale survey [8] that produced in an aggregate grit score.

2.1 Data Collection and Data Mining

Seventh grade students from two school districts participated in a research study. After combining the two datasets, there were 456 rows of Grit survey responses representing thirty-eight students. Sixty-eight students used MathSpring, producing 3,012 rows of data, each representing a student-math problem interaction. Variables were discretized into Booleans, indicating high/low or true/false. We created the negation of each variable (e.g., for GUESS, we also created a counterpart NoGUESS variable with the opposite truth value) to be considered also. Along with Guess, other variables included Hi/Low Grit, is/is not Solved, Hi/Low Mistakes, Hi/Low Hints, Yes/No Finished, Not/Likely Read (the problem).

We used Association Rule Mining to discover rules, a non-parametric method for exploratory data analysis, which finds associations that occur more frequently than expected from random sampling. The four critical parameters and minimum thresholds used are the following: Support 0.05, Confidence 0.84, Lift 1.15, Conviction 1.75. Last, we subjected the most important rules to a Chi-Square statistical test, those with solely "High Grit" or "Low Grit" as a consequent or antecedent.

3 Results

The mean Grit Score for the N=38 students in the sample was *M=3.07*, *SD=0.51, Median=3, Range= [1,5]*. This means the student grit assessment had some variability but the distribution is centered on a neutral grit value. A median split was done, classifying students as low or high grit, so that half of the students were considered gritty or not. Interestingly, we found that High-Grit students had much more activity, 71% of the student-problem interactions in the dataset vs. 29% for the non-gritty students. Table 4 shows the number and percent of cases for notable variable in detail, after the discretization process.

Due to a low support threshold of 0.05, thousands of rules were created. Only a selected subset of rules was chosen for interpretation, mainly those rules with a single consequent or antecedent, and those which met thresholds and had highest values for the metrics of confidence, conviction and lift.

Table 1. Name, number of Cases and Percent Cases for all Variables in the final dataset

Variable Name	N cases	% High (or True)	Counterpart Variable	N cases	% High (or True)
HiGrit	2146	71.25%	LowGrit	866	28.75%
GUESS	368	12.22%	NoGUESS	2644	87.78%
DNFINISH	261	8.67%	FINISHED	2751	91.33%
NOTREAD	86	2.86%	LIKELYREAD	2926	97.14%
isSolved	1655	54.95%	NotSolved	1357	45.05%
HiMistakes	1343	44.59%	LowMistakes	1669	55.41%
HiHints	822	27.29%	LowHints	2190	72.71%

A notable finding was that no rules with *LowGrit* as a consequent appeared at all according to our criteria specified in the parameter thresholds. This made us realize that, due to the much lower number of math problems seen by Low Grit students, the *confidence* for any rule with *LowGrit=1* as a consequent would be at chance level at 0.288 (as opposed to 0.5). We realized how the *confidence* metric is not very reliable in this case due to the imbalanced dataset. On the other hand, the metric that balances the rarity of the premises of a rule and their confidence is the '*conviction'* parameter. We thus set conviction as our first priority for selection of rules.

Table 5 shows the rules that had the highest conviction, confidence and lift. These rules also are the most complete rules (as generally subsequent rules that met the parameter thresholds had similar premises, but combined subsets of the propositions). **Rule A** is the rule with highest confidence, conviction and lift, and states that *if a student made a high amount of mistakes in a math problem, and asked for many hints as a way to help them solve the problem, then it means the student has a high level of Grit*. This joint condition happened in 19% of the total student-problem interactions examined. The significance of the effect for each rule was verified with a Chi-Square test by computing cross-tabulations between the premise being true/false vs. High/Low Grit (p<0.0001 for rules 1, 2, and 3).

Table 2. Grit as a Consequent: Association Rules with highest Conviction, Confidence, Lift

Rule	**Confidence**	**Conviction**	**Lift**	**Support**
Rule A. HiMistakes ^ HiHints → HiGrit *	0.89	2.56	1.25	0.19
Rule B. LowMistakes ^ isSolved → Low Grit *	0.45	1.29	1.56	0.10

* Significant difference at p<0.0001, χ^2 (1, N=3012)

On the other hand, no rules were found that met the thresholds of confidence, lift and conviction for *LowGrit* as a consequent. Still, we show the rule that has the best outcome for those metrics. The implication *LowMistakes ^ isSolved → Low Grit* has a confidence level of 0.45, which is low, however, it is higher than chance as stated earlier (chance level for any *LowGrit* row is 0.288). The rule suggests that if a student solves problems by making a low number of mistakes, then the student is NOT gritty.

Table 6 summarizes the found rules with Low/High Grit as a premise. This time, it was easier to find rules with *LowGrit* as an antecedent that met the thresholds of confidence, lift and conviction but not for *HiGrit*. Rule C is the main rule found for Low Grit as an antecedent (other similar rules are variations of this same effect), suggesting that if a student has low grit, then they will likely ask for few hints in a problem.

The rule that contains *HiGrit* as an antecedent is Rule D. While Rule D does not meet the lift and conviction thresholds we had set, it does meet the confidence threshold, and is the rule found with the highest values of confidence and conviction. This rule captures that *if a student is gritty, then the student will not quick-guess the correct answer to a problem.* Remember that guessing implies that a student entered many answers incorrectly and did not ask for help/hints, until they manage to solve it correctly (the multiple-choice format in most questions in MathSpring probably favors this type of disengagement behavior in general). We consider that students who guess are avoiding help when they should instead be asking for it, as they are answering incorrectly, as stated in previous research [1,2]. Rushing to get the right answer without fully understanding why, and avoiding seeking help.

Table 3. Grit as an Antecedent: Association Rules with highest Conviction, Confidence, Lift

Rule	**Confidence**	**Conviction**	**Lift**	**Support**
Rule C. Low Grit → Low Hints *	0.88	2.27	1.21	0.18
Rule D. Hi Grit → NoGUESS *	0.89	1.16	1.02	0.7

* Significant difference at p<0.0001, χ^2 (1, N=3012)

4 Discussion

This research starts unpacking how grit may be expressed in student behaviors inside an intelligent tutor, and on learning how fostering gritty-like behaviors might eventually improve a students' grit. In general, the results of Association Rule Mining suggest that there are differences students' behaviors depending on their assessed level of grit. Apparently, students who are gritty tend to neither quick-guess answers to problems, nor making lots of mistakes while avoiding help. At the same time, rules found with grit as a consequent suggest that if a student is in a situation of conflict, making mistakes but resolving them by asking for hints (or videos or examples), we can predict that the student has high grit. This is a desirable behavior when facing challenge in interactive learning environments, as specified by a review on help seeking and help provision in interactive learning environments [2].

It was harder to find Association Rules that associate students with low grit with behaviors (there are not as many systematic behavior patterns that could be associated to students of low grit). Still, the few rules found suggest that when a student has low levels of grit, they will seek for a low amount of hints. Conversely, the behavior that a student is NOT gritty is that he/she makes a low number of mistakes and eventually solves the problems correctly. Given the agency that MathSpring allows (more than most other learning environments) this does not necessarily mean that low-grit students tend to solve problems correctly (otherwise solve-on-first would have been part of the rules found). Students who skip problems or give-up will receive easier problems in an adaptive tutor. Also, students could choose material that is easier, or already mastered, to guarantee higher levels of success. Further analyses could help discern if this is the case, by analyzing the level of difficulty of the problems students received. Grit is a construct that will predetermine students to have different kinds of self-regulatory behaviors while learning in interactive learning environments.

Acknowledgements

This material is based upon work supported by the National Science Foundation under NSF award #1551594 INT: Collaborative Research: Detecting, Predicting and Remediating Student Affect and Grit Using Computer Vision. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Aleven, V., McLaren, B., Roll, I., and Koedinger, K., Toward Meta-Cognitive Tutoring: A Model of Help-Seeking with a Cognitive Tutor. International Journal of Artificial Intelligence in Education (IJAIED), 2006. 16(2): 101- 128.
2. Aleven, V., Roll, I., McLaren, B., Ryu, E. J., and Koedinger, K. (2005) An Architecture to Combine Meta-Cognitive and Cognitive Tutoring: Pilot Testing the Help Tutor. Artificial Intelligence in Education: Supporting Learning Through Intelligent And Socially Informed Technology, 125: 17.
3. Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R.M. (2003). Help Seeking and Help Design in Interactive Learning Environments. Review of Educational Research, 73(2), 277-320
4. Arroyo, I. and Woolf, B. (2005) Inferring Learning and Attitudes from a Bayesian Network of Log File Data. In Proceedings of the 12th International Conference on Artificial Intelligence in Education, 33-40.
5. Arroyo, I., Beal, C., Murray, T., Walles, R., and Woolf, B. P. (2004) Web-Based Intelligent Multimedia Tutoring for High Stakes Achievement Tests Intelligent Tutoring Systems. In Proceedings of the 7th International Conference on Intelligent Tutoring Systems (ITS'04), Maceo, Brazil, 142-169.
6. Arroyo, I., Woolf, B. P., and Beal, C. R. (2006) Addressing Cognitive Differences and Gender During Problem Solving. International Journal of Technology, Instruction, Cognition and Learning, 4: 31-63.
7. Arroyo, I., Woolf, B. P., Burelson, W., Muldner, K., Rai, D., & Tai, M. (2014). A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *International Journal of Artificial Intelligence in Education*, *24*(4), 387-426.
8. Corno, L., & Snow, R. E., Adapting teaching to individual differences among learners, in Handbook of research on teaching, M.C. Wittrock, Editor. 1986, MacMillan: New York.
9. Craig, S., Graesser, A., Sullins, J., and Gholson, B., Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. Journal of Educational Media, 2004. 29(3): 241-250.
10. Csikszentmihalyi, M., Flow: the psychology of optimal experience. 1st ed. 1990, New York: Harper & Row. xii, 303 p.

11. D'Mello, S. and Graesser, A., Mind and body: Dialogue and posture for affect detection in learning environments, in Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work. 2007, IOS Press. p. 161-168.
12. Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. Journal of Personality and Social Psychology, 92 (6), 1087-1101.
13. Duckworth, Angela & Quinn, Patrick. (2009). Development and validation of the Short Grit Scale (GRIT-S). Journal of personality assessment. 91. 166-74. 10.1080/00223890802634290.
14. Eskreis-Winkler, L., Gross, J. J., & Duckworth, A. L. (2016). Grit: Sustained self-regulation in the service of superordinate goals. *Handbook of self-regulation: Research, theory and applications. New York, NY: Guilford.*
15. Eskreis-Winkler, L., Shulman, E. P., Beal, S. A., & Duckworth, A. L. (2014). The grit effect: Predicting retention in the military, the workplace, school and marriage. *Frontiers in psychology*, *5*.
16. Goleman, D., Emotional Intelligence: why it can matter more than IQ. Bloomsbury. 1996, London.
17. Martin, N. and Halperin, S., Whatever it takes: How twelve communities are reconnecting out-of school youth. 2006, American Youth Policy Forum: Washington, D.C.
18. Pekrun, R., Goetz, T., Daniels, L., Stupinsky, R., and Perry, R., Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. Journal of Educational Psychology, 2010. 102(3): 531-549.
19. Reed, J., Pritschet, B. L., & Cutton, D. M. (2013). Grit, conscientiousness, and the transtheoretical model of change for exercise behavior. *Journal of health psychology*, *18*(5), 612-619.
20. Saltman, K. J. (2014). The austerity school: Grit, character, and the privatization of public education. *symploke*, *22*(1), 41-57.
21. Schrum, L., & Levin, B. B. (2009). *Leading 21st-century schools: Harnessing technology for engagement and achievement*. Corwin Press.
22. Woolf, B., Arroyo, I., Muldner, K., Burleson, W., Cooper, D., Dolan, R., and Christopherson, R. (2010) The Effect of Motivational Learning Companions on Low Achieving Students and Students with Disabilities. In Proceeding of the10th International Conference on Intelligent Tutoring Systems (ITS'10), Pittsburgh, PA, 327-337.

Disengagement Detection Within an Intelligent Tutoring System

Su Chen[1,2], Anne Lippert[1,2], Genghu Shi[1,2],Ying Fang[1, 2] and Arthur C. Graesser[1, 2]

[1] University of Memphis, Memphis TN 38111, USA
[2] Institute for Intelligent Systems, Memphis, TN
schen4@memphis.edu

Abstract. This paper describes a novel automated disengagement tracing system (DTS) that detects mind wandering in students using AutoTutor, an Intelligent Tutoring System (ITS) with conversational agents. DTS is based on an unsupervised learning method and thus does not rely on any self-reports of disengagement. We analyzed the reading time and response accuracy of 52 low literacy adults who interacted with AutoTutor to learn reading comprehension strategies. Our results show that students completing a lesson with 20 questions tend to start mind wandering at the 11th ~15th question. Question chunks with mind-wandering have an accuracy of 20%, in contrast to 70% in accuracy for non-mind wandering.

