AIED 2013 Workshops Proceedings
Volume 1

Workshop on Massive Open Online Courses (moocshop)

Workshop Co-Chairs:

Zachary A. Pardos
Massachusetts Institute of Technology

Emily Schneider
Stanford University

http://www.moocshop.org

Preface

The moocshop surveys the rapidly expanding ecosystem of Massive Open Online Courses (MOOCs). Since late 2011, when enrolment for Stanford's AI class went viral, MOOCs have been a compelling and controversial topic for university faculty and administrators, as well as the media and blogosphere. Research, however, has played a relatively small role in the dialogue about MOOCs thus far, for two reasons. The first is the quickly moving landscape, with course scale and scope as the primary drivers for many stakeholders. The second is that there has yet to develop a centralized space where researchers, technologists, and course designers can share their findings or come to consensus on approaches for making sense of these emergent virtual learning environments.

Enter the moocshop. Designed to foster cross-institutional and cross-platform dialogue, the moocshop aims to develop a shared foundation for an interdisciplinary field of inquiry moving forward. Towards this end, we invited researchers, technologists, and course designers from universities and industry to share their work on key topics, from analytics to pedagogy to privacy. Since the forms and functions of MOOCs are continuing to evolve, the moocshop welcomed submissions on a variety of modalities of open online learning. Among the accepted papers and abstract-only submissions, four broad categories emerged:

- Position papers that proposed lenses for analyses or data infrastructure required to lower the barriers for research on MOOCs
- Exploratory analyses towards designing tools to assess and provide feedback on learner knowledge and performance
- Exploratory analyses and case studies characterizing learner engagement with MOOCs
- Experiments intended to personalize the learner experience or affect the psychological state of the learner

These papers and abstracts are an initial foray into what will be an ongoing dialogue, including discussions at the workshop and a synthesis paper to follow based on these discussions and the proceedings. We are pleased to launch the moocshop at the joint workshop day for AIED and EDM in order to draw on the expertise of both communities and ground the workshop discussions in principles and lessons learned from the long community heritage in educational technology research. Future instantiations of the moocshop will solicit contributions from a variety of different conferences in order to reflect the broad, interdisciplinary nature of the MOOC space.

June, 2013
Zachary A. Pardos & Emily Schneider

Program Committee

Co-Chair: Zachary A. Pardos, *MIT, USA* (pardos@mit.edu)
Co-Chair: Emily Schneider, *Stanford, USA* (elfs@cs.stanford.edu)

Ryan Baker, *Columbia Teacher's College, USA*
Amy Collier, *Stanford University, USA*
Chuong Do, *Coursera, USA*
Neil Heffernan, *Worcester Polytechnic Institute, USA*
Jack Mostow, *Carnegie Mellon University, USA*
Una-May O'Reilly, *Massachusetts Institute of Technology, USA*
Zach Pardos, *Massachusetts Institute of Technology, USA*
David Pritchard, *Massachusetts Institute of Technology, USA*
Emily Schneider, *Stanford University, USA*
George Siemens, *Athabasca University, USA*
John Stamper, *Carnegie Mellon University, USA*
Kalyan Veeramachaneni - *Massachusetts Institute of Technology, USA*

Table of Contents

Position / Overview

Two Models of Learning: Cognition Vs. Socialization 1
Shreeharsh Kelkar

Welcome to the moocspace: a proposed theory and taxonomy for massive open online courses 2
Emily Schneider

Roll Call: Taking a Census of MOOC Student 10
Betsy Williams

Developing Data Standards and Systems for MOOC Data Science 17
Kalyan Veeramachaneni, Franck Dernoncourt, Colin Taylor, Zachary Pardos and Una-May O'Reilly

Assessment

Syntactic and Functional Variability of a Million Code Submissions in a Machine Learning MOOC 25
Jonathan Huang, Chris Piech, Andy Nguyen and Leonidas Guibas

Revisiting and Extending the Item Difficulty Effect Model 33
Sarah Schultz and Trenton Tabor

Using Argument Diagramming to Improve Peer Grading of Writing Assignments 41
Mohammad H. Falakmasir, Kevin D. Ashley and Christian D. Schunn

Interventions

Improving learning in MOOCs with Cognitive Science 49
Joseph Jay Williams

Measurably Increasing Motivation in MOOCs 55
Joseph Jay Williams, Dave Paunesku, Benjamin Heley and Jascha Sohl-Dickstein

Controlled experiments on millions of students to personalize learning 56
Eliana Feasley, Chris Klaiber, James Irwin, Jace Kohlmeier and Jascha Sohl-Dickstein

Engagement

Analysis of video use in edX courses 57
Daniel T. Seaton, Albert J. Rodenius, Cody A. Coleman, David E. Pritchard and Isaac Chuang

Exploring Possible Reasons behind Low Student Retention Rates of Massive Online Open Courses: A Comparative Case Study from a Social Cognitive Perspective 58
Yuan Wang

Using EEG to Improve Massive Open Online Courses Feedback Interaction 59
Haohan Wang, Yiwei Li, Xiaobo Hu, Yucong Yang, Zhu Meng and Kai-Min Chang

Collaborative Learning in Geographically Distributed and In-person Groups 67
Rene Kizilcec

Two Models of Learning: Cognition Vs. Socialization

Shreeharsh Kelkar[1]

[1] Massachusetts Institute of Technology
United States
skelkar@mit.edu

Abstract. In this paper, I bring out the contrasts between two different approaches to student learning: that of computational learning scientists and socio-cultural anthropologists, and suggest some implications and directions for learning research in MOOCs. Computational learning scientists see learning as a matter of imbibing particular knowledge propositions, and therefore understand teaching as a way of configuring these knowledge propositions in a way that takes into account the learner's capacities. Cultural anthropologists see learning as a process of acculturation or socialization--the process of becoming a member of a community. They see school itself as a social institution and the process of learning at school as a special case of socialization into a certain kind of learning style (Lave 1988); being socialized into this learning style depends on the kinds of social and cultural resources that a student has access to.

Rather than see these approaches as either right or wrong, I see them as productive leading to particular kinds of research: thus, while a computational model of learning leads to research that looks at particular paths through the course material that accomplish the learning of a concept, an anthropological approach would look at student-student and student-teacher forum dialog to see how students use language, cultural resources and the affordances of the forum itself to make meaning. I argue that a socialization approach to learning might be more useful for humanities courses where assignments tend to be essays or dialogue. Finally, I bring up the old historical controversy in Artificial Intelligence: between the Physical Symbol Systems hypothesis and situated action. I argue that some of the computational approaches taken up by the proponents of situated action may be useful exemplars to implement a computational model of learning as socialization.

Keywords: cultural anthropology, learning models, socialization

welcome to the moocspace: a proposed theory and taxonomy for massive open online courses

Emily Schneider[1]

[1] Lytics Lab, Stanford University,
Stanford, CA
elfs@cs.stanford.edu

Abstract. This paper describes a theoretical framework and feature taxonomy for MOOCs, with the goal of developing a shared language for researchers and designers. The theoretical framework characterizes MOOC design goals in terms of stances towards knowledge, the learner, and assessment practices, taking as a starting point the affordances of the Web and digital learning environments. The taxonomy encompasses features, course structures, and audiences. It can be mapped onto the theoretical framework, used by researchers to identify similar courses for cross-course comparisons, and by instructional designers to guide design decisions in different dimensions. Both the theory and the taxonomy are intended in the spirit of proposal, to be refined based on feedback from MOOC researchers, designers, and technologists.