Keywords: CSAL AutoTutor, Mind Wandering, Disengagement.

1 Introduction

In many respects, intelligent tutoring systems (ITS) live up to their reputation as "next generation" learning environments. Well-designed ITSs are technology driven, automated, and offer a personalized and adaptive instruction that is difficult, if not impossible to implement in a traditional classroom setting. In other respects, ITSs are no more advanced than human instructors when it comes to challenges in student learning. For instance, both ITS designers and human teachers struggle with how best to keep learners focused, interested, and stimulated by material. Regardless of whether they learn from an ITS or in a classroom, students are likely to become disengaged due to various reasons such as fatigue, distraction by environment, loss of interest or falling behind in a course. Though there have been efforts by ITS designers and developers to make systems more generally attractive and interactive to users (Graesser, Cai, Morgan, & Wang, 2017), it is likely that effective interventions will need to be personalized. Studies have only recently been conducted with personalized interventions to prevent or interrupt disengagement activities and guide an individual learner back on track (D'Mello & Graesser, 2012). A critical component of such an intervention is an ITS built-in disengagement tracing algorithm which can capture "mind-wandering"(MW) promptly and accurately.

We define MW to be the disengagement of attention from an assigned task, which is largely involuntary and related to “off-track” behaviors such as boredom and distraction. Besides leading to low performance, MW can present a problem for researchers because it may contaminate the actual reading time (or time spent on one question) and thus confound the true signal/pattern in the data. MW students usually take too long (thinking about something irrelevant to the reading task) or too short (quickly finish the session without comprehension) on one question chunk (i.e. a chunk that a student spends on one question). A disengaged reader is extremely slow or fast with low performance, depending on how readers handle the frustration of underperforming. Data analyzed without addressing the abnormal reading time due to MW may lead to unreliable and misleading results. It is well established that MW is negatively related to reading comprehension (Mills, Graesser, Risko&D'Mello, 2017).

Existing MW detection methods applied supervised learning approaches to train models using self-reported MW (Mills, Graesser, Risko&D'Mello, 2017). The participants are probed during reading with a stimulus signal, upon which they report whether or not they are MW. Self-reported MW is not always available for a concurrent disengagement monitoring system; such self-reports are collected at the end of training sessions and used for post-hoc research. However, these judgments may have a response bias to the extent that disengaged students may feel guilty and prefer not to admit that they have been MW. Beck (2005) proposed an approach using item response theory to detect whether a student is engaged in answering questions. The estimated probability of disengagement depended on the response time and accuracy of the responses. However, Beck’s method requires a reasonably large sample size to build a model that accounts for inter-student and -question type variability since a large number of parameters were introduced. Apparently, the required size is difficult to obtain, even for Beck, who was unable to test the approach due to insufficient data.

In this paper, we propose an unsupervised self-learning algorithm to monitor whether a student is engaged in answering questions within AutoTutor lessons. Disengagement is measured in terms of the time that a student spends on a question, as well as his or her relative short-term performance. Disengaged students tend to spend too long or short time on a particular question and thereby perform poorly on the question. The algorithm utilizes the first 3 to 5 well-performed questions to learn a student’s pace in a specific lesson and then tracks his/her learning process for questions for which they exhibit disengagement.

2 Description of CSAL Auto Tutor

CSAL AutoTutor is a derivative of AutoTutor developed to help adult learners with low literacy skills improve reading comprehension as part of an intervention led by the Center for the Study of Adult Literacy (CSAL, http://csal.gsu.edu). AutoTutor teaches comprehension strategies by holding conversations called “trialogues” between two computer agents (a tutor and peer) and the human student (Graesser, Li, & Forsyth, 2014; Lehman & Graesser, 2016). The 35 lessons of AutoTutor focus on one or more specific theoretical levels of reading comprehension. The lessons are adaptive in the sense that they present reading material of varying difficulty depending on the student's performance. Typically, the system will first present students a medium level text and ask 8-12 questions about the text. Depending on students' performance on the questions, they will subsequently get a hard (if above a threshold) or easy (if below a threshold) level text and assessment (Graesser, Feng, and Cai, 2017). Some lessons only provide one medium level text followed by up to 30 questions.

3 Method

3.1 Participants and Design

Participants were 52 adult students from literacy classes in Atlanta and Toronto. They completed a 100-hour intervention over four months. Their ages ranged from 16–69 years (M = 40, SD = 14.97) and 73.1% were female. All participants read at 3.0–7.9 grade levels. On average, the 52 participants completed 23 lessons (ranging from 2 to 29 lessons1), and each lesson contained 14.6 questions (medium level) ranging from 6 to 30 questions. The lessons were scaled on different levels of text and discourse analysis. Specifically, Graesser and McNamara's multilevel theoretical framework of comprehension specifies six theoretical levels: word (W), syntax (Syn), the explicit textbase (TB), the referential situation model (SM), the genre/rhetorical structure (RS), and the pragmatic communication level. AutoTutor taps all of these levels except for syntax and pragmatic communication. The 29 lessons were assigned a primary level (but typically had a secondary or even tertiary level, but these were not considered in this paper). The word level addresses topics such as word meaning clues, learning new words, and multiple meaning words. TB lessons focus on pronouns, punctuation, and main ideas. The SM lessons concern connecting ideas and making inferences from text, whereas RS lessons cover the structure of different genres, such as steps in procedures and problems and solutions. Of the 29 lessons, only 12 provide a single medium level text assessed by 15 to 30 questions. The other 17 lessons start with a medium level text (~15 questions) and then branch to an easy/hard level text according to a student's performance on the first text. The counts of lessons from each theoretical level and branch status are provided in Table 1.

3.2 Disengagement Tracing Algorithm

A disengagement tracing system (DTS) in AutoTutor is expected to automatically learn a student's reading ability and set it as a reference of the participant for disengagement detection. Capturing behavior that as "off-track" will allow us to identify whether a student is "mind-wandering"(MW) on a specific question. The amount of time a student takes to respond to a question, namely "response time" (RT) can be used to determine when a student is off-track. MW students will involuntarily shift

[1] 6 of the 35 AutoTutor lessons were not in the scope of the intervention curriculum so students did not receive these lessons.

Table 1. Distribution of Theoretical Levels Across the 29 lessons (Number of lessons)

Theoretical Level	W	TB	SM	RS
One Text	1	1	6	4
Two Texts (Branch to easy/hard)	3	4	5	5

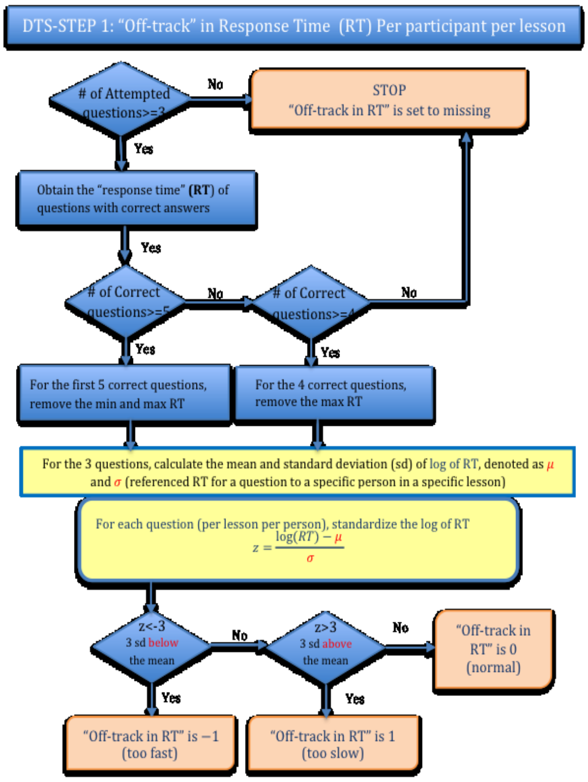

DTS-STEP 1: "Off-track" in Response Time (RT) Per participant per lesson
# of Attempted questions>=3
No
STOP
"Off-track in RT" is set to missing
Yes
Obtain the "response time" (RT) of questions with correct answers
Yes
# of Correct questions>=5
No
# of Correct questions>=4
No
Yes
Yes
For the first 5 correct questions, remove the min and max RT
For the 4 correct questions, remove the max RT
For the 3 questions, calculate the mean and standard deviation (sd) of log of RT, denoted as μ and σ (referenced RT for a question to a specific person in a specific lesson)
For each question (per lesson per person), standardize the log of RT
$z = \frac{\log(RT) - \mu}{\sigma}$
z<-3
3 sd below
the mean
No
z>3
3 sd above
the mean
No
"Off-track in RT" is 0 (normal)
Yes
Yes
"Off-track in RT" is −1 (too fast)
"Off-track in RT" is 1 (too slow)

Fig. 1. DTS-Step 1: "off-track" in response time (RT)

attention away from the targeted task towards task-unrelated thoughts and comprehension is likely to suffer. When we assume students' reading abilities are unlikely to improve or worsen in a short time period, one indication that students are off-track is if they try to compensate for a lack of comprehension by answering "too fast" or "too slow" (relative to their personalized "normal" RT) on a question. **The DTS algorithm consists of two steps:** the first step (illustrated in Fig. 1) in detecting disengagement on a question is to identify off-track in RTs. To this end, we make two assumptions: (1) students tend to be engaged at the beginning of a lesson when answering the first few questions and (2) if a student correctly answered a question, he/she most probably was engaged. This means the average time a student spends on the first few correctly answered questions of a lesson reflect RTs while the student is "on-track" or engaged. However, what is on-track for one student may be off-track for another, Furthermore, an individual's reading ability may vary depending on the characteristics of the texts (e.g. difficulty, type) included in each lesson. Because of these sources of variation, it is necessary to establish multiple baselines or reference behaviors for each learner in each lesson. Therefore, we extract a set of references, or a "reference library" of RTs for each participant for each lesson from the log files of AutoTutor, which contain this information. Specifically, we computed the mean and standard deviation of the log of RT of the first m correctly answered questions ($\mathrm{m} = 5$ in this study) and treated it as a "reference RT" for a certain individual at a specific lesson. It is possible that one question is answered correctly by accident. To take this into consideration, we dropped the highest (and lowest) reading time before calculating the benchmark statistics. If the student has less than 3 (correctly answered) questions, the algorithm lacks information to learn for the reference library and will return "missing" until the student answers enough questions correctly. Response time is naturally right-skewed. Our numerical study shows that a log transformation takes the data to a normal distribution. Given the normal distribution, the "3-standard deviation rule" applies. Once the reference library is created, we can say 'a student is off-track on a question (too fast or too slow)' if the log of reading time is below or above 3 standard deviations from the reference engaged data sample.

Disengagement detection only based on response time would lead to a large number of "false positive". Some lessons start with a very easy or "confidence-boosting" question, which means learners will respond more quickly to this question than others with high accuracy. Disengaged students usually perform poorly since they are not focusing on the question. However, a student with an overall accuracy of 80% for a lesson may still answer 3 questions incorrectly in a sequence and take more time than usual to do so. This indicates a high chance that this student is off-track while working on these 3 questions. Some questions in a lesson are very straightforward (or complicated). Students may take significantly less (or longer) time than their reference engaged time. Our target MW questions are those with off-track response times, poor local performance, but possibly adequate overall performance. Overall performance of a lesson per participant is measured by the overall correct proportion for the lesson. Local performance of a question per participant is given by moving average of correctness proportion. The k^{th} order moving average of t^{th} question is given by $\frac{\sum_{\mathrm{i=t-k}}^{\mathrm{t+k}} \mathrm{X_i}}{2\mathrm{k}+1}$, where $\mathrm{X_i}$ is 1 if the i^{th} question is correctly answered and 0 otherwise. In this study, we take $\mathrm{k} = 1$. Step 2 of DTS refines results from Step 1 by filtering out well-performed questions for students who spent too long (or short) time on a questions.

4 Results

We applied the proposed DTS algorithm to the data extracted from AutoTutor (18,863 question-chunks, 52 participants) and identified 900 mind-wandering question-chunks from 51 participants. We were interested in, first, which "questionID"s (questions are answered sequentially) in a lesson are most likely to lead to disengagement? Second, do the patterns of MW differ across the four theoretical levels? We plotted the proportions of MW by each "questionID" for lessons in each of the four theoretical levels

(Fig. 2). The number of question chunks is different for each "questionID". For example, there are more observed question chunks in Question#1 than Question#12 due to the facts that (a) some lessons have less questions than others or (b) some students did not complete all the questions in a lesson. DTS algorithm assumes that the response time of questions within one lesson is from the same distribution. We are mainly interested in differences of MW pattern between theoretical levels although response time may vary between lessons within a theoretical category. In Fig. 2, we also plotted the frequency of question chunks for each "questionID". Fig. 2 suggests different trends in MW for the different theoretical levels. In general, an increasing number of MW is observed as "questionID" goes from 1 ~ up to 30.

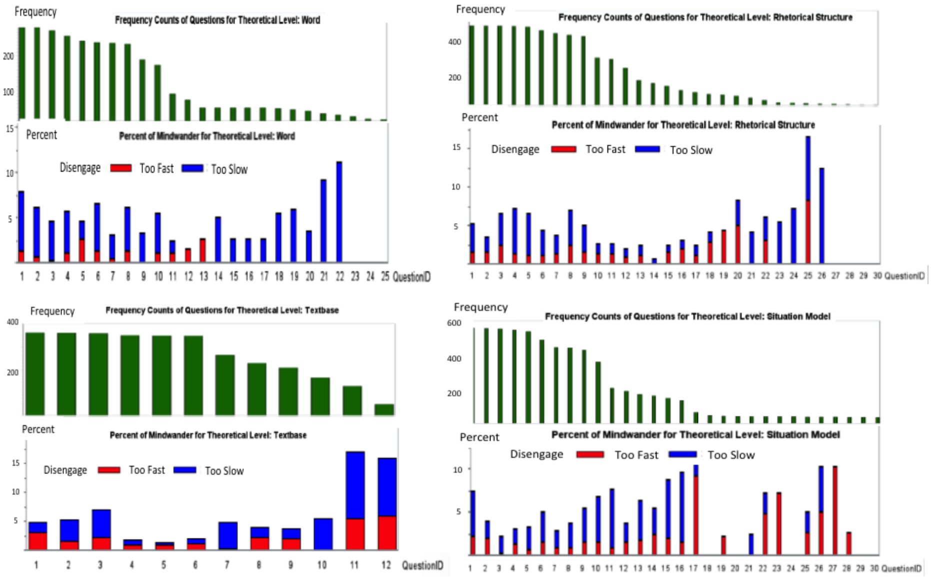


Fig. 2. Disengaged proportion versus question ID at four theoretical levels

Higher proportions of question chunks are identified as "disengaged" in terms of response time and performance for larger "questionID", which coincide with common sense. Students may get tired. Surprisingly, there is a small peak (in MW rate) at the first question of lessons. However, we note the first question is a special case, which may not truly reflect disengagement. For instance, participants may require additional time to adjust to the text/ lesson or they may encounter confusion in using the technology. This appears to be the case for the TB level where the first question has a high rate of participants answering "too fast" followed by a slight drop in the disengagement rate. We also see that some theoretical levels show an increase in disengagement between Question 2 to 4, which may indicate that students were learning skills or getting familiar with the content. For lessons at the SM level, students required a longer learning period (too slow rate increased until Question 6). For all levels it appears that after five to seven questions students gradually gained necessary skills for the lesson, and disengagement decreased until 11th ~15th question, after which it increased again. The question chunks with high MW rate before 11th question should not be considered as "true" disengagement. Students tended to start MW at the 11th ~15th question, after which engagement rates steadily increased. Participants may have felt fatigue or got bored, which could slow their speed in problem solving or induce quick answers without deep thinking. To our surprise, the disengagement rate of the **last** 1~3 questions suddenly dropped to zero (except TB), which contradicts common sense. We checked the frequency counts of questions for each "questionID" and found that the total counts of these last 1~3 questions are very small (nearly zero). Thus, we would not be able to observe disengaged question-chunks with such a small sample size. Furthermore, we found fewer disengaged chunks after question 11 because some of the lessons in our sample contained less than 12 questions. Naturally, any contribution from these lessons to the frequency of MW becomes zero after question 11. Another explanation is that some of the students did not complete all the questions and quit in the middle of the lesson. We can see how this explanation makes sense particularly for lessons containing only one text since they may ask students up to 30 questions. It may be that giving students > 11 questions leads to boredom/fatigue or frustration (if questions are too difficult) and so they voluntarily disengage from the tutoring system. For lessons with two unique texts and ~ 12 questions per text, the story is likely to be different. When a student is presented with a second text, he or she spends extra time constructing a new mental model to make sense of this new information- similar to what occurs at the beginning of a lesson when material is first presented. We are likely to see this additional time show up as increased response times and increased MW for the first few questions pertaining to the second text. After this, the mental model is somewhat stable and response times should level out.

To determine the effectiveness of the DTS proposed in Section 3.2, we compared the accuracy of the responses given while MW versus not MW. Out of the 900 MW question-chunks, 178 (20%) questions were correctly answered. In contrast, 12,657 (70%) of the 17,867 non-MW question-chunks were correctly answered. The accuracy of non-MW question-chunks is 70%, which is significantly higher than the 20% for the MW group ($\chi^2 = 40.6, \mathrm{p} < .001$). To better illustrate the power of the proposed DTS algorithm, we predicted the off-track reading time (Step 1 of DTS) by classical outlier detection method, i.e. 3 IQR (Interquartile range) rule. An extreme outlier is detected when the data is below Q1 (first quartile) $-3 * IQR$ or above Q3 (third quartile)$+3$ $*$IQR. To fairly compare the proposed DTS algorithm, we filtered out the poorly performed questions identified in Step 2 of DTS from the questions with extreme outliers in reading time. The accuracy of non-MW versus MW questions was 69% and 55% respectively, indicating our DTS algorithm performs better in predicting MW question-chunks.

5 Discussion and Summary

This paper provides an intelligent self-learning algorithm to monitor student engagement during instruction. The algorithm learns a student's baseline reading ability from his/her first 3~5 well performed questions in a specific lesson and then creates a personalized reference RT. An off-track question chunk is identified if abnormal deviation from the reference is found. The proposed method does not require any self-reported MW evaluation from the participants and can provide disengagement feedback promptly during the lesson. Furthermore, the proposed algorithm is simple and fast, which makes it amenable for use on projects with massive data. The DTS algorithm assumes that questions in a lesson are similar/exchangeable in terms of difficulty and context. Additional adjustments are needed if questions in a lesson are designed to be in different levels. In addition, DTS may report "false disengagement" in the first 10 questions. Users should be cautious in interpreting the early signal of "disengagement" by DTS algorithm.

Disengagement/MW detection and monitoring is critical in improving the efficiency of intelligent tutoring systems. Feedback from the proposed disengagement monitoring system can elucidate factors that lead to distractions. Accordingly, effective interventions can help engage the off-track learner at the right time. For example, once the disengagement is identified, a pop-up window with a kind reminder like "It seems like that you are mind wandering. Do you need a break? Or would you like to read more details about XX?" Or we could have the agents say something shocking when mind wandering is detected. Then users will turn their attention back to the lesson. These types of human-like interactions can be integrated into ITS to grasp the user's attention. The DTS technique "cares" about the student in that it looks for situations when the student is bored or frustrated and can adapt material or prompts to the student.

References

1. Beck, J.E. (2005). Engagement tracing: using response times to model student disengagement. In Proceedings of the 2005 conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology. IOS Press, Amsterdam, The Netherlands, The Netherlands, 88-95.
2. D'Mello, S. K. & Graesser, A. (2012). AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems*, 2(4), 23: 1-38.
3. Graesser, A.C., Feng, S., & Cai, Z. (2017). Two technologies to help adults with reading difficulties improve their comprehension. In E. Segers and P. Van den Broek (Eds.), Developmental perspectives in written language and literacy. In honor of Ludo Verhoeven (pp. 295-313). John Benjamin Publishing Company.
4. Graesser, A.C., Li, H. & Forsyth, C. (2014) Learning by Communicating in Natural Language with Conversational Agents. Curr Dir Psychol Sci 23:374–380
5. Graesser, A.C., Cai, Z., Morgan, B., & Wang, L. (2017). Assessment with computer agents that engage in conversational dialogues and trialogues with learners. *Computers in Human Behavior*, 76: 607-616.
6. Lehman, B., & Graesser, A.C. (2016). Arguing your way out of confusion. In F. Paglieri (Ed.), The Psychology of argument: Cognitive approaches to argumentation and persuasion. London: College Publications.
7. Mills, C., Graesser, A., Risko, E. F., & D'Mello, S. K. (2017). Cognitive coupling during reading. *Journal of Experimental Psychology: General*, 146(6), 872.

Assessments That Care About Student Learning

Stephen E. Fancsali & Steven Ritter

Carnegie Learning, Inc., Pittsburgh PA 15219, USA
{sfancsali, sritter}@carnegielearning.com

Abstract. We argue that an important requirement of assessments that care is that they focus on student learning. Intelligent tutoring systems (ITSs) are a basis for such assessments; they provide a means by which to continually assess what students know *as they learn*. Given widespread dissatisfaction with high-stakes assessments, we present a review of recent work targeted at replacing high-stakes exams with regular use of an ITS. We conclude by discussing some areas for future research and development.

Keywords: Intelligent Tutoring Systems, Mathematics Education, High-Stakes Testing, Formative Assessment, Summative Assessment, Instruction-Embedded Assessment.

1 Introduction

1.1 Characteristics of Assessments that "Care"

John Self's [1] description of ITSs as systems that "care" about students focused on the way that the personalization in such systems allows them to care about students in a way that other systems cannot. With respect to caring assessments, we agree with Zapata-Rivera [2] that personalization can enable assessments to address students at their individual level of understanding. Personalization in caring assessments might also enable students to demonstrate their knowledge in different ways and, perhaps, at different times. However, the most important characteristic of a caring assessment is not a result of personalization but of the goal of the assessment. For an assessment to be "caring," the experience must be beneficial to the student. Summative assessments are typically, though not always, designed to benefit institutions by providing them with information about the effectiveness of some aspect of instruction (e.g., the teacher, institution, or materials). Students are merely measurement instruments in this process. In contrast, caring assessments are fundamentally formative and directly assist the students in learning.

We posit that an exciting opportunity exists wherein ITSs, augmented by several tools and affordances that still need to be developed, are used *as caring assessments*. Such assessments are fundamentally formative, focused on student learning, and adaptive to student differences, but they also can serve a summative purpose to the institution.

In what follows, we argue that the time has come, both technologically and politically, to push forward with innovative approaches to assessment that use technologies like ITSs, embedded within the learning process, to provide continual, on-going, formative assessment *while students learn* to replace high-stakes, end-of-year summative assessment approaches. Accomplishing this goal, relying on systems like ITSs that attend to Self's notion of "caring" about students (e.g., by having a student model of what learners know and do not know *during the learning process*), will better allow a broad swath of educators, courseware and ITS developers, and others to (eventually) bask "in the positive glow associated with the term" caring [1]. More importantly, innovative approaches will increase instructional time, provide better measures of what students actually know, and improve learning outcomes. We detail recent work in developing statistical models that predict students' end-of-year test scores in mathematics using data from an "ITS that cares," namely Carnegie Learning's MATHia ITS, based on its Cognitive Tutor technology [3].

1.2 The Problem(s) with High-Stakes, Summative Assessments

High-stakes summative assessments, by design and implementation, often contradict what we know to be beneficial to instruction [4]. The fact that only the student's knowledge on the particular day of the test is important leads to cramming, which optimizes short-term performance, at the expense of long-term memory [5,6]. Item Response Theory (IRT) assumes that student knowledge is fixed for the period of the exam, and so the examination environment is set up to minimize student learning (even though we do know that prompted memory retrieval, as practiced in tests, does improve learning [7]).

Most high-stakes assessments only provide coarse measures of learning like multiple-choice items, which, even when well designed (e.g., with demonstrated validity and reliability), provide minimal opportunities to illuminate student misconceptions or the extent to which learners have mastered particular micro-competencies, skills, or knowledge components (KCs [8]).

In addition to the aforementioned shortcomings, standardized, high-stakes, summative assessments crowd out instructional time. Not only does taking the tests take time, but also teachers often spend several instructional periods (and in many cases weeks' worth of instructional time) preparing for such high-stakes assessments. Further, there are often numerous tests given. The Council of Great City Schools reports that, among large school districts recently surveyed in the U.S., the typical eighth grader, in a typical academic year, spends 25.3 hours taking 10.3 *district-administered* tests, which alone would consume 2% of instructional time in a 180 instructional-day academic year, without accounting for preparation time and other summative assessments [9].

1.3 Responses to the Problem(s)

Public backlash to perceived and actual shortcomings of high-stakes, standardized testing reflects perceptions that testing takes up too much instructional time while not being well-aligned to such instruction [10]. On a national level in the U.S., the Every

Student Succeeds Act (ESSA) encourages innovative assessment approaches, demonstrating recognition that the existing framework is less than satisfactory. At a state and local level, so-called “opt-out” movements [11] have led to parents and students exercising their rights to not be required to take certain high-stakes, standardized assessments. As we noted in [12], in 2017, 27% of students in the U.S. state of New York opted out of high-stakes math testing [13], and so many students in Minneapolis recently opted-out of state exams for 10th and 11th grade math that the state does not believe that the exam results can be judged to be reliable [14]. Officials and legislators in Georgia (and elsewhere) are presently working to pursue possible alternatives to high-stakes, end-of-year assessments via possibilities like more frequent, formative assessments via short quizzes and other possible alternatives [15]. What these response tend to have in common is a recognition that accountability and assessment of learning and knowledge are important but that the methods presently employed to assess such learning and knowledge are inadequate.