Keywords: taxonomy, knowledge organization, MOOCs, online learning theory

1 Introduction

If learning is the process of transforming external information into internal knowledge, the Internet offers us a universe of possibilities. In this context, MOOCs are simply a well-structured, expert-driven option for openly accessible learning opportunities. As of mid-2013, the boundaries of the moocspace[1] remain contested, with opinions (data-driven or no) generated daily in the blogosphere, the mainstream media, and an increasing number of academic publications. Meanwhile, decisions being made at a breakneck speed within academic institutions, governmental bodies, and private firms. What of the earlier forms of teaching and learning should we bring forward with us into networked, digital space, even as its interconnected and virtual

[1] Other types of open online learning opportunities that lend themselves to be named with similar wordplay include the DIYspace (e.g. Instructables, Ravelry, MAKE Magazine), the Q-and-Aspace (e.g. Quora, StackOverflow), the OERspace (indexed by such services as OERCommons and MERLOT), the coursespace (freely available course syllabi and instructional materials that are not officially declared or organized as OER), and the gamespace (where to even begin?). Then there is Wikipedia, the blogosphere and newsites, curated news pages (both crowdsourced, e.g. Slashdot, and personalized, e.g. Pinterest), and the great morass of affinity groups and individual, information-rich webpages.

nature allow us to develop new forms? How can an interdisciplinary, distributed group of researchers, course designers, administrators, technologists, and commentators make sense of our collective endeavor?

Towards a shared language for the *how* and *what* we are creating with MOOCs, I offer two frameworks. Firstly, for orientation towards the goals we have when we design MOOCs, I propose a theoretical framework that characterizes our assumptions about knowledge, the learner, and assessments. The framework takes as a starting point the affordances of the Web and digital learning environments, rather than those of brick-and-mortar learning environments.

Secondly, for grounding in the concrete, I offer a taxonomy of MOOC features, structures, and audiences, designed to capture the broad scope of MOOCs in terms of lifelong learning opportunities. Each element of the taxonomy can be mapped onto the theoretical framework to make explicit the epistemological stances of designers. The taxonomy can be used by researchers as a way of identifying similar courses for cross-course comparisons, and by instructional designers as a set of guideposts for potential design decisions in different dimensions. Finally, in the closing section of the paper, I provide an example of mapping the theory onto features from the taxonomy and introduce an application of the taxonomy as the organizing ontology for a digital repository of research on MOOCs, also referred to as the moocspace. Each framework is meant as a proposal to be iterated upon by the community.

2 A Proposed Theory *(Orientation)*

MOOC criticism and design decisions have largely been focused on comparisons with brick-and-mortar classrooms: how do we translate the familiar into these novel digital settings? Can classroom talk be replicated? What about the adjustments to teaching made by good instructors in response to the needs of the class? It is imperative to reflect on what we value in in-person learning environments and work to maintain the nature of these interactions. But to properly leverage the networked, digital environment to create optimal learning opportunities for MOOC participants, we also need to compare the virtual to the virtual and explore opportunities to embody the core principles of cyberspace in a structured learning environment.

Techno-utopian visions for the Web have three dominant themes: participatory culture, personalization, and collective intelligence. Participatory culture highlights the low cost of producing and sharing digital content, enabled by an increasing number of authoring, curatorial, and social networking tools [1]. In this account, personal expression, engagement, and a sense of community are available to any individual with interest and time—an ideal that MOOCs have begun to realize with well-facilitated discussion boards, and somewhat, with peer assessment. Some individual courses have also encouraged learners to post their own work in a portfolio style. But overall there are not many activities in this vein that have been formalized in the moocspace.

Participatory culture's elevation of the self is echoed in the personalized infrastructure of Web services from Google to Netflix, which increasingly seek to use recommendation engines to provide customized content to all users. The algorithmic principles of this largely profit-driven personalization are extendable to learning environments, though desired outcomes for learning are more complex than the metrics used for business analytics--hence the need for learning analytics to develop robust and theory-driven learner models for adaptive environments. Visions of personalized digital learning include options for learners to engage with the same content at their own pace, or to be treated to differentiated instruction based on their preferences and goals [2]. In MOOCs this will require robust learner models based on interaction data and, likely, self-reported data as well. Analytics for this level of personalization in MOOCs have yet to be achieved but personalization is occurring even without adaptive algorithms, as distributed learners are primarily interfacing with content at their own machines, at their own pace. Finally, collective intelligence focuses on the vast informational network that is produced by and further enables the participatory, creative moments of the users of the Web [3]. Each individual learner in a MOOC enjoys a one-to-many style of communication that is enabled by discussion boards and other tools for peer-to-peer interaction. In the aggregate, this becomes many-to-many, a network of participants that can be tapped into or contributed to by any individual in order to share knowledge, give or get assistance with difficult problems, make sense of the expectations of faculty, or simply to experience and add to the social presence of the virtual experience.

These themes are embodied in a range of epistemological stances towards two core dimensions of learning environments: the location of knowledge and conceptions of the learner. Assessment is the third core dimension of the learning environment [4]. The technology enables a wide number of assessment types but the stances towards assessment follow not from the affordances of the Web but from the standard distinction between formative and summative assessments. However, instead of using this jargon, I choose language that reflects the nature of the interaction enabled by each type of assessment, as the central mechanism of learning in online settings are interactions among learners, resources, and instructors [5] Finally, it is important to note that this framework treats the instructor as a designer and an expert participant, which also leaves room for the expert role to be played by others such as teaching assistants.

***Knowledge*: Instructionist-participatory**

Where are opportunities to acquire or generate knowledge? Does knowledge live purely with the instructor and other expert participants or does it live in the broad universe of participants? Who has the authority to create and deliver content? Is the learning experience created solely by the course designers or is it co-created by learners?

***Learner*: Personalized-Collectivist**

Are learners cognitively and culturally unique beings, or members of a network? Do the learning opportunities in the course focus on the individual learner or on the interactions of the group?

***Assessment*: Evaluation-Feedback**

> What opportunities are provided for learners to make explicit their progress in knowledge construction? Are assessments designed to tell learners if they're right or to give them guidance for improvement?

The poles of each stance, as named above, are opposed to each other epistemologically, but one end is not necessarily preferable to the other. The choice between each stance is predicated on what is valued by the designer in a learning environment or learning experience, and what is known about effective instruction and learning activities from the learning sciences. Each feature of the course can be characterized along one or more of these dimensions (see Section 4.1). This means that multiple stances can exist in the same course.

3 A Proposed Taxonomy *(Grounding)*

The proposed taxonomy includes two levels of descriptive metadata. The first level characterizes course as a whole and is meant to evoke the broad set of opportunities available for sharing knowledge with MOOCs. The second level takes in turn each element of the interactive learning environment (ILE) and develops a list of possible features for the implementation of these elements, based on current and potential MOOC designs. The features on this level can also serve as a set of guidelines of options for course designers. In multiple iterations of the course, many of these fields will stay the same but others will change. Most fields will be limited to one tag but others could allow multiple (e.g. target audience in General Structure).

The architecture and options for metadata on learning objects has been a subject in the field for quite some time, as repositories for learning objects and OER have become more common. While I am somewhat remiss to throw yet another taxonomy into the mix, I believe that it is important to represent the unique role of MOOCs in an evolving ecosystem of lifelong learning opportunities. Because the content and structure of a MOOC is not limited by traditional institutional exigencies of limited seats or approval of a departmental committee and accreditation agencies, it becomes a vessel for knowledge sharing, competency development, and peer connections across all domains, from computer science to music production and performance.[2] As a technology it is agnostic to how it is used, which means that it can be designed in any way that our epistemological stances guide us to imagine. Education has goals ranging from knowledge development to civic participation and MOOCs can be explicitly designed to meet any of these goals.

3.1 General MOOC Structure

On the highest level, each MOOC needs to be characterized in terms of its subject matter, audience, and use. Table 1 presents the proposed categories and subcategories for the General MOOC Structure. With an eye towards future interoperability, where

[2] That said, there is an ongoing conversation about integrating MOOCs back into the pre-existing educational institutions, so the taxonomy must ne conversant with these efforts while also representing the vagaries of the moocspace as a separate ecosystem.

possible I use the terminology from the Learning Resources Metadata Initiative (LRMI) specification [7], or note in parentheses which LRMI field the moocspace categories could map onto.

Table 1. Categories and Subcategories for General MOOC Structure

- Name (LRMI)
- Numeric ID (auto-generated)
- Author (LRMI)
  - Faculty member
- Publisher (LRMI)
  - Affiliated university or other institution
- Platform
- inLanguage (LRMI)
  - primary language of resource
- Domain (*about*)
  - Computational /STEM – CS, math, science, computational social sciences, etc.
  - Humanist – humanities, non-computational social sciences, etc.
  - Professional – business, medicine, law, etc.
  - Personal – health, thinking, speaking, writing, art, music, etc.
- Level (*typicalAgeRange* or *educationalRole)*
  - Pre-collegiate; basic skills (i.e. gatekeeper courses, college/career-ready); undergraduate; graduate; professional development; life skills
- Target audience (*educationalRole*)
  - Current students, current professionals, lifelong learners
- Use (*educationalUse* or *educationalEvent*)
  - Public course (date(s) offered), content for "wrapped" in-person course (location and date(s) offered)
- Pace
  - Cohort-based vs. self-paced (*learningResourceType* or *interactivityType*)
  - Expected workload for full course (total hours, hours/week) (*timeRequired*)
- Accreditation
  - Certificate available
  - Transfer credit

3.2 Elements of the Interactive Learning Environment (ILE)

The ILE is made up of a set of learning objects, socio-technical affordances, and instructional and community design decisions. These features are created by the course designers -- instructors and technologists – and interpreted by learners throughout their ongoing interaction with the learning objects in the course, as well as the other individuals who are participating in the course (as peers or instructors).[3] The features of the ILE can be sorted into four distinct categories: instruction, content, assessment, and community. Table 2 lists out the possible features of the ILE, based on the current trends in MOOC design. As stated, this is a descriptive list - based on

[3] The individual- and group-level learning experiences that take place in the ILE are enabled by the technological infrastructure of the MOOC platform and mediated by learner backgrounds (e.g. prior knowledge, self-regulation and study habits) and intentions for enrolling [8] as well as the context in which the MOOC is being used (e.g. in a "flipped" classroom, with an informal study group, etc.). The relationship of these psychological and contextual factors to learning experiences and outcomes is a rich, multifaceted research area, which I put aside here to foreground the ILE and systematically describe the dimensions along which it varies.

the current generation of MOOCs – but will be expanded in the future, both to reflect new trends in MOOC design and to take a normative stance on potential design choices that are based in principles of the learning sciences or interface design. Some of the features are mutually exclusive (i.e. lecture types) but others could occur simultaneously in the same MOOC (i.e. homework structure). Most features will need to be identified by spending some time exploring the course, ideally while it is taking place.

Table 2. Features of ILE

Instruction • Lecture ▪ "traditional": 1-3 hrs/wk, 20+ mins each ▪ "segmented": 1-3 hrs/wk, 5-20 mins each ▪ "minimal": <1 hr/wk • Readings • Simulations/inquiry environments/virtual labs • Instructor involvement – range from highly interactive to "just press play"	**Content** • Domain (in General Structure) • Modularized ▪ Within the course ▪ connected with other MOOCs/OER • Course pacing ▪ Self-paced ▪ Cohort-based
Assessment • In-video quizzes ▪ multiple choice vs. open-ended • Homework structure ▪ Multiple-choice ▪ Open-ended problems ▪ Performance assessments ▪ Writing assignments or programming assignments ▪ Videos, slides, multimedia artefacts • Group projects • Practice problems (non-credit bearing) ▪ Grading form–Quantitative, Qualitative • Grading structure (relevant to all credit-bearing assessments) ▪ Autograded ▪ Peer assessment, self-assessment, both ▪ Multiple submissions	**Community** • Discussion board • Social Media - Facebook group, Google+ community, twitter hashtag, reddit, LinkedIn, etc. • Blogs / student journals (inside or outside of platform) • Video chat (G+ hangout, Skype) • Text chat

4 The Taxonomy, Applied

4.1 Example of course mapping

Each course feature can be mapped onto one or more epistemological stances. The course overall can then be characterized by the overall epistemological tendencies of the course features. Table 3 provides an example.

Table 3. Mapping “Crash Course in Creativity” to the Taxonomy

General	Name: Crash Course in Creativity Author: Tina Seeling Publisher: Stanford Platform: NovoEd Domain: personal-thinking	Level: life skills Target audience: lifelong learners Use: public course (fall 2012) Pace: cohort-based - **collectivist** Certificate: yes
ILE and Stances		
Instruction	Lecture: minimal – 5-10 mins/wk to inspire group projects – **participatory** Readings: free, from her book - **instructionist**	
Content	Not modularized - **instructionist**	
Assessment	One individual creative projects – **participatory, individualist** Three group creative projects – **participatory, collectivist** Peer grading with qualitative comments–**participatory, feedback, collectivist**	
Community	Discussion board – **participatory, collectivist**	
OVERALL	**Participatory, collectivist, feedback**	

4.2 Stances to guide best practices and analytics.

The stances are not normative but do help specify which traditions of instructional and interface design should be turned to for guidance in best practices for designing resources. For example: instructionist lecture videos should follow the principles of multimedia learning, including balancing and integrating visual and verbal representations, relying on segmented (and learner-paced) narratives, and providing signaling mechanisms for the upcoming structure and content of a lecture. [9] The underlying epistemologies can also provide guidance about the type of analytics that are appropriate to for characterizing success in the design of the MOOC. For example, group-level outcomes may be more compelling for a collectivist MOOC – what is the overall level of interaction between learners, what kind of social networks form, with group projects can we characterize group composition or dynamics that lead to higher grades?

4.3 Centralizing distributed science: a short description of the moocspace

The taxonomy is a high-level, qualitative categorization of MOOCs that will allow for meaningful comparison across shared metrics about the courses. The taxonomy will be most usefully implemented in the *moocspace* – a digitized repository of knowledge about the research and production of massive open online courses – so named because it is an abstraction and reflection of the larger moocspace. The MOOC, abstracted, will be the central object of the moocspace, attached to standard metrics about the course, as well as reports on any research that has been done with data from that MOOC.[4] Variations in metrics could be related to aspect of the course

[4] Developing a small, meaningful set of shared metrics for MOOCs is currently an open question. Higher education in the US is characterized by enrollment rates at the beginning of

design, which are formalized in the taxonomy. Beyond descriptive data, a transparent, well-organized research base will enable an incremental and cumulative set of evidence from both exploratory studies (e.g. building learner models based on observational data) and experiments on the multiplicity of instructional and interface design features. A well-documented experiment in a small number of MOOCs could be replicated elsewhere by other researchers, and the findings could be synthesized by a third group by comparing results across variations in course features.

The moocspace could also be expanded to include the content of the MOOC itself, if licensing decisions are made that will allow MOOCs to become re-usable and re-mixable pieces of OER. This implementation would involve paradata on the uses of MOOC materials and incorporate a community aspect where faculty who use the materials could talk about what worked or didn't work in their courses. Finally, the MOOC object could also be attached to open datasets on MOOCs. The individuals who using such datasets may not be inside the academy, which underscores the need to build a structure for sharing newly developed knowledge back with the community.

If the moocspace is to be implemented, we will need develop consensus on the features in the taxonomy, as well as a strategy for tagging existing courses (crowdsourced? local experts?) and for adding new features to the taxonomy.

Acknowledgments. Much thanks to Kimberly Hayworth for her role in the initial brainstorming of MOOC features; thanks also to the Lytics Lab, Roy Pea, Zach Pardos, and the anonymous moocshop reviewers for feedback on earlier versions.

4 References

1. Jenkins, H. (2009) *Confronting the challenges of participatory culture: Media education for the 21st century*. Cambridge, MA: MIT Press.