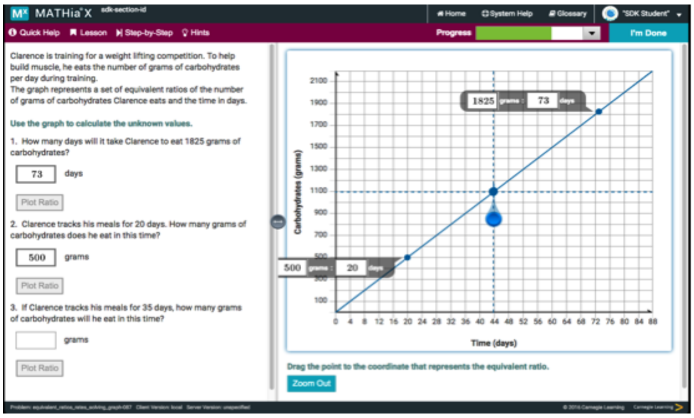
1.4 MATHia & Cognitive Tutor

MATHia is an ITS for middle school and high school mathematics, based on Carnegie Learning’s Cognitive Tutor technology, that typically is a part of a blended mathematics curriculum. Carnegie Learning generally recommends that the instructional mix of this blended curriculum be a 60%-40% split between instructor-facilitated, student-centered classroom activities that facilitate collaborative learning and deep conceptual understanding (60% of the time) and individual student work in a computer lab or classroom with the MATHia ITS (40% of the time).

MATHia is based on an adaptive, mastery learning [16] approach and relies on a fine-grained model of KCs (e.g., Grade 6 mathematics comprised of approximately 700 KCs) that students must master to make progress through content. Content is presented to students in topical “workspaces,” each of which focuses on a set of KCs that must be mastered to move on to the next workspace. Within each workspace, students work on multi-step, complex, real-world problems (see Fig. 1), and student responses at each step provide rich data about student problem-solving strategies and a fine-grained understanding of what students know and do not know.

2 Using MATHia Data to Predict Standardized Test Scores

Recent efforts [12, 17] have focused on using student MATHia performance data to predict standardized test scores in large school districts in the U.S. states of Virginia (VA) and Florida (FL). This work follows in the tradition of work using data from the ASSISTments system [18] and considers the relative contributions of various measures of MATHia performance (and transformations thereof) (e.g., workspaces mastered per hour, hints requested, errors made), prior year test performance or a pre-test score (i.e., prior knowledge), and socio-demographic data (e.g., socio-economic status via free/reduced-price lunch status, English language learner status, etc.).



Clarence is training for a weight lifting competition. To help build muscle, he eats the number of grams of carbohydrates per day during training.
The graph represents a set of equivalent ratios of the number of grams of carbohydrates Clarence eats and the time in days.

Use the graph to calculate the unknown values.

1. How many days will it take Clarence to eat 1825 grams of carbohydrates?

73 days

2. Clarence tracks his meals for 20 days. How many grams of carbohydrates does he eat in this time?

500 grams

3. If Clarence tracks his meals for 35 days, how many grams of carbohydrates will he eat in this time?

grams

Drag the point to the coordinate that represents the equivalent ratio.

Problem: equivalent_ratios_rates_solving_graph-087 Client Version: local Server Version: unspecified

Fig. 1. A screenshot of problem-solving in the MATHia platform.

Specifics of model construction, specification, and selection are beyond the scope of the present discussion (see [12, 17]), but Table 1 provides a brief summary of results to demonstrate our success so far. While various model goodness-of-fit metrics are considered in detail in the original work reporting these results, we rely on the relatively simple to interpret adjusted R^2 values of the best models for particular academic years in Table 1.

In FL, the Florida Comprehensive Assessment Test (FCAT) was used in 2013-14, and the Florida Standards Assessment (FSA) was used in 2014-15 and 2015-16. Results for FL are reported are for the best model learned on data from another academic year's data, so in each case, results reported are for the situation in which an academic year's data served as a held-out test set for the statistical model learned [12]. In VA, models were learned to predict scores on the Standards of Learning (SOL) exam for mathematics [17], but data were only available for a single academic year. R^2 values reflect the proportion of variance in SOL exam scores explained by a model learned on data for 7th graders. Cross-validation results indicated that these values do not seem to reflect substantial over-fitting.

Table 1 shows that we can account for up to 73% of the variation in FSA scores, and we see the relative contribution of different categories of variables, starting with a model including pre-test scores (M1) and progressively increasing the complexity of models through M5. Importantly, we see that there are relatively small differences between M5 and M6 (which does not include demographics), so demographic variables do not provide for substantial predictive power. Ideally, we would be able to rely on process variables (i.e., MATHia performance) only, and especially for predicting FCAT/FSA, we explain over 50% of the variation in these scores with process/performance data alone.

Table 1. Adjusted R^2 values for best linear regression models reported in [12, 17]. Variable categories are pre-test performance (*pre-test*), MATHia process data (*process*), and demographic data (*demog*). An M6 model was not considered by [17].

Model	Variables	VA SOL	FL FCAT/FSA		
		2011-12	***2013-14***	***2014-15***	***2015-16***
		n=940	n=7,491	n=7,368	n=8,065
M1	*pre-test*	.5	.6001	.6035	.6528
M2	*process*	.43	.5271	.5393	.593
M3	*process* + *demog*	.45	.5443	.5656	.6185
M4	*pre-test* + *demog*	.51	.6059	.629	.6684
M5	*pre-test* + *demog* + *process*	.57	.6642	.689	.7349
M6	*pre-test* + *process*		.6707	.6326	.7258

3 Future R&D

Being able to predict standardized test scores with reasonable success using performance data from systems like MATHia is insufficient for such systems to replace such tests. Further, systems like MATHia are designed to be used and generally, though not exclusively, are used as a part of a blended curriculum. To transition to using such systems in an assessment role, we see several important areas of R&D to pursue both for Carnegie Learning and the broader community of ITS and assessment researchers working on developing caring assessments. In addition to improving models like those for which we have here briefly reported results, we need to identify minimally sufficient sets of content that contribute to successful predictive models. This will help to identify subsets of content that should be used as a part of assessments in ITSs like MATHia. Content management, assessment design, and editing tools will be required to allow for state-by-state and possibly local customization. Security tools will be required to insure that students do their own work. More work needs to be done to establish the validity and reliability of this approach to assessment, likely by continuing to build bridges between traditional IRT approaches and the knowledge tracing approaches of systems like MATHia.

References

1. Self, J.A.: The distinctive characteristics of intelligent tutoring systems research: ITSs care, precisely. International Journal of Artificial Intelligence in Education 10, 350–364 (1999).
2. Zapata-Rivera, D.: Toward caring assessment systems. In: Tkalcic, M, Thakker, D., Germanakos, P., Yacef, K., Paris, C., Santos, O. (eds.) Adjunct Publication of the 25th Conf. on User Modeling, Adaptation and Personalization, UMAP '17, pp. 97–100. ACM, New York (2017).
3. Ritter, S., Anderson, J.R., Koedinger, K.R., Corbett, A.T.: Cognitive Tutor: applied research in mathematics education. Psychonomic Bulletin & Review 14, 249–255 (2007).

4. Snow, R.E., Lohman, D.F.: Implications of cognitive psychology for educational measurement. In: Linn, R.L. (ed.) Educational Measurement, 3rd ed., pp. 263–331. American Council on Education/Macmillan, New York (1989).
5. Bloom, K.C., Shuell, T.J.: Effects of massed and distributed practice on the learning and retention of second-language vocabulary. Journal of Educational Research 74(4), 245–248 (1981).
6. Rea, C.P., Modigliani, V.: The effect of expanded versus massed practice on the retention of multiplication facts and spelling lists. Human Learning: Journal of Practical Research & Applications 4(1), 11–18 (1985).
7. Roediger, H.L., Karpicke, J.D.: The power of testing memory: Basic research and implications for educational practice. Perspectives on Psychological Science 1, 181–210 (2006).
8. Koedinger, K.R., Corbett, A.T., Perfetti, C.: The Knowledge-Learning-Instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning. Cognitive Science 36(5), 757–798 (2012).
9. Hart, R., Casserly, M., Uzzell, R., Palacios, M., Corcoran, A., Spurgeon, A.: Student testing in America's great city schools: An inventory and preliminary analysis. Council of Great City Schools, Washington, DC (2015).
10. PDK/Gallup.: 47th annual PDK/Gallup poll of the public's attitudes toward the public schools: Testing doesn't measure up for Americans. Phi Delta Kappan 97(1), (2015).
11. Bennett, R.E.: Opt out: An examination of issues. ETS Research Report No. RR-16-13 (ETS Research Report Series). Educational Testing Service, Princeton, NJ (2016) doi:10.1002/ets2.12101
12. Fancsali, S.E., Zheng, G., Tan, Y., Ritter, S., Berman, S.R., Galyardt, A.: Using embedded formative assessment to predict state summative test scores. In: Proceedings of the 8th International Conf. on Learning Analytics and Knowledge, pp. 161–170. ACM, New York (2018).
13. Moses, S.: State testing starts today; opt out CNY leader says changes are 'smoke and mirrors.' Syracuse.com (28 March 2017). http://www.syracuse.com/schools/index.ssf/2017/03/opt-out_movement_ny_teacher_union_supports_parents_right_to_refuse_state_tests.html, last accessed 2018/03/29.
14. State of Minnesota, Office of the Legislative Auditor: Standardized student testing: 2017 evaluation report. State of Minnesota, Office of the Legislative Auditor, St. Paul, MN (2017).
15. Tagami, T.: Smaller tests could replace state's big Milestones exams. The Atlanta Journal-Constitution (02 February 2018). https://www.myajc.com/news/local-education/smaller-tests-could-replace-state-big-milestones-exams/xbdXop4VvI2Tmf6EFl7fVN/, last accessed 2018/03/29.
16. Bloom, B.S.: Learning for mastery. Evaluation Comment 1(2), (1968).
17. Ritter, S., Joshi, A., Fancsali, S.E., Nixon, T.: Predicting Standardized Test Scores from Cognitive Tutor Interactions. In: Proceedings of the Sixth International Conference on Educational Data Mining, pp. 169–176. (2013).
18. Junker, B.W.: Using on-line tutoring records to predict end-of-year exam scores: experience with the ASSISTments project and MCAS 8th grade mathematics. In: Lissitz, R.W. (ed.) Assessing and modeling cognitive development in school: intellectual growth and standard settings. JAM, Maple Grove, MN (2006).

Optimizing Human Learning

Workshop eliciting Adaptive Sequences for Learning (WeASeL)

Fabrice Popineau [1], Michal Valko [2] and Jill-Jênn Vie [3]

[1] CentraleSupélec, France
[2] Inria Lille, France
[3] RIKEN Center for Advanced Intelligence Project, Japan

Preface

This section contains the papers presented at WeASeL 2018: Optimizing Human Learning – Workshop eliciting Adaptive Sequences for Learning held on June 12, 2018 in Montréal.

Each submission was reviewed by at least 3 program committee members. The committee decided to accept 3 papers. The program also includes 2 invited talks and 1 tutorial.

What should we learn next? In this current era where digital access to knowledge is cheap and user attention is expensive, a number of online applications have been developed for learning. These platforms collect a massive amount of data over various profiles, that can be used to improve learning experience: intelligent tutoring systems can infer what activities worked for different types of students in the past, and apply this knowledge to instruct new students. In order to learn effectively and efficiently, the experience should be adaptive: the sequence of activities should be tailored to the abilities and needs of each learner, in order to keep them stimulated and avoid boredom, confusion and dropout.

Educational research communities have proposed models that predict mistakes and dropout, in order to detect students that need further instruction. There is now a need to design online systems that continuously learn as data flows, and self-assess their strategies when interacting with new learners. These models have been already deployed in online commercial applications (ex. streaming, advertising, social networks) for optimizing interaction, click-through-rate, or profit. Can we use similar methods to enhance the performance of teaching in order to promote lifetime success?

We thank the workshop chairs, Nathalie Guin and Amruth Kumar.