2. National Educational Technology Plan. (2010). *Transforming American Education: Learning Powered by Technology.* Washington, DC: US Department of Education, Office of Educational Technology.

3. Lévy, P., & Bonomo, R. (1999). *Collective intelligence: Mankind's emerging world in cyberspace*. Perseus Publishing.

4. Brown, A. L., & Cocking, R. R. (2000). *How people learn*. J. D. Bransford (Ed.). Washington, DC: National Academy Press.

5. Anderson, T. "Towards a theory of online learning." (2004) *Theory and practice of online learning*: 3-31. Athabasca University, retrieved from http://cde.athabascau.ca/online_book/ch2.html

6. LRMI Specification Version 1.1 (Apr 28, 2013). www.lrmi.net/the-specification

7. Grover, S., Franz, P., Schneider, E. and Pea, R. (2013) "The MOOC as Distributed Intelligence: Dimensions of a Framework for the Design and Evaluation of MOOCs." In *Proceedings of the 10th International Conference on Computer Supported Collaborative Learning* (Madison, WI, June 16-19).

8. Mayer, R E., ed. (2005) *The Cambridge handbook of multimedia learning*. Cambridge University Press.

the semester, and persistence rates and completion rates over time. In addition to enrollment and activity rates initially and over time, for open courses it may be more appropriate to examine levels of engagement, time-on-task, or participation on the discussion forum.

Roll Call:
Taking a Census of MOOC Students

Betsy Williams[1],

[1] Stanford University, Graduate School of Education, 520 Galvez Mall,
CERAS Building, 5th Floor, Stanford, CA, 94305, USA
{betsyw@stanford.edu}

Abstract. This paper argues for spending resources on taking a high quality census or representative survey of students on who enroll with all major MOOC platforms. Expanded knowledge of current students would be useful for business and planning, instruction, and research. Potential concerns of cost, privacy, stereotype threat, and maladaptive use of the information are discussed.

Keywords: MOOC population, education, data collection, survey, demography

1 Introduction

Quantitative education researchers are accustomed to piecing together complex analyses from the rather lifeless data available from administrative records and test scores. The fine-grained data collected by MOOCs—including detailed knowledge of students' attendance and attention patterns, response on formative and summative assessments, and discussions with instructors and fellow students—offer an opportunity for much greater understanding of teaching and learning.

Unfortunately, MOOCs are not making the most out of their big data because they are not collecting enough data on students' backgrounds. Borrowing Bayesian terms, platforms have few priors on students, even though these priors can have great predictive power if paired with existing knowledge, from fields like developmental psychology and higher education theory.

The major platforms optimize sign up to make becoming part of the platform as quick as possible, leaving students mostly mysterious. EdX requests a few valuable pieces of demographic data upon registration, asking for voluntary identification by gender, year of birth, level of education completed, and mailing address without a clear reason why.[1] Coursera's information gathering is more like social media or a dating site, encouraging students who visit the profile page to share their age, sex, and location. As part of its "About Me" prompt, Coursera suggests that among other things users might share "what you hope to get out of your classes," while EdX asks the question in an open-ended text box upon registration. While these questions yield some of the data that is valuable for improving courses, the platforms, and education

[1] No one reads terms of service [1].

research, I argue that the platforms should collect more key data, clearly identified as information that will not be sold or used for targeted marketing or for student evaluation.

The paper first describes the fields most useful for analysis based on priors, and then it explains the benefits to platform development, instructional quality, and research. Potential drawbacks are discussed, including cost, privacy concerns, the risk of invoking stereotype threat, and the potential for undesirable changes to arise from this information.

2 Prior Information about Students

Given infinite data storage and infinite indulgence on the part of MOOC students, knowing every scrap of data about students might allow for inspired analyses and eerily predictive machine learning exercises. However, a more humble conception of student data would ably fulfill our research needs.

Core demographic information includes year of birth, gender, and race/ethnicity.[2] Asking users for their current city or place of residence should generate more accurate location results than IP address tracing or the information provided to appear on a semi-public profile. Combined with place of origin and native language, these questions provide a sketch of a student's likely history and culture.

A MOOC-run survey would also provide the opportunity to ask questions less often available in administrative education data, although extremely useful for understanding who enrolls. Although sensitive, questions about socioeconomic status and living situation would be tremendously helpful; for instance, is a student living with family, and to which generation does that student belong?

Adult students' lives are increasingly complex, and questions about work and education history should do their best to capture this. If a student's highest degree is a high school diploma (or equivalent), then have they ever enrolled in higher education? If so, in how many institutions? How many years and months would they estimate this spanned? Were they primarily taking full time or part time loads? What was the name of their primary institution, and what was their most recent course of study pursued? Those who have earned bachelor's degrees or higher should face similar questions. For all students, questions about previous or concurrent MOOC use would be very valuable. Work history can get a similar treatment, identifying such things as area of employment, and full- and part-time scheduling.

Although students themselves may not be entirely clear on the point yet, questions about educational and career goals, along with goals for the course, would be extremely useful. This information is captured to some extent in existing questions or for particular research. However, this may be incomplete or collected only in a piecemeal fashion. For instance, a study on learner patterns surveys the students in

[2] Race and ethnicity are social constructs whose meaning greatly varies by national context. For instance, being white in Norway has a different social meaning than being white in South Africa. And Belgium is split by a key ethnic marker—Walloon versus Fleming—that does not matter in other countries. Thus, choices for race/ethnicity should be based on the selected country of origin and/or country of residence.

one course, asking for intentions in the course, current employment status, years of work history, and highest degree attained [2].

Valuable information from surveys need not all be based on recall or opinion. Meaningful priors about academic preparation in particular fields can be generated by computer adaptive test questions in key content domains, based on existing work in psychometrics. Behavioral economics shows that survey questions can measure levels of risk aversion (asking for preferences between a gamble for $X and receiving $Y with certainty) and time discounting on money (asking about preferences for receiving $X now or $Y at a certain point in the future).[3]

Finally, there's a useful realm of information about how students use the platform. Within a class, how much time do they plan to devote, how do they plan to interact with peers, and will they use external supports, such as tutors, websites, and textbooks? What modes of access to the course are available to them? In particular, what electronic devices are available to them, is their use of the devices limited, and what kind of Internet access is available?

3 Value for Planning and Strategy

The background information on users discussed above provides extremely valuable data for the operations of the course platforms. Let us stipulate that there are limitations on the data being used for targeted marketing purposes. Even so, having aggregate background information on who is using which MOOCs is a huge advance.

In a traditional business mindset, the primary questions would be who is willing and who is able to pay. However, more advanced uses could help a course recommendation engine distinguish who is taking the course as a consumption good versus as investment in their future; the follow-up courses the students are interested in may be vastly different.

The survey may also suggest a greater than anticipated demand for classes taught at a certain level or on a certain topic. Students' locations, educations, and work histories might help the platform identify other institutions that may be good partners, either because they are very well-represented or under-represented.

4 Value for Instructional Design

A strong finding in educational research is that there is not a single correct way to teach or structure a course. Instead, learners matter, and knowledge about the students and their characteristics is important for teaching well [3]. Knowing more about the students also allows instructors to effectively call on their existing knowledge and address likely misconceptions; this is part of Pedagogical Content Knowledge [3] and a prominent contribution of Piagetian constructivism [4].

[3] For the most accurate answers, survey takers would actually receive the payout they say they prefer, subject to a gamble or delay as the case may be.

For example, knowing the age distribution and native languages of students can improve instruction. Instructors may choose allusions, words, and examples better.

An inherent challenge within the online classroom is that some feedback that is obvious in a physical context is not available. One student falling asleep in a lecture hall is far more obvious and effective of a signal than a thousand who never rewind the recorded lectures. While learning analytics is tackling this paucity of data in clever ways, we would also benefit greatly from leaning on priors. Imagine two students who do not watch the second week lecture by the beginning of week three: one has a doctoral degree in the field, while the other is a high school graduate who has attended several different institutions of higher education and intends to take a course for professional development. Applying theory to this prior knowledge, we might think the former finds the course matter unnecessary to review, while the latter may be struggling to stay motivated in the class.

In short, better prior knowledge can be paired with data collected in courses to better identify how students are learning the course content and improve the course.

5 Value for Education Research

MOOC populations are so wildly self-selected, and the field so new, that external validity is extremely questionable. At best, we might extend findings in a class to perhaps the same class the next time it is taught or use the results to develop hypotheses and learning theory.

While there is great value in using research to improve a single course, ideally the lessons could be transferred more broadly, so that the effort of analysis pays greater dividends. However, results cannot generalize until the population of the study is understood; once more is known about incoming characteristics of MOOC students who were studied, researchers can seek other classes that resemble them in salient details.

More concretely, MOOCs offer radical levels of access to education, and so they include many non-traditional and out-of-school learners. These nontraditional learners can be elusive research subjects, and there is also great diversity among their numbers. Having additional background data allows us to tag them and better understand their behavior. If a course platform is successful with a particular college level course and is contemplating recommending it to a partner community college, it would be wise to understand how students of different backgrounds performed. The inference is not direct, but it is far more useful than a recommendation based on coarser data.

The MOOC is also a fantastic platform for learning about how everyone learns, not just how self-selected MOOC users learn. The large number of students and the computerized means of instruction mean that MOOCs are very amenable to experimentation and careful observation. In addition, the very design of MOOCs strips down the traditional classroom; greater insight about learning and traditional instruction can come from adding back in some of these elements that are taken for granted in other classrooms.

Yet again, the great advantages of MOOCs as a place for learning research have the caveat that results are hard to generalize. However, if researchers control for the observable background data of the students who opt into MOOCs, their results will be far more plausibly applicable to a wide array of classes.

A key challenge within the online classroom is that feedback that may be obvious in a physical context, such as real-time indications of student engagement or confusion, is usually not available. One student falling asleep in a lecture hall is far more obvious and effective of a signal than a thousand who never rewind the recorded lectures. While learning analytics is tackling this paucity of data in clever ways, we would also benefit greatly from leaning on priors. Imagine two students who do not watch the second week lecture by the beginning of week three: one has a doctoral degree in the field, while the other is a high school graduate who has attended several different institutions of higher education and intends to take a course for professional development. Applying theory to this prior knowledge, we might think the former finds the course matter unnecessary to review, while the latter may be struggling to stay motivated in the class.

In short, better prior knowledge can be paired with data collected in courses to better identify how students are learning the course content.

6 Concerns and Limitations

There are genuine concerns with collecting this much data. Here, I discuss cost, privacy, stereotype threat, and maladaptive use. I present these cursorily not to dismiss these points, but to begin what must be a larger discussion.

6.1 Cost

Course platforms are in a unique position It can be extremely costly to ask survey questions. User attention is limited and a choice to ask an additional question may implicitly limit their engagement later during the session, or even drive them away from the service at the extreme. Higher quality survey data can be generated by using internal resources to follow up with non-responders; higher response rates can also be generated by incentives, such as monetary payment, entry in a lottery, or access to a premium site feature. In addition, comprehensive surveys offered by a platform itself can be more easily embedded in the site, making the survey more available and more salient.

Administering a vast survey at the site level also better captures students who might be over-sampled if asked class-by-class. Cross-course analyses can be conducted more easily if the relatively permanent, detailed background information is available at the platform level, rather than asked for in individual courses.

Stratified sampling methods could be used to reduce the burden on students and the cost burden on the platform. For instance, core questions could be asked of the main sample of students, while additional long forms of the survey ask different questions of different students. The aggregate picture can be pieced together with a smaller burden on most students and a lower cost to the platform. While this is less than ideal, it may be a necessary tradeoff in some cases.

6.2 Privacy

Privacy concerns are important and complex, and researchers are used to the question of balancing privacy concerns against the benefit of the research. The more background data a platform collects, the more risk that personally identifiable information about subjects is available through composite reports or if the data are intercepted. Access security and care in reporting results are thus crucial and should be considered ahead of time.

Because of these concerns or others, some students may wish not to provide information, which could systematically bias the survey sample, making our inferences worse. Some students who wish to opt out may be reassured if the reasons for the research and the protection of the data are made clear. Others may be more comfortable with anonymized options for responding or techniques designed for collecting sensitive data. [5]

6.3 Stereotype Threat and Maladaptive Use of Information

Arguably, the Internet provides one of the few places in society where people are not forced to reveal information about their social position, which may be of value in itself.[4] A powerful strand of research in social psychology suggests that invoking identities that are attached to negative stereotypes can hinder educational performance; people are especially vulnerable to this "stereotype threat" if they feel there is a power imbalance and that they are being defined by others' judgments [8]. This threat could both change answers provided and potentially harm the student. However, a sustained harm seems unlikely to result from the trigger of a few questions on a survey; rather, the underlying negative social context or vulnerability might be in play. It would be unfortunate if a detailed survey triggered stereotype threat, even temporarily, but making sure the questions are seen as low-stakes could help.

There may also be a risk that instructors change their courses in unintended ways if they find out more about the students. An instructor might make a college-level course less rigorous if he finds out high school students are enrolled, for instance. While this raises concerns, it is ultimately up to policy and instructors' judgment.

[4] Perhaps the Internet is the place where students "will not be judged by the color of their skin, but by the content of their character." [6] Less seriously, "On the Internet, nobody knows you're a dog." [7]

7 Conclusion

Platform operations, instructional design, and educational research would all benefit from collecting more systematic background data about students. Better knowledge about who takes MOOCs is crucial at this stage in their lifetime. I propose not only a census of MOOC users on each platform, capturing a snapshot of users today, but an ongoing effort to capture these detailed demographic snapshots at least every three years.

Acknowledgments. Many thanks to Eric Bettinger, Susanna Loeb, Tom Dee, Roy Pea, Mitchell Stevens, and participants in the Lytics Lab for valuable discussions that have influenced this paper.

References

1. Gindin, S.E.: Nobody Reads Your Privacy Policy or Online Contract? Lessons Learned and Questions Raised by the FTC's Action Against Sears. 8 Nw. J. Tech & Intell. Prop. 1, 1--38 (2009)
2. Kizilcec, R.F, Piech, C., Schneider, E.: Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses. In: 3rd Conference on Learning Analytics and Knowledge. Leuven, Belgium (2013)
3. Shulman, L.S.: Knowledge and Teaching: Foundations of the New Reform. Harvard Educational Rev. 57, 1--22 (1987)
4. Ackermann, E.K.: Constructing Knowledge and Transforming the World. In: Tokoro, M., Steels, L. (eds.) A Learning Zone of One's Own: Sharing Representations and Flow in Collaborative Learning Environments. pp. 15--37. IOS Press, Amsterdam, Berlin, Oxford, Tokyo, Washington, DC (2004)
5. Du, W., Zhan, Z.: Using Randomized Response Techniques for Privacy-Preserving Data Mining. In: 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 505-510. ACM, New York (2003)
6. King, M.L.: I Have a Dream. In: Carson, C., Shepard, K. (eds.) A Call to Conscience: The Landmark Speeches of Dr. Martin Luther King, Jr. IPM/Warner Books, New York (2001)
7. Steiner, P.: On the Internet, Nobody Knows You're a Dog. The New Yorker LXIX, 20, p. 61 (1993)
8. Walton, G.M., Paunesku, D., Dweck, C.S.: Expandable Selves. In: Leary, M.R., Tangney, J.P. (eds.) The Handbook of Self and Identity, Second Edition, pp. 141--154. Taylor and Francis, New York (2012)

MOOCdb: Developing Data Standards for MOOC Data Science

Kalyan Veeramachaneni, Franck Dernoncourt,
Colin Taylor, Zachary Pardos, and Una-May O'Reilly