May 24, 2018
Tokyo, Japan

Michal Valko, Fabrice Popineau and
Jill-Jênn Vie

Optimizing Human Language Learning

Masato Hagiwara

Duolingo, Pittsburgh, PA, USA
masato@duolingo.com
http://masatohagiwara.net

Abstract. Learning foreign languages has become an essential skill in the globalized economy - English alone is estimated to have a total of 1.5 billion learners worldwide. As computer-based language learning apps increase in popularity, they generate vast amounts of student learning/behavioral data, opening up entirely new possibilities to optimize human language learning on an unprecedented scale.
In the first part of this talk, we introduce results from our user behavioral analysis and performance prediction projects using the learner data from Duolingo to find out the key traits of successful language learners. Secondly, we review some of the recent development to maximize second language learning through optimizing spaced repetition. Finally, we present the task of second language acquisition modeling (SLAM), which is a task to predict errors made by second language learners based on their past performance, along with some of the key findings from the SLAM shared task we hosted recently.

Keywords: Language learning · Spaced repetition · Second language acquisition modeling.

Where's the Reward? A Review of Reinforcement Learning for Instructional Sequencing

Shayan Doroudi

Computer Science Department
Carnegie Mellon University, Pittsburgh, PA, USA
shayand@cs.cmu.edu
http://www.cs.cmu.edu/~shayand/

Abstract. Since the 1960s, researchers have been trying to optimize the sequencing of instructional activities using the tools of reinforcement learning (RL) and sequential decision making under uncertainty. Many researchers have realized that reinforcement learning provides a natural framework for optimizing and personalizing instruction given a particular model of student learning, and excitement towards this area of research is as alive now as it was over fifty years ago. But does it actually help students learn? If so, when and where might we expect it to be most helpful? To help answer these questions, I will take three approaches. First, I will present a historical narrative of attempts to optimize instructional sequencing using RL. By looking to the past, we hope to better understand why researchers from different communities have worked on this problem and discover some trends that might tell us where the field is going. Second, I will present a case study of two experiments that we ran in a fractions intelligent tutoring system that showed no significant differences between various instructional policies. Finally, I will systematically review the empirical research in this area. We find that in many cases where RL has been applied to rich domains and environments, such as our intelligent tutoring system, it has not been very successful. However, I will show that it has been successful in settings that are constrained in one or more ways. Based on insights we draw from these three approaches, I make suggestions for how the field should proceed if we want to make the most out of reinforcement learning and if we want to quickly identify how rewarding this line of research might be. In particular, I suggest that data-driven RL approaches be informed by and constrained with ideas and theories from the learning sciences and that researchers perform more robust evaluations of instructional policies derived using reinforcement learning before testing them on students. I present on work conducted with Emma Brunskill and Vincent Aleven.

Keywords: Reinforcement learning · Sequential decision making · Student learning.

Knowledge Tracing Machines: towards an unification of DKT, IRT & PFA

Jill-Jênn Vie

RIKEN Center for Advanced Intelligence Project, Tokyo, Japan
vie@jill-jenn.net
https://jilljenn.github.io

Abstract. The goal of this tutorial is to make you compare typical baselines for predicting student performance (item response theory, performance factor analysis) on famous datasets, and replace some blocks of their architectures with deep neural networks (deep knowledge tracing, deep factorization machines). Hopefully we can understand where neural networks improve the predictions substantially, and where they do not. No knowledge of educational models is needed, an experience of Python is preferred. All code can be retrieved at https://github.com/jilljenn/ktm.

Keywords: Item response theory · Deep knowledge tracing · Predicting student performance.

SARLR: Self-adaptive Recommendation of Learning Resources

Liping Liu, Wenjun Wu and Jiankun Huang

State Key Lab of Software Development Environment Department of Computer Science and Engineering, Beihang University, Beijing, China
{liuliping,wwj,hjk}@nlsde.buaa.edu.cn

Abstract. Personalized recommendation is important for online students to select rich learning resources and make their own learning schedules. We propose SARLR, a new self-adaptive recommendation algorithm of online learning resources. The SARLR algorithm integrates an IRT-based learning cognitive model named T-BMIRT into the recommendation framework and is able to adaptively adjust learning path recommendations based on dynamic of individual learning process. The experimental results show that the SARLR algorithm outperforms the existing recommendation algorithms.

Keywords: Online Education, Learning Recommendation, ITS

1 Introduction

With the growing prevalence of online education, students have access to all kinds of electronic learning resources, including electronic books, exercises and learning videos. Given the diversity of students' background, learning styles and knowledge levels, it is essential to have personalized recommendation tools to facilitate students in choosing their own learning paths to satisfy their individual needs [1]. Previous studies have introduced personalized learning recommendation algorithms following the two major approaches including rule-based recommendation and data-driven recommendation.

Most Intelligent Tutor Systems (ITS) such as [2], primarily adopt the rule-based approach to design their recommendation algorithms, which requires domain experts to evaluate learning scenarios for different kinds of students and define extensive recommendation rules accordingly. Apparently, such a labor-intensive approach can only be applied in specific learning domains. For modern online educational systems, designers often take the data-driven approach by utilizing collaborative filtering methods to implement learning recommendation algorithms. These data-driven recommendation algorithms [3] attempt to identify suitable learning resources for students by comparing similarity among students and learning objects.

Although the data-driven recommendation approach is more scalable and general than the rule-based approach, current proposed solutions have common problems in achieving highly adaptive recommendation towards students' latent learning state. They often focus on either searching for similar learning resources based on content or identifying similar student groups based on their learning behaviors. The recommended learning objects or paths fail to consider the impact of difficulty of learning objects and dynamic change in students' learning states.

In this paper, we propose a novel learning recommendation algorithm named SARLR, which attempts to integrate an IRT-based learning cognitive model into the recommendation framework and to adaptively adjust learning path recommendations based on dynamics of individual learning process. Specifically, we introduce a temporal, multidimensional IRT-based model named as T-BMIRT, which can accurately infer student proficiency of multiple latent skills and difficulties of exercise assessments. In addition, the T-BMIRT model incorporates the parameter of video learning, which can describe the improvement in student skills after their interactions with video lectures. Based on the T-BMIRT model, the SARLR algorithm can comprehensively analyze every student's skill progress at each learning step and recommend to them a personalized learning path with the matching online video lectures and homework problems.

The contributions of this paper are the two-fold. First, we introduce the T-BMIRT model, to estimate students' latent skill levels and difficulties of learning resources for recommendation. Second, we propose the SARLR algorithm by integrating the T-BMIRT model in the adaptive recommendation process of learning resources. The experimental results confirm that the SARLR outperforms regular recommendation algorithms. Lastly, we present an evaluation strategy for recommendation algorithms in terms of rationality and effectiveness.

2 Related Work

Data-driven learning recommendation algorithms often utilize common recommendation methods widely adopted in the e-Commence area, including Collaborative Filtering (CF) and Latent Factor Model (LFM). CF can be further divided into UCF (User-based Collaborative Filtering) and ICF (Item-based Collaborative Filtering). The core idea of LFM is to connect users and items through latent features [4].

EduRank [5] is a collaborative filtering based method for personalization in e-learning. It can generate a difficulty ranking of questions for a target student by aggregating the ranking of similar students. Although this method is able to rank the available exercise questions based on their difficulties for similar students, it doesn't integrate cognitive learning models in its framework for estimating the ability of individual students. Thus, it can't generate the matching learning paths for students based on their state of latent skills.

The most related work to our research in previous studies is the Latent Skill Embedding (LSE) model [6], which also presents a probabilistic model of students and lessons. Although the LSE model provides a good foundation for designing a recommendation framework for personalized learnir [illegible] ıper [6] doesn't propose a detailed recommendation algorithm. Our T-BMIR [illegible] is more fine-grained than the LSE model because it defines a video learning parameter to capture student progress through their interaction with video lectures. Moreover, we present the SARLR algorithm that utilizes the T-BMIRT model to identify similar students for a target student and recommend their learning paths according to the dynamic state of the target student's latent skills. We also extend the recommendation evaluation criteria expected gain by incorporating two more metrics including relevance accuracy and difficulty accuracy. These new metrics can support more comprehensive performance evaluation for learning recommendation algorithms.

Recently, reinforcement learning has been explored in personalized study planning in ITS [7-9]. Most of them have not evaluated their approaches in real online learning scenarios and compared their performance to existing problem selection strategies used in current systems. Moreover, calculating an optimal personalized learning path in a POMPD is often time-consuming and even becomes intractable as the dimensions of the knowledge state and strategy spaces increase. Therefore, our SARLR algorithm adopts the collaborative filter based approach and we plan to investigate the possibility of utilizing reinforcement learning in our framework in future work.

3 SELF-ADAPTIVE RECOMMENDATION

Fig.1 illustrates the major components in the SARLR algorithm. First, it uses the T-BMIRT model to estimate every student's skill levels and difficulties of learning resources. Second, it searches for similar students based on their skill vectors from the outputs of the T-BMIRT model. Third, it extracts the learning path of the best student, whose skill level is the highest among the similar students after learning related knowledge. Lastly, it recommends the learning path to the target student and sets up two pre-warning conditions to adaptively adjust his recommended contents. The target student's latest behavior data are collected instantly and used as a feedback to update the T-BMIRT model. Thus, all of the modules form a closed loop, which constantly optimizes our model.

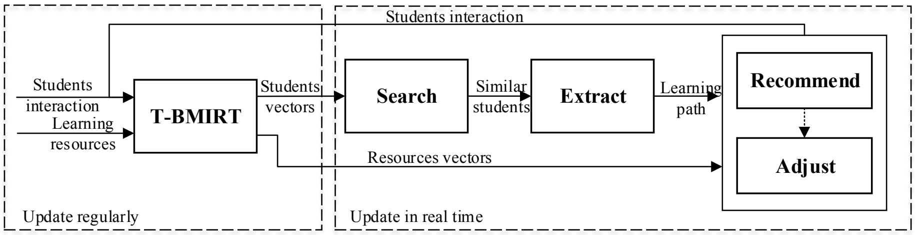


Fig. 1. The Overall architecture of the SARLR algorithm

3.1 The T-BMIRT model

The T-BMIRT model aims to model students and learning resources to infer students' latent skills and learning resources' attributes on multiple knowledge components. We define the model based on IRT, T-IRT and MIRT model [10]. In a two-parameter IRT model, the probability of the student s correctly answering the question q is given by:

$$p_{sq} = \frac{1}{1+exp[-(\alpha_q(\theta_s - \beta_q))]},\ P(\theta_{t+\tau}|\theta_t) = \phi_{\theta_t, \upsilon^2\tau}(\theta_{t+\tau}) \quad (1)$$

Where α_q is the question discrimination, β_q is the question difficulty, θ_s is the student's ability value. The Temporal IRT (T-IRT) model [11] extends the original IRT and MIRT model by modeling a student's latent skills over time as a Wiener process, where $\theta_{t+\tau} - \theta_t \sim N(\theta_t, v_2\tau)$. The model indicates the ability value of the student at the next moment is only relevant to his current ability value.

The T-IRT model only considers interactions between students and assessments, ignoring their interactions with learning videos. However, we believe that the students' ability can be significantly improved after completing a learning video. Therefore, in [12], we introduce a new model T-BMIRT by incorporating learning video parameters to describe the impact of students' interaction with learning videos. The major equations are defined in Eq (2):

$$P(\vec{\theta}_{s,t+\tau}|\vec{\theta}_{s,t}, \vec{l}_{s,t}) = \phi_{\vec{\theta}_{s,t}+\vec{l}_{s,t}, \upsilon^2\tau}(\vec{\theta}_{s,t+\tau}), \vec{l}_{s,t} = \frac{d_{s_t}}{d_t} \cdot \vec{g}_t \cdot \frac{1}{1+exp\left(-\left(\frac{\vec{\theta}_{s,t} \cdot \vec{h}_t}{\|\vec{h}_t\|} - \|\vec{h}_t\|\right)\right)} \quad (2)$$

Where $\vec{l}_{s,t}$ represents knowledge that student s gains from the video t, $\vec{g}_t$ represents knowledge of the video t, $\vec{h}_t$ is the prerequisites of video t, d_{s_t} is the duration in which student s watches video t and d_t is the total length of the video t. In Eq (2), both student ability and learning video requirements have been expanded from one-dimensional to multidimensional. We utilize the vector projection method to determine whether the relevant abilities of the student exceed the relevant skill requirements of the video lectures.

The T-BMIRT model enables us to infer every student's current ability θ, video knowledge g and video skill requirements h through the student's responses of assessment questions. The detailed model fitting process of the T-BMIRT can be found in [12]. An approximation technique makes it possible to train the T-BMIRT in an online way. As a result, the T-BMIRT can be effectively used in the framework of the SARLR algorithm to estimate the parameters of learning resources and students' ability levels.

3.2 Similar Students Search and Learning Path Extraction

SARLR Phase 1 describes the process of searching similar students and extracting a suitable learning path for a target student. At Step 1, the algorithm identifies the students *MS* with the similar skill levels to the target student s_X through k-nearest neighbor search method over the k-dimension tree (kd-tree) structure and k-nearest neighbor search method. At Step 2-4, the algorithm selects the best student $s_b \in MS$ with the highest ability level at the moment when they complete learning specific knowledge units. At Step 5, the algorithm extracts the learning path p of s_b to the target student s_X.