Massachusetts Institute of Technology, USA.
{kalyan,francky,colin_t,zp,unamay}@csail.mit.edu

1 Introduction

Our team has been conducting research related to mining information, building models, and interpreting data from the inaugural course offered by edX, *6.002x: Circuits and Electronics*, since the Fall of 2012. This involves a set of steps, undertaken in most data science studies, which entails positing a hypothesis, assembling data and features (aka properties, covariates, explanatory variables, decision variables), identifying response variables, building a statistical model then validating, inspecting and interpreting the model. In our domain, and others like it that require behavioral analyses of an online setting, a great majority of the effort (in our case approximately 70%) is spent assembling the data and formulating the features, while, rather ironically, the model building exercise takes relatively less time. As we advance to analyzing cross-course data, it has become apparent that our algorithms which deal with data assembly and feature engineering lack cross-course generality. This is not a fault of our software design. The lack of generality reflects the diverse, ad hoc data schemas we have adopted for each course. These schemas partially result because some of the courses are being offered for the first time and it is the first time behavioral data has been collected. As well, they arise from initial investigations taking a local perspective on each course rather than a global one extending across multiple courses.

In this position paper, we advocate harmonizing and unifying disparate "raw" data formats by establishing an open-ended standard data description to be adopted by the entire education science MOOC oriented community. The concept requires a schema and an encompassing standard which avoid any assumption of data sharing. It needs to support a means of sharing *how the data is extracted, conditioned and analyzed.*

Sharing scripts which prepare data for models, rather than data itself, will not only help mitigate privacy concerns but it will also provide a means of facilitating intra and inter-platform collaboration. For example, two researchers, one with data from a MOOC course on one platform and another with data from another platform, should be able to decide upon a set of variables, share scripts that can extract them, each independently derive results on their own data, and then compare and iterate to reach conclusions that are cross-platform as well as cross-course. In a practical sense, our goal is a standard facilitating insights

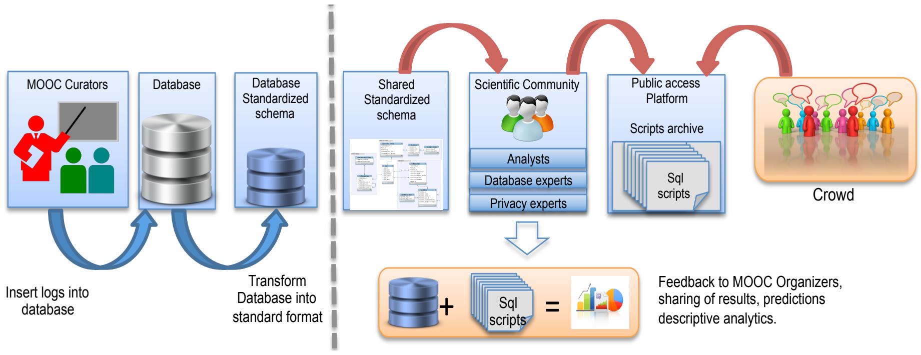

MOOC Curators
Database
Database Standardized schema
Insert logs into database
Transform Database into standard format
Shared Standardized schema
Scientific Community
Analysts
Database experts
Privacy experts
Public access Platform
Scripts archive
Sql scripts
Crowd
+
Sql scripts
=
Feedback to MOOC Organizers, sharing of results, predictions descriptive analytics.

Fig. 1. This flowchart represents the context of a standardized database schema. From left to right: Curators of MOOC course format the raw transaction logs into the schema and populate either private or public databases. This raw database is transformed into a standard schema accepted by the community, (like the one proposed in this paper) and is exposed to the analytics community, mostly researchers, who develop and share scripts, based upon it. The scripts are capable of extracting study data from any schema-based database, visualizing it, conditioning it into model variables and/or otherwise examining it. The schema is unifying while the scripts are the vehicle for cross-institution research collaboration and direct experimental comparison.

from data being shared without data being exchanged. It will also enable research authors to release a method for recreating the variables they report using in their published experiments.

Our contention is that the MOOC data mining community - from all branches of educational research, should act immediately to engage in consensus driven discussions toward a means of standardizing data schema and building technology enablers for collaborating on data science via sharing scripts, results in a practical, directly comparable and reproducible way. It is important to take initial steps now. We have the timely opportunity to avoid the data integration chaos that has arisen in fields like health care where large legacy data, complex government regulations and personal privacy concerns are starting to thwart scientific progress and stymy access to data. In this contribution, we propose a standardized, cross-course, cross-platform, database schema which we name as *"MOOCdb"*. [1]

We proceed by describing our concept and what it offers in more detail in Section 2. Section 3 details our proposed the data schema systematically. Section 4 shows, with a use case, how the schema is expressive, supportive and reusable. Section 5 concludes and introduces our current work.

[1] We would like to use the MOOCshop as a venue for introducing it and offering it up for discussion and feedback. We also hope to enlist like minded researchers willing to work on moving the concept forward in an organized fashion, with plenty of community engagement.

2 Our Concept and What it Offers

Our concept is described as follows, and as per Figure 1:

- It identifies two kinds of primary actors in the MOOC eco-system: *curators* and *analysts*. Curators collect raw behavioral data expressing MOOC students' interaction with online course material and then transfer it to a database, often as course content providers or course platform providers. *Analysts* reference the data to examine it for descriptive, inferential or predictive insights. The role of the analysts is to visualize, descriptively analyze, use machine learning or otherwise interpret some set of data within the database. Analysts extract, condition (e.g. impute missing values, de-noise), and create higher level variables for modeling and other purposes from the data. To perform this analysis, they first transform the data into the standard schema and compose *scripts* or use publicly available scripts when it suffices. They also contribute their scripts to the archive so others can use.
- It identifies two types of secondary actors: the *crowd*, and the data science experts (database experts and privacy experts). When needs arise, the community can seek the help of the *crowd* in innovative ways. Experts contribute to the community by providing state-of-the art technological tools and methods.
- A common standardized and shared schema into which the data is stored. The schema is agreed upon by the community, generalizes across platforms and preserves all the information needed for data science and analytics.
- A shared community-oriented repository of data extraction, feature engineering, and analytics scripts.
- Over time the repository and the schema, both open ended, grow.

This concept offers the following:
The benefits of standardization: The data schema standardization implies that the raw data from every course offering will be formatted the same way in its database. It ranges from simple conventions like storing event timestamps in the same format to common tables, fields in the tables, and relational links between different tables. It implies compiling a scientific version of the database schema that contains important events, fields, and dictionaries with highly structured data is amenable for scientific discovery. Standardization supports cross-platform collaborations, sharing query scripts, and the definition of variables which can be derived in exactly the same way for irrespective of which MOOC database they come from.
Concise data storage: Our proposed schema is "loss-less", i.e. no information is lost in translating raw data to it. However, the use of multiple related tables provides more efficient storage.
Savings in effort: A schema speeds up database population by eliminating the steps where a schema is designed. Investigating a dataset using one or more existing scripts helps speed up research.
Sharing of data extraction scripts: Scripts for data extraction and descriptive statistics extraction will be open source and can be shared by everyone.

Some of these scripts could be very general and widely applicable, for example: "For every video component, provide the distribution of time spent by each student watching it?" and some would be specific for a research question, for example generation of data for Bayesian knowledge tracing on the problem responses. These scripts could be optimized by the community and updated from time to time.

Crowd source potential: Machine learning frequently involves humans identifying explanatory variables that could drive a response. Enabling the crowd to help propose variables could greatly scale the community's progress in mining MOOC data. We intentionally consider the data schema to be independent of the data itself so that people at large, when shown the schema, optional prototypical synthetic data and a problem, can posit an explanatory variable, write a script, test it with the prototypical data and submit it to an analyst. The analyst can assess the information content in the variable with regards to the problem at hand and rank and feed it back to the crowd, eventually incorporating highly rated variables into learning.

A unified description for external experts: For experts from external fields like"Very Large Databases/Big Data" or "Data Privacy", standardization presents data science in education as unified. This allows theme to technically assist us with techniques such as new database efficiencies or privacy protection methods.

Sharing and reproducing the results: When they publish research, analysts share the scripts by depositing them into a public archive where they are retrievable and cross-referenced to their donor and publication.

Our concept presents the following challenges:

Schema adequacy: A standardized schema must capture all the information contained in the raw data. To date, we have only verified our proposed schema serves the course we investigated. We expect the schema to significantly change as more courses and offerings are explored. It will be challenging to keep the schema open ended but not verbose. While a committee could periodically revisit the schema, a more robust approach would be to let it evolve through open access to extension definitions then selection of good extensions via adoption frequency. This would embrace the diversity and current experimental nature of MOOC science and avoid standard-based limitations. One example of a context similar to the growth of MOOCs is the growth of the internet. HTML and Web3.0 did not rein in the startling growth or diversity of world wide web components. Instead, HTML (and its successors and variants) played a key role in delivering content in a standardized way for any browser. The semantic web provides a flexible, community driven, means of standards adoption rather than completely dictating static, monolithic standards. We think there are many lessons to learn from the W3C initiative. To whit, while we provide ideas for standards below, we propose that, more importantly, there is a general means of defining standards that allow interoperability, which should arise from the examples we are proposing.

Platform Support: The community needs a website defining the standard data template and a platform assisting researchers in sharing scripts. It requires tests for validating scripts, metrics to evaluate new scripts and an repository of scripts with efficient means of indexing and retrieval.

Motivating the crowd: How can we encourage large scale script composition and sharing so the crowd will supply explanatory variables? How can we provide useful feedback when the crowd is not given the data? KAGGLE provides a framework from which we can draw inspiration, but it fundamentally differs from what we are proposing here. KAGGLE provides a problem definition, a dataset that goes along with it, whereas we are proposing that we share the schema, propose a problem, give an example of a set of indicators and the scripts that enabled their extraction, and encourage users to posit indicators and submit scripts. Such an endeavor requires us to: define metrics for evaluation of indicators/features given the problem, provide synthetic data (under the data schema) to allow the crowd to test and debug their feature engineering scripts, and possibly visualizations of the features or aggregates over their features (when possible), and most importantly a dedicated compute resource that will perform machine learning and evaluate the information content in the indicators.

3 Schema description

We surveyed a typical set of courses from Coursera and edX. We noticed three different modes in which students engage with the material. Students observe the material by accessing all types of resources. In the second mode they submit material for evaluation and feedback. This includes problem check-ins for lecture exercises, homework and exams. The third mode is in which they collaborate with each other. This includes posting on forums and editing the wiki. It could in future include more collaborative frameworks like group projects. Based on these three we divide the database schema into three different tables. We name these three modes as *observing*, *submitting* and *collaborating*. We now present the data schema for each mode capturing all the information in the raw data.

3.1 The observing mode

In this mode, students simply browse and observe a variety of resources available on the website. These include the *wiki*, *forums*, *lecture videos*, *book*, *tutorials*. Each unique resource is usually identifiable by a *URL*. We propose that data pertaining to the observing mode can be formatted in a 5-tuple table: *u_id* (*user id*), *r_id* (*resource id*), *timestamp*, *type_id*, *duration*. Each row corresponds to one click event pertaining to student. Two tables that form the dictionaries accompany this event table. The first one maps each unique *url* to *r_id* and the second one maps *type_id* to resource type, i.e., *book*, *wiki*. Splitting the tables into event and dictionary tables allows us to reduce the sizes of the tables significantly. Figure 4 shows the schema and the links.

3.2 The submitting mode

Similar to the table pertaining to the observing mode of the student, we now present a structured representation of the problem components of the course.

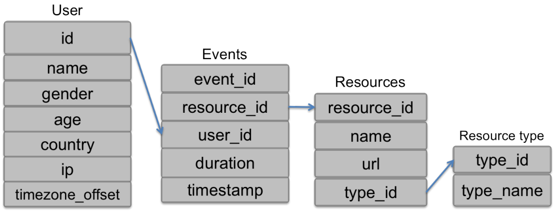


Fig. 2. Data schema for the observing mode

A typical MOOC consists of assignments, exams, quizzes, exercises in between lectures, labs (for engineering and computer science). Unlike campus based education, students are allowed to submit answers and check them multiple times. Questions can be multiple choice or a student can submit an analytical answer or even a program or an essay. Assessments are done by computer or by peers to evaluate the submissions [1]. We propose the following components:

Submissions table: In this table each submission made by a student is recorded. The 5 tuple recorded is *u_id*, *p_id*, timestamp, the answer, and the attempt number.

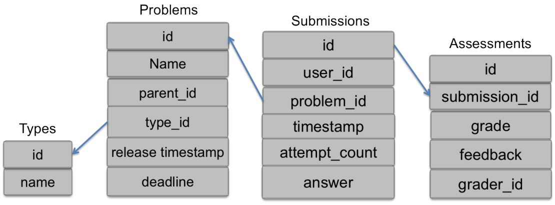


Fig. 3. Data schema for the submitting mode.

Assessments table: To allow for multiple assessments this table is created separately from the submissions table. In this table each assessment for each submission is stored as a separate row. This separate table allows us to reduce the size since we do not repeat the *u_id* and *p_id* for each assessment.

Problems table: This table stores the information about the problems. We *id* the smallest problem in the entire course. The second field provides the name for the problem. The problem is identified if it is a sub problem within another problem by having a parent *id*. Parent *id* is a reflective field in that its entries are one of the problem *id* itself. Problem type *id* stores the information about whether it is a homework, exercise, midterm or final. The table also stores the problem release date and the problem submission deadline date as two fields. Another table stores the id for problem types.

3.3 The Collaborating mode

Student interact and collaborate among themselves throughout the course duration through forums and wiki. In forums a student either initiates a new thread or responds to an existing thread. Additionally students can up vote, and down vote the answers from other students. In wiki students edit, add, delete and initiate a new topic. To capture this data we form the following tables with the following fields:

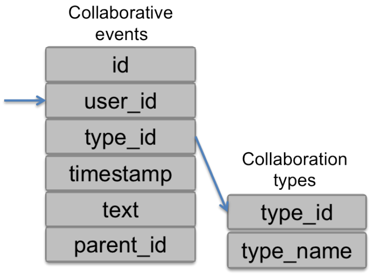


Fig. 4. Data schema for collaborating mode

Collaborations table: In this table each attempt made by a student to collaborate is given an *id*. The 5 fields in this table are *u_id*, collaboration type (whether wiki or forum), timestamp, the pointer to the text inserted by this user, and the parent *id*. The last field is a reflective field as well.

Collaboration type table: In this table the collaboration type *id* is identified with a name as to whether it is a wiki or a forum.

4 The edX 6.002x case study

edX offered its first course *6.002x: Circuits and Electronics* in the Fall of 2012. 6.002x had 154,763 registrants. Of these, 69,221 people looked at the first problem set, and 26,349 earned at least one point on it. 13,569 people looked at the midterm while it was still open, 10,547 people got at least one point on the midterm, and 9,318 people got a passing score on the midterm. 10,262 people looked at the final exam while it was still open, 8,240 people got at least one point on the final exam, and 5,800 people got a passing score on the final exam. Finally, after completing 14 weeks of study, 7,157 people earned the first certificate awarded by MITx, showing that they successfully completed 6.002x.

The data corresponding to the behavior of the students was stored in multiple different formats and was provided to us. These original data pertaining to the observing mode was stored in files and when we transcribed in the database with fields corresponding to the names in the "*name-value*" it was about the size of around 70 GB. We imported the data into a database with the schema we described in the previous subsections. The import scripts we had to build fell into two main categories:

- reference generators, which build tables listing every user, resource and problem that were mentioned in the original data.
- table populators, which populate different tables by finding the right information and converting it if needed.