SARLR Phase 1: Search and Extraction

```
INPUT:
    Set of students $S = \{s_1, s_2, \ldots, s_n\}$, target student $s_X \in S$
    Matrix of abilities $A = [\theta_{s,t}]$, where $\theta_{s,t}$ is the ability value of student s at time t
    Set of learning resources $E = \{e_1, e_2, \ldots, e_m\}$
    The time in this paper is the index of learning resources with the student just completed learning.
OUTPUT: learning path $p$
    search for similar students $MS$, where $s_k \in MS$ and $\theta_{s_k,t_0}$ is similar to $\theta_{s_X,t_0}$
    for each $s_i \in MS$ do
        find $s_b = argmax(distance(\theta_{s_i,T_{s_i}} - \theta_{s_i,t_0}))$, where $T_{s_i}$ is the time of $s_i$ completing learning
    end for
    extract the learning path $p = (e_{i_1}, e_{i_2}, \ldots e_{i_T})$ of $s_b$
    return $p$
```

3.3 Adaptive Adjustment

Because each individual student has his/her inherent learning style, even when he follows the recommended learning path generated in SARLR phase 1, the learning outcome may not be as good as expected by the recommendation algorithm. In order to deal with this problem, we set up the two conditions in Eq (3) to initiate the Adaptive Re-planning phase, which is defined in *SARLR Phase 2*.

$$p_{sq} = \frac{1}{1+exp\left(-(\vec{\theta}_{s,i}\cdot\vec{\alpha}_q-b_q)\right)},\ p_{se} = \frac{1}{1+exp\left(-\left(\frac{\vec{\theta}_{s,i}\cdot\vec{h}_e}{\|\vec{h}_e\|}-\|\vec{h}_e\|\right)\right)} \tag{3}$$

Eq (3) specifies p_{sq} and p_{sl} to evaluate the progress of the target student in the learning path. p_{sq} indicates the probability of student s correctly answering exercise q, where $\vec{\theta}_{s,i}$, $\vec{\alpha}_q$ and b_q represent the same symbols as the T-BMIRT model in Eq (1-2). p_{se} indicates the degree of knowledge that student s can acquire from the video e, where $\vec{q}_e$ represents the level of knowledge required for the learning video.

When p_{sq} becomes less than the threshold C_{sq}, it means that the difficulty of the exercise q in the recommended learning path has significantly exceeded the student's ability. When p_{se} becomes less than the threshold C_{se}, it means that the skill level of the target student is lower than the requirement of the recommend video e , thus he can only acquire little knowledge from the video. When either condition is met, the SARLR determines that the original recommended path has to be re-planned to match the student's knowledge state.

SARLR Phase 2: Adaptive Re-planning

```
INPUT:
    Target student $s_X$, recommended learning path $p = (e_{i_1}, e_{i_2}, \ldots e_{i_T})$
    Result of $s_X$ interacted with learning resources in $p$
OUTPUT: new learning path
    for each $e \in p$ do
        if $e$ is a video and $p_{se} < C_{se}$ do
            return SARLR Phase 1 to re-plan path $p$
        else if $e$ is an exercise and $s_X$ failed it and $p_{sq} < C_{sq}$ do
            return SARLR Phase 1 to re-plan path p
        end if
    end for
```

4 EXPERIMENTS

We selected two datasets to perform our experiments, the public “Assistments”, including 224,076 interactions, 860 students, 1,427 assessments and 106 skills, and a blended learning data from our learning analysis platform including 14,037,146 learning behavior data from 140 schools and 9 online educational companies.

4.1 Experiments for T-BMIRT

We divided each data set into two parts, one part only contains single skill assessments, and the other part contains multiple skills assessments. The IRT, T-IRT are single skill models, and the MIRT and T-BMIRT are multiple skills models. The dimensions for models are related to the numbers of knowledge components. The values in Table 1 are average results of the cross-validation. It shows that T-BMIRT outperforms the other models on each dataset, especially on the multidimensional dataset.

Table 1. Prediction Results of each model

Models	**Assistments**				**Blended learning data**			
	One-dimensional		Multidimensional		One-dimensional		Multidimensional	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Frequency method	0.694	N/A	0.683	N/A	0.702	N/A	0.688	N/A
IRT	0.716	0.779	0.701	0.758	0.721	0.784	0.706	0.752
MIRT	0.714	0.771	0.721	0.786	0.718	0.775	0.722	0.783
T-IRT	0.738	0.805	0.712	0.769	0.744	0.801	0.717	0.764
T-BMIRT	0.743	0.815	0.738	0.803	0.757	0.820	0.748	0.816

4.2 Rationality Evaluation

The rationality evaluation verifies whether the algorithm can recommend the suitable learning resources that meet the student’s needs and ability levels. We set the following two indicators for it.

$$\mathrm{RC}_{\mathrm{s_x}} = \frac{\sum_{e_i}^{p} similarity(h_{e_i}, KC_{s_x})}{m}, \ \mathrm{DC}_{\mathrm{s_x}} = \frac{\sum_{e_i}^{p} similarity(h_{e_i}, \theta_{s_{x,i}})}{m} \tag{4}$$

Where $e_i \in p$ is the learning resources in a recommended path, m is the length of the path, KC_{s_x} is the knowledge components which s_x is learning in the current chapter, function *similarity()* calculates the adjusted cosine similarity of the two vectors in the parentheses. The relevance accuracy RC_{s_x} is used to evaluate whether the difficulties of the recommended learning resources for the target student s_x are matched with his ability. The difficulty accuracy DC_{s_x} is set to evaluate whether the difficulties of the recommended learning resources for the target student can match his current ability levels.

We selected the blending data to do this experiments. Table 2 shows the average of the 10-fold cross-validation results. It can be seen that the UCF and ICF have a similar effect, but the UCF works better on the relevance accuracy, while the ICF is better at the difficulty accuracy. The LFM performs better than the first two algorithms in terms of both indicators. The SARLR algorithm performs best among all these algorithms.

Table 2. Results of Rationality Experiment.

Model	Relevance accuracy	Difficulty accuracy
UCF	0.86	0.77
ICF	0.71	0.83
LFM	0.87	0.84
SARLR	0.97	0.92

4.3 Effectiveness Evaluation

The effectiveness evaluation verifies whether the students' abilities can be improved by the recommendation algorithm. We clustered the students into six groups according their ability levels. We calculated "expected gain" $G = \frac{E\left(R_{S'}\right) - E(R_S)}{E(R_S)}$ by using PCA and K-means method to further split the students of the same group into two parts based on their learning paths [6]. One part is the students whose learning paths are strictly recommended, denoted as S' , and the other part is the students whose learning path are randomly selected, denoted as S. $E(R_{S'})$ and $E(R_S)$ and indicate that the students' average score in the last online assessment. We sorted the six groups of the students ascendingly based on their ability levels: group 1 has the lowest skill level, group 2 has a higher skill level than group 1, and group 6 has the highest.

Table 3. Results of Effectiveness Experiment

Model	Expected gain					
	1	2	3	4	5	6
UCF	-0.04	-0.06	0.07	-0.03	0.08	0.01
ICF	0.05	0.04	-0.03	0.07	-0.02	0.05
LFM	0.04	0.12	0.09	0.10	0.03	-0.05
SARLR	0.11	0.27	0.24	0.23	0.17	0.06

We selected the public data "Assistments" to do this experiments. Table 3 shows that the SARLR algorithm performs much better than the other three algorithms. Especially for the students in group 2 to group 5, the SARLR algorithm helps them to achieve noticeable progress from the recommendation learning paths. It indicates that SARLR is more effective on improving learning gain of students with average ability levels.

5 CONCLUSIONS

We developed a self-adaptive recommendation algorithm of learning resources (SARLR) to personalize students' learning path. It contains the T-BMIRT, a temporal blended multidimensional IRT model, which performs well on the prediction task of multi-dimensional skills assessments, especially when the study process contains learning video interactions. Based on the T-BMIRT model, the SARLR algorithm adopts a reasonable recommendation strategy and establishes conditions to adaptively adjust recommendations towards the dynamic needs of the students. In addition, we extend the evaluation criteria for personalized learning recommendation in term of rationality and effectiveness. Experimental results prove that the SARLR algorithm outperforms the other recommendation algorithms based on CF and LFM.

References

1. Akbulut, Y., Cardak, C. S.: Adaptive educational hypermedia accommodating learning styles: A content analysis of publications from 2000 to 2011. Computers & Education. 58(2), 835-842 (2012).
2. Vesin, B., Ivanović, M., KlašNja-MilićEvić, A., Budimac, Z.: Protus 2.0: Ontology-based semantic recommendation in programming tutoring system. Expert Systems with Applications. 39(15), 12229-12246 (2012).
3. Wu, D., Lu, J., Zhang, G.: A fuzzy tree matching-based personalized e-learning recommender system. IEEE Transactions on Fuzzy Systems. 23(6), 2412-2426 (2015).
4. Jenatton, R., Roux, N. L., Bordes, A., Obozinski, G. R.: A latent factor model for highly multi-relational data. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, pp. 3167-3175. ACM, California (2012).
5. Segal, A., Katzir, Z., Gal, K., Shani, G., Shapira, B.: Edurank: A collaborative filtering approach to personalization in e-learning. In: Proceedings of the 7th International Conference on Educational Data Mining, pp. 68–75. EDM, London (2014).
6. Reddy, S., Labutov, I., Joachims, T.: Latent skill embedding for personalized lesson sequence recommendation. *arXiv preprint arXiv:1602.07029* (2016).
7. Theocharous, G., Beckwith, R., Butko, N., Philipose, M.: Tractable POMDP Planning Algorithms for Optimal Teaching in SPAIS. In: Workshop on Plan Activity, and Intent Recognition (PAIR), IJCAI (2009).
8. Folsom-Kovarik, J. T., Sukthankar, G., Schatz, S. L., Nicholson, D. M.: Scalable POMDPs for Diagnosis and Planning in Intelligent Tutoring Systems. In: AAAI Fall Symposium: Proactive Assistant Agents. AAAI Press, Virginia (2010).
9. Brunskill, E., Russell, S.: Partially observable sequential decision making for problem selection in an intelligent tutoring system. In Educational Data Mining (EDM), pp. 327–328 (2011).
10. Reckase, M.: Multidimensional item response theory. Springer, New York (2009).
11. Ekanadham, C., Karklin, Y.: T-skirt: Online estimation of student proficiency in an adaptive learning system. arXiv preprint arXiv:1702.04282 (2017).
12. Huang, J.,Wu, W.: T-BMIRT: Estimating representations of student knowledge and educational components in online education. In: 2017 IEEE International Conference on Big Data, pp. 1301-1306. IEEE Press, Massachusetts (2017).

An Adaptive Tutor to Promote Learners' Skills Acquisition during Procedural Learning

Joanna Taoum, Anaïs Raison, Elisabetta Bevacqua and Ronan Querrec

Lab-STICC, UMR 6285, CNRS, ENIB, France
{taoum,raison,bevacqua,querrec}@enib.fr

Abstract. Our research work proposes an adaptive and embodied virtual tutor based on intelligent tutoring systems. The domain model is represented in our work by a virtual environment meta-model and the interface by an embodied conversational agent. Our main contribution concerns the tutor model, that is able to adapt the execution of a pedagogical scenario according to the learner's level of knowledge. To achieve such a goal, we rely on the inference of the learner's memory content.

Keywords: Adaptive Pedagogical Behavior · Virtual Environment · Learner's Memory · Pedagogical Scenario · Embodied Conversational Agent

1 Introduction

The work presented in this paper is applied to the domain of procedural learning in a virtual environment for industrial systems. According to Anderson [1], procedural learning is considered to be complex and this complexity requires the use of practice (repetition). In order to be able to manage the interaction between a tutor and a learner during these repetitions, we choose to describe this information using pedagogical scenarios. These scenarios define the activities that should be carried out by the tutor and the learner, their sequencing, as well as the pedagogical objectives that should be achieved.

However, these scenarios remain general. They can be effective at the beginning of learning (during the first repetitions), but not in the following repetitions. Considering that each learner evolves differently, during repetitions, it is important to adapt the execution of these pedagogical scenarios according to the learner's evolution.

The real-time adaptation of the pedagogical situation to a learner is one of the major objectives of Intelligent Tutoring Systems (ITSs). In order to adapt the situation to the learner, a fundamental goal of an ITS is to model the learner. In procedural learning domain, Corbett and Anderson [2] propose some general concepts to model the learner during the acquisition of procedural skills. These concepts are too theoretical to be applied to teaching procedures in industrial systems. As we are dealing with teaching human activities in industrial systems, the cognitive knowledge that our student model infers is related to memorization. Atkinson and Shiffrin [3] proposed a general theoretical framework which divides human memory into three structural components: sensory memory, working memory and long-term memory. To implement this general framework of memory, several ITSs have been built using the cognitive architecture Act-r [4]. The goal of Act-r is to simulate the realization of complex tasks by human beings. It is mainly designed around two concepts: declarative and procedural knowledge. Declarative knowledge is represented by a set of chunks and procedural knowledge by a set of production rules (*if-then* statements). In Act-r, information processing of memory is a *Black Box*. It can be used to generate the tutor behavior but not to represent the knowledge flow in the learner model.

In this work, we propose a tutor behavior that adapts the execution of the pedagogical scenario according to the learner's inferred knowledge (see section 3.1). To represent such a knowledge, we propose a cognitive architecture based on Act-r [4]. In section 2, we introduce Mascaret [5] that we use to represent the domain model and the pedagogical scenario. To realize pedagogical assistances in a human-like way, we propose an interface model based on a virtual environment and an Embodied Conversational Agent (ECA).

2 Domain and Interface Model

The domain model is formalized in our work by Mascaret, a virtual reality meta-model based on UML. It allows to describe and simulate technical systems and human activities in a virtual environment. The domain expert uses class diagrams to describe the different types of entities, their properties and the structure of the environment. Procedures are designed as predefined collaborative scenarios through UML activity diagrams, which represent plans of actions. It is the role of the interface model to recognize when the student executes these actions. Using a meta-model to formalize the domain model 1) allows domain experts to provide the knowledge themselves in the ITS, and 2) keeps domain data explicit during the simulation, thus they can serve agents as the knowledge base.

In Mascaret, pedagogy is considered as a specific domain model. Pedagogical scenarios are implemented through UML activity diagrams containing a sequence of actions. These actions can be either pedagogical actions, like explaining a resource, or domain actions, like manipulating an object. For the definition of pedagogical scenarios and actions, we rely [6]. In Mascaret five types of pedagogical actions are implemented:

1. Pedagogical actions on the virtual environment: highlighting an object, playing an animation.
2. Pedagogical actions on user's interactions: changing the viewpoint, locking the position, letting the student navigate.
3. Pedagogical actions on the structure of the system: describing the structure, displaying a documentation about an entity.
4. Pedagogical actions on the system dynamics: explaining the procedure's objectives, explaining an action.

5. Pedagogical actions on the pedagogical scenario: displaying a pedagogical resource, making an evaluation (e.g. a quiz).

These pedagogical actions are realized through the interface model, that is represented in our work by an ECA, using GRETA platform [7]. This ECA is able to select and perform multi-modal communicative and expressive behaviors in order to interact naturally with the user. In MASCARET, any entity which acts on the environment is considered as an *agent*. Particularly, the ECA and the human user are *embodied* agents. An embodied agent is able to recognize as well as perform basic actions, like:

1. Verbal communication (e.g. giving an information)
2. Non-verbal actions (e.g. facial expression) and actions on the environment (e.g. manipulating an object)
3. Navigation (e.g. observing)

These basic actions are used to implement the domain and pedagogical actions involved the pedagogical scenario. Through the interface model, the tutor is able to recognize the realization of each of these actions performed by the user to evaluate the evolution of the pedagogical scenario and to adapt it if necessary.

3 Adaptive Tutor Model

The tutor model uses the knowledge of the domain model and the actions done by the learner in order to choose pedagogical actions that will be realized through the interface model. More precisely, the tutor behavior takes into account the actions done (or inaction) by the student by recognizing them through the interface. The goal of our proposed tutor model is to adapt the execution of the pedagogical scenario according to the student model represented in our work by the student's memory.

In what follows, we first describe the student model that is used to decide which adaptation to perform and then how the tutor behavior detects the need for adaptation.

3.1 Student Model

We propose a reimplementation of the generic framework of memory proposed by Atkinson and Shiffrin [3] in the context of learning procedures. Our contributions to this framework consist in making explicit the *Black Box* by 1) formalizing the user's memory information, and 2) implementing the transformation of the stimuli into knowledge and the knowledge flow between the three components of the human memory. In our work, incoming stimuli from the virtual environment and the virtual tutor are restricted to those related to vision and hearing. Thus, the student can see 3D objects and hear instructions uttered by the tutor about activities to realize. Therefore, we encode data about objects and activities. To formalize the encoding of information, we rely on MASCARET.

Objects are considered in MASCARET as `Entity`. An `Entity` can be hierarchical, thus it can be composed of `Entity` and represented by a name, geometric properties (position, orientation and shape) and domain model properties (as a meta class `Class` attribute). As for activities, they are represented by the meta class `Activity`, they can also be hierarchical and composed of several `Activity`, `Role`, `Action` and `Flow` between actions and objects. MASCARET data formalism is hierarchical, which allows to instantiate the content of the memories according to the knowledge level of the learner.

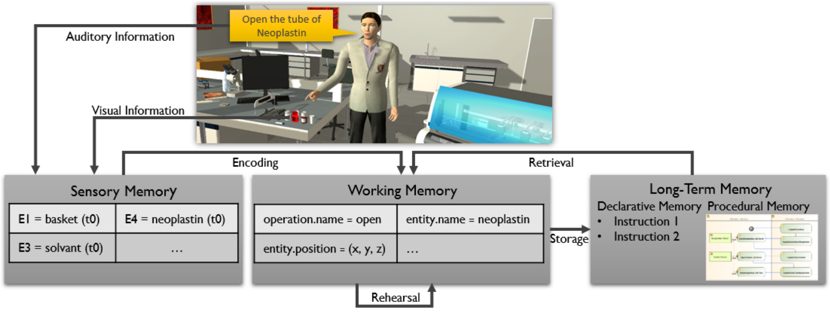


Fig. 1. Formalization of the encoding and structuring of instructions in the memory.

In this work, we therefore distinguish three structural components in human memory in which a sequence of cognitive processes is implemented to process information (encoding, storage, retrieval). The first operation involved in the information processing is the encoding of information. It is the transformation of incoming stimuli from the virtual environment and the virtual tutor to a formal representation that can be stored in the working memory. As mentioned previously, incoming stimuli are visual (set of objects in the student's field of view) and auditory (uttered by the tutor). Only prominent information (e.g. objects that have been highlighted by the tutor) is transferred from the sensory memory to the working memory. The working memory stores and manipulates information based on the content of the sensory memory and the long-term memory (prior knowledge). The level of complexity of stored information in the working memory depends on the student's prior knowledge (by complexity of information we mean the level of the formal representation in MASCARET hierarchical formalism). This prior knowledge is retrieved from the long-term memory. The transfer of some knowledge from the working memory to the long-term memory, takes place when the student completes an action [8].

This student model is used as an input in the tutor behavior.

3.2 Tutor Behavior

The tutor behavior takes into account the actions done by the learner and the inferred student model to adapt the execution of the pedagogical scenario. This adaptation can be a modification of the student model (modification of the memory content) and/or the execution of a pedagogical action. The decision making of the tutor behavior is represented in Figure 2.

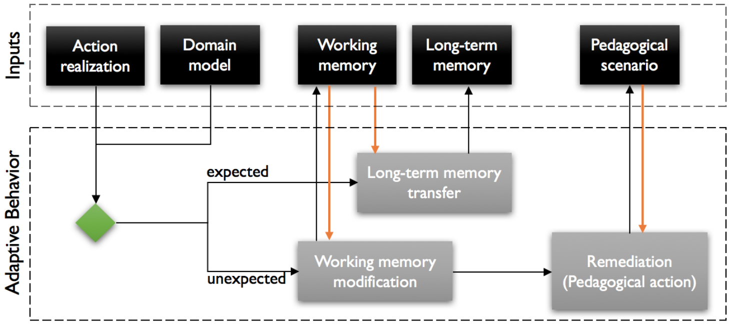


Fig. 2. Tutor behavior decision making.

The execution of a pedagogical scenario is a set of interaction between the tutor and the learner. As explained in section 2, the tutor actions (pedagogical actions) are realized through the interface, and this latter is also able to recognize the actions realized by the learner in the context of this interaction.

Our tutor behavior categorizes the actions done by the learner, based on two types of actions:

1. related to the domain model: an action can be either a domain action on a specific object or an answer to the tutor's questions. The tutor relies on the domain model to check if these actions are considered as errors or not.
2. related to the interaction: actions done by the learner can also be a feedback to the tutor's action (e.g. a facial expression, a question, observing the environment or an inaction). In this case, instead of using the domain knowledge, the tutor evaluates whether this feedback is negative or not.

If the learner's action is considered as an error or as a negative feedback, this means that this action is unexpected in the context of the executed scenario. In this case a new pedagogical action is needed and the content of the learner's memory must be reevaluated.

For example, if according to the pedagogical scenario the tutor explains the next action that the student has to do, we instantiate two chunks in the working memory, one for the `Action` and the other one for the `Entity`. If the student realizes an unexpected action (for example he/she shows a negative facial expression), then the tutor behavior considers that the student does not know the object position, contrary to what the tutor inferred. In this case the tutor remedies to this situation by re-evaluating the content of the student's working memory and then realizes a new pedagogical action to highlight the object.

4 Conclusion and Future Work

The model that we propose here allows an embodied conversational agent, playing the role of tutor, to execute a predefined pedagogical scenario written by a trainer in a virtual environment and especially to adapt its execution according to the individual evolution of students. To do this, the ECA infers the student's knowledge by estimating the content of his/her memories involved in procedural learning. The tutor behavior that we propose is a simple behavior that allows us to show the usability of the memory model that we have implemented to define a pedagogical behavior. In the same way that in Mascaret it is the trainer who describes the pedagogical scenario using a dedicated language (based on UML activities), we consider that it would be more interesting if it is the pedagogue which describes the tutor's behavior using the same language. We aim to make the concepts defined in our model accessible and formalized in this language.

In order to evaluate the impact of our model on the student's performance, we plan to carry an experiment that will involve two groups of participants. In the first group, a non-adaptive virtual tutor will be present in the virtual environment. The non-adaptive tutor will apply a single pedagogical scenario during repetitions. If the student asks for help, the tutor announces the action to be performed, its goal and highlights the object to manipulate. In the second group, an adaptive tutor will guide the learner. Based on our model, the tutor will be able to adapt the execution of the pedagogical scenario according the evolution of the learner's level of expertise. In this experiment, we expect that learners interacting with an adaptive tutor perform the procedure without errors and without the need for help, earlier than those who are interacting with a non-adaptive tutor.

References

1. Anderson, J.: A spreading activation theory of memory. Journal of verbal learning and verbal behavior **22** (1983) 261–295
2. Corbett, A., Anderson, J.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction **4** (1995) 253–278
3. Atkinson, R., Shiffrin, R.M.: Human memory: A proposed system and its control processes. In K. W. Spence and J. T. Spence (Eds.), The Psychology of learning and motivation: Advances in research and theory (vol. 2). (1968) 89–105
4. Anderson, J.: The architecture of cognition. 2nd edn. Psychology Press (2013)
5. Chevaillier, P., Trinh, T., Barange, M., Devillers, F., Soler, J., De Loor, P., Querrec, R.: Semantic modelling of virtual environments using Mascaret. In: Proceedings of the 4th Workshop SEARIS, IEEE VR, Singapore. (2001)
6. Le Corre, F., Hoareau, C., Ganier, F., Buche, C., Querrec, R.: A pedagogical scenario language for virtual environment for learning based on uml meta-model. application to blood analysis instrument. In: International Conference CSEDU. (2014) 301–308
7. Niewiadomski, R., Bevacqua, E., Mancini, M., Pelachaud, C.: Greta: an interactive expressive eca system. In: 8th International Conference AAMAS. (2009) 1399–1400
8. Ganier, F.: Factors Affecting the Processing of Procedural Instructions: Implications for Document Design. IEEE Transactions on Professional Communication **47**(1) (2004) 15–26

Optimizing Recommendation in Collaborative E-Learning by Exploring DBpedia and Association Rules

Samia Beldjoudi[1], Hassina Seridi[2]

[1]Superior School of Industrial Technologies, Annaba, Algeria
[1, 2]Laboratory of Electronic Document Management LabGED Badji Mokhtar University, Annaba, Algeria
[1]s.beldjoudi@epst-annaba.dz [2]seridi@labged.net

Abstract. Social tagging activities allow users to add free annotations on resources to express user interests, preferences and automatically generate folksonomies. This paper demonstrates how structured content available through DBpedia can be leveraged to support recommendation of resources in folksonomies. A limitation of resources' recommendation is the content overspecialization conducting in the incapability to recommend relevant resources different from the ones that the learner already knows. To address this issue, we proposed to take advantage of the richness of the open and linked data graph of DBpedia and association rules to learn learners' behavior. The proposed approach demonstrates the efficiency of using DBpedia to enhance diversity and novelty when recommending resources to users in folksonomies. The basic idea is to iteratively explore the RDF data graph to produce novel and diverse relevant recommendations.