The sizes and the format of the resulting tables is as follows: submissions: 341 MB (6,313,050 rows); events: 6,120 MB (132,286,335 rows); problems: 0.08 MB; resources: 0.4 MB; resource types: 0.001 MB; users: 3MB. We therefore reduced the original data size by a factor of 10 while keeping most of the information. This allows us to retrieve easily and quickly information on the students' activities. For example, if we need to know what is the average number of pages in the book a student read, it would be around 10 times faster. Also, the relative small size of the tables in this format allows us to do all the work in memory on any relatively recent desktop computer. For more details about the analytics we performed as well as the entire database schema we refer the reader to [2] [2]

5 Conclusions and future work

In this paper, we proposed a standardized data schema and believe that this would be a powerful enabler for ours and others researchers involved in MOOC data science research. Currently, we after building databases based on this schema we are developing a number of analytic scripts that extract multiple attributes for a course. We intend to release them in the near future. We believe it is timely to envision an open data schema for MOOC data science research.

Finally, we propose that as a community we should come up with a shared standard set of features that could be extracted across courses and across platforms. The schema facilities sharing and re-use of scripts. We call this the "feature foundry". In the short term we propose that this list is an open, living handbook available in a shared mode to allow addition and modification. It can be implemented as a google doc modified by the MOOC community. At the moocshop we would like to start synthesizing a more comprehensive set of features and developing the handbook. Feature engineering is a complex, human intuition driven endeavor and building this handbook and evolving this over years will be particularly helpful.

References

1. Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., Koller, D.: Tuned models of peer assessment in MOOCs. In: Proceedings of The 6th International Conference on Educational Data Mining (EDM 2013). (2013)
2. Dernoncourt, F., Veeramachaneni, K., Taylor, C., O'Reilly, U.M.: Methods and tools for analysis of data from MOOCs: edx 6.002x case study. In: Technical Report, MIT. (2013)

[2] For the full MOOCdb database schema, see `http://bit.ly/MOOCdb`)

Syntactic and Functional Variability of a Million Code Submissions in a Machine Learning MOOC

Jonathan Huang, Chris Piech, Andy Nguyen, and Leonidas Guibas

Stanford University

Abstract. In the first offering of Stanford's Machine Learning Massive Open-Access Online Course (MOOC) there were over a million programming submissions to 42 assignments — a dense sampling of the range of possible solutions. In this paper we map out the syntax and functional similarity of the submissions in order to explore the variation in solutions. While there was a massive number of submissions, there is a much smaller set of unique approaches. This redundancy in student solutions can be leveraged to "force multiply" teacher feedback.

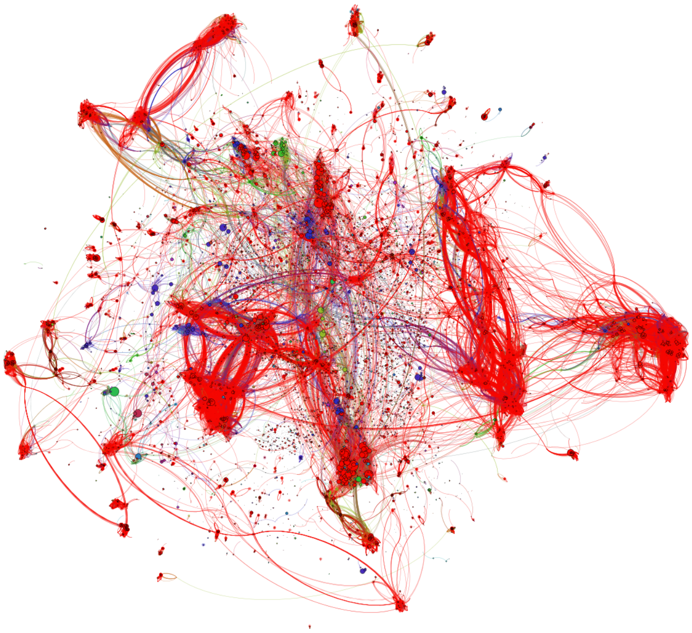


Fig. 1. The landscape of solutions for "gradient descent for linear regression" representing over 40,000 student code submissions with edges drawn between syntactically similar submissions and colors corresponding to performance on a battery of unit tests (red submissions passed all unit tests).

1 Introduction

Teachers have historically been faced with a difficult decision on how much personalized feedback to provide students on open-ended homework submissions such as mathematical proofs, computer programs or essays. On one hand, feedback is a cornerstone of the educational experience which enables students to learn from their mistakes. On the other hand, giving comments to each student can be an overwhelming time commitment [4]. In contemporary MOOCs, characterized by enrollments of tens of thousands of students, the cost of providing informative feedback makes individual comments unfeasible.

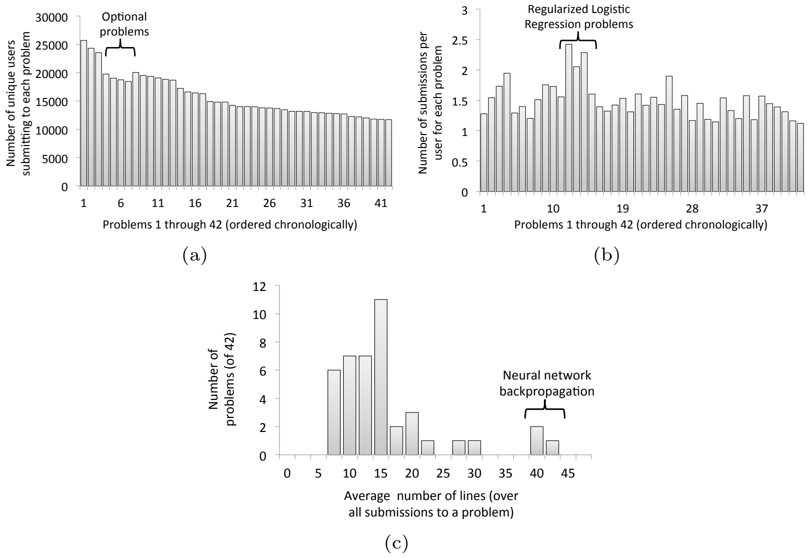
Interestingly, a potential solution to the high cost of giving feedback in massive classes is highlighted by the volume of student work. For certain assignment types, most feedback work is redundant given sufficiently many students. For example, in an introductory programming exercise many homework submissions are similar to each other and while there may be a massive number of submissions, there is a much smaller variance in the content of those submissions. It is even possible that with enough students, the entire space of reasonable solutions is covered by a subset of student work. We believe that if we can organize the space of solutions for an assignment along underlying patterns we should be able to "force multiply" the feedback work provided by a teacher so that they can provide comments for many thousands of students with minimal effort.

Towards the goal of force multiplying teacher feedback, we explore variations in homework solutions for Stanford's Machine Learning MOOC that was taught in Fall of 2011 by Andrew Ng (ML Class), one of the first MOOCs taught. Our dataset consists of over a million student coding submissions, making it one of the largest of its kind to have been studied. By virtue of its size and the fact that it constitutes a fairly dense sampling of the possible space of solutions to homework problems, this dataset affords us a unique opportunity to study the variance of student solutions. In our research, we first separate the problem of providing feedback into two dimensions: giving output based feedback (comments on the functional result of a student's program) and syntax based feedback (comments on the stylistic structure of the student's program). We then explore the utility and limitations of a "vanilla" approach where a teacher provides feedback only on the k most common submissions. Finally we outline the potential for an algorithm which propagates feedback on the entire network of syntax and output similarities. Though we focus on the ML Class, we designed our methods to be agnostic to both programming language, and course content.

Our research builds on a rich history of work into finding similarity between programming assignments. In previous studies researchers have used program similarity