Keywords: Collaborative E-learning, Recommendation, DBpedia, Diversity, Novelty.

1 Introduction

Social tagging systems have achieved a great success over the web in the last years, especially in recommendations approaches. The problem of a precise recommender system is that the entire set of recommended resources may be obvious as one considers the case of a film recommendation algorithm that only returns films of the same actor. To overcome this problem, novelty and diversity should be also considered in the evaluation of a recommender system, as precision only offers an incomplete description of the system's effectiveness.

The main focus of our study is how to exploit the semantic aspect of DBpedia to enhance resource recommendation within social tagging systems. We propose a new method for analyzing learner profiles according to their tagging activities in order to improve the recommendation of resources. The effectiveness of results depends on the resolution of social tagging drawbacks. In our process, we demonstrate how we can reduce the tags ambiguity problem by taking into account social similarities calculated on folksonomies combined with similarities between resources in DBpedia. We used also the force of Linked Open Data (LOD) to enhance resource recommendation by exploring the interlinked entities in LOD cloud. We base up on the iterative exploration of the DBpedia graph to obtain novel and diverse recommendations that should satisfy the learner and create the effect of surprise by recommending resources that the user did not expect at the beginning.

This paper is organized as follows: Section 2 is an overview of the main contributions related to our work. Section 3 is dedicated to the presentation of our approach. In section 4 we present and discuss the results of some experiments we conducted to measure the performance of our approach. Conclusion and future works are described in Section 5.

2 Related works

Social web based approaches, like folksonomies, have achieved a high level of improvement even in E-learning practice. In this section, an overview about some contributions attached to this field is proposed. [Kopeinik et al., 2017] investigated the application of two tag recommenders that are inspired by models of human memory. The authors find that displaying tags from other group members helps significantly in semantic stabilization in the group, as compared to a strategy where tags from the students' individual vocabularies are used. In [Beldjoudi et al., 2016], the authors proposed a new approach for personalizing and improving resources retrieval in collaborative learning with tackling tags ambiguity and event detection impact on resourced retrieved by ranking. In another contribution [Beldjoudi et al., 2017] proposed a method to analyze user profiles according to their tags in order to predict interesting personalized resources and recommend them. The authors proposed a new approach to reduce tag ambiguity and spelling variations in the recommendation process by increasing the weights associated to web resources according to social similarities. They base upon association rules for discovering interesting relationships among a large dataset on the web. [Karabadji et al., 2018] proposed to focus mainly on the growing of the large search space of users' profiles and to use an evolutionary multi-objective optimization-based recommendation system to pull up a group of profiles that maximizes both similarity with the active user and diversity between its members. In such manner, the recommendation system will provide high performances in terms of both accuracy and diversity. In our work we want to leverage the social and semantic web in order to enhance educational resources recommendation in collaborative e-learning.

3 Approach description

In this paper, we propose a method to analyze learner profiles according to their tags in order to predict interesting personalized resources and recommend them. We argue that the automatic sharing of resources strengthens social links among learners and we exploited this idea to reduce tag ambiguity in the recommendation process by increasing the weights associated to web resources according to social similarities. We based upon association rules that are a powerful method for discovering interesting relationships among a large dataset on the web. Our goal was to find correlations between tags, i.e. to find tags frequently appearing together, in order to extract those which are not used by one particular learner but which are often used by other users close to him in the social network.

The effectiveness of the recommendation depends on the resolution of the problems of folksonomies. In our approach we tackle the problems of tag ambiguity, diversity and novelty. To resolve the problem of tag ambiguity in recommendation, we propose to measure the similarity between learners to identify those who have similar preferences and therefore adapt the recommendation to learner profiles.

- First step: For each extracted association rule (Tags *A* → *Tags B*) whose antecedent applies to an active learner *lx*, we measure the similarities between this learner and the learners of his social network who use the tags occurring in the consequent of the rule. The resources associated to these tags are recommended to the learner depending on these similarities. To measure similarity between two learners (*l1* and *l2*), both are represented by a binary vector representing all their tags and we compute the cosines similarity between the two vectors.

- Second step: To avoid the cold-start problem which generally results from a lack of data required by the system in order to make a good recommendation, when the learner of the recommender system is not yet similar to other users, we propose to exploit semantic links between resources in DBpedia. DBpedia can be a reliable and rich source of content information that supports recommender systems to overcome problems, such as the cold-start problem and limited content analysis that restrict many of the existing systems, by building on a robust measurement of the similarities between resources using DBpedia. In this approach, we use the Linked Open Data to assess the similarity between folksonomies resources using their corresponding resources on DBpedia (i.e. we measure the similarity between the resources that would be recommended by the system, as related to a tag occurring in the consequent of an association rule, and those that are already recommended to the learner). The similarity between two resources is calculated using Jaccard index.

In another hand, when using a recommender system such as those of online stores, the results are mainly expected by the users. In this case, it is clear that the recommendation is not very helpful in the sense of the lack of diversity and novelty. To solve this dilemma in folksonomies-based collaborative learning, we propose extracting the most popular features found in the resources-based learner profile (i.e. the characteristics that interest the learner when they tag their resources) and then explore the LOD to extract resource linked with these features.

Let us consider a learner profile composed from the resources (*R1, R2, R3 and R4*), Thus the intersection between the resources' features must be calculated (*R1*∩ *R2* ∩ *R3* ∩ *R4*), this is done because we want to extract the most popular characteristics that interest the learner when they choose tagging their resources. Then for each feature (*Pi*) in the result of intersection we will explore the LOD graph in the first level to extract other resources (*R5*) having these features or having a direct/ indirect link with these later (*R6, R7* resp).

Supposing that (*R1*∩ *R2* ∩ *R3* ∩ *R4*) = {[domain: informatics]; [author: …]; [year: …]; [edition: …]…}. By exploring the LOD graph we find that the resource "informatics" is linked with other resources (for example: "University, Formation, Bio-Informatics…") via the predicates (*P1, P2, P3…*). In its turn the resources "University, Formation, Bio-Informatics…" are linked via other predicates (*Pj*) with other resources (for example: "Boston University…"). Therefore, it appears relevant to recommend some courses of the Boston University to the current user.

Our approach is based on the iterative exploration of the DBpedia graph, where each step depends on the result of the previous steps. In order to obtain relevant and personalized recommendations for each learner, we calculate the occurrence number of the {domain, author, year, edition…} characteristics and then we choose the ones that best reflect the learner interest to exploit them later in the exploration of the RDF graph of DBpedia.

The purpose of the graph exploration is to obtain recommendations that should not only satisfy the learner but also to have diversity and a novelty in the recommendation, to create the effect of surprise by recommending resources that the learner did not expect at the beginning. The learner evaluates the recommended resources in real time in each iteration. The process stops when none of the recommended resources has satisfied the user.

If the learner liked at least one resource among those in the proposed list, in the second iteration, we focus on these ones. Thus, we re-explore the LOD graph again starting from these items by using the query language SPARQL to return more educational resources connected with them; this technique allows us to propose a list of diverse and novel resources to ensure the surprise effect.

The real-time evaluation process as well as the exploration of the graph is iterative. At each iteration, we explore the graph based on the positive ratings assigned to the resources previously recommended. Indeed, the evaluation is an essential step to determine the new pattern of requests for the re-exploration of the graph to generate another list of recommendations. At each step, we propose to the user 10 resources, if he assigns a rating more or equal to three, we consider that he liked the recommended resource, and so we record it in his profile, otherwise we move to another resource.

After evaluating the 10 resources, the program suggests to the user to recommend after the 10 resources have been evaluated, the program suggests to recommend some more to the user. If he accepts then another list of resources is generated from his profile, otherwise, we stop and return the list of resources liked. With this method, we ensure that the recommended list of resources is diverse, where every user can obtain diverse resources even if they do not appear in the profile of his neighbors in the social network.

4 Experimental Results

In this section, experiment over a popular dataset is described and results are analyzed and discussed. The dataset exploited in our test is del.icio.us. In this experiment, we were interested in data generated from users who tagged resources about education. Thus, our database comprises 1128 tag assignments involving 95 users, 432 tags containing ambiguous tags and 314 resources.

4.1 Experimental Methodology

To evaluate the quality of a recommender system, we must demonstrate that the recommended resources are really being accepted and added by the users. Because the knowledge of this information requires asking the users of the selected databases if they appreciated the proposed set of resources, which is impossible in our case because we do not have access to this community, we have used a cross-validation where we have randomly removed some resources from the profile of each user, and we applied our approach on the remainder dataset in order to show if it can recommend the removed resources to their corresponding users or not. If it is the case, so we can conclude that our approach enables to extract the user preferences.

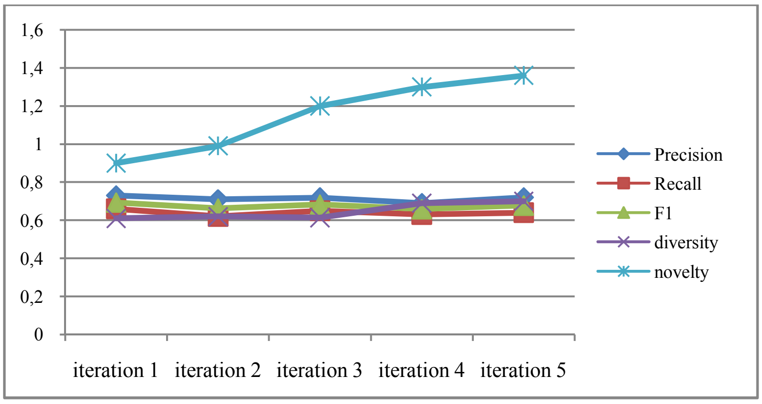


Fig. 1. Average precision, recall, F1, diversity and novelty of the recommendations

The curve presented in figure 1 show average values of precision, recall, F1, diversity and novelty measures in the five iterations. We notice that the precision achieved a good value in all iterations, this is due to the fact that the system recommends exactly the items wanted by the user i.e. those that match his profile. Sometimes the system begins to deteriorate in terms of precision but always with a value that exceeds 0.6. This decrease is quite normal since the system begins to recommend items according to different attributes (domain, year ...) which is known as diversity of recommendation. Learners sometimes accept the recommended resources and other times it was not the case. Recall and F1 measure achieved all both good values in the all iterations.

To calculate individual diversity and novelty, we used the metrics proposed in [Zhang and Hurley, 2009] and [Vargas, 2014] respectively. Figure 1 showed promising values of both diversity and novelty in the five iterations. This demonstrates the importance of DBpedia to extract more diversified and novel resources in the recommendation. It is clear that the effectiveness of recommendation depends of preserving both precision and diversity. Results demonstrate that our approach preserving both them in all iterations.

5 Conclusion

In this contribution we have exploited the strength of social aspect in folksonomies to let members in the community benefit from the educational resources tagged by other users, based on the recommendation of resources. The proposed approach is based on DBpedia, the objective was to overcome the problem of diversity and novelty in recommendation. Primary results show also the utility of exploring LOD graph in ensuring diversity when recommending personalized educational resources in social tagging systems. In order to continue and improve our work, we aim at using others principles like event detection, for example, to help capturing and analyzing the behavior of learners when new events come, this can improve recommendation and even resources ranking.

References

1. Beldjoudi, S., Seridi, H. & Benzine, A.: The Impact of Social Similarities and Event Detection on Ranking Retrieved Resources in Collaborative E-Learning Systems. In: Koch F., Koster, A., Primo, T., Guttmann, C. (eds) Advances in Social Computing and Digital Education. CARE 2016, SOCIALEDU 2016. Communications in Computer and Information Science, vol 677. Springer, Cham (2016)
2. Beldjoudi, S., Seridi, H. & Faron-Zucker, C.: Personalizing and Improving Resource Recommendation by Analyzing Users Preferences in Social Tagging Activities. Computing and Informatics 36(1): 223-256 (2017)
3. Karabadji, NEI., Beldjoudi, S., Seridi, H., Aridhi, S. & Dhifli, W.: Improving memory-based user collaborative filtering with evolutionary multi-objective optimization. Expert Syst. Appl. 98: 153-165 (2018)
4. Kopeinik, S., Lex, E., Seitlinger, P., Albert, D. & Ley, T.: Supporting collaborative learning with tag recommendations: a real-world study in an inquiry-based classroom project. In: LAK '17 Proceedings of the Seventh International Learning Analytics & Knowledge Conference Pages 409-418 (2017).
5. Vargas, S.: Novelty and diversity enhancement and evaluation in recommender systems and information retrieval. In Proceedings of the 37th International ACM SIGIR Conference on Research &#38; Development in Information Retrieval, SIGIR '14, pages 1281–1281, New York, NY, USA, (2014).
6. Zhang, M. and Hurley, N.: Novel item recommendation by user profile partitioning. In 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, volume 1, pages 508–515, Sept (2009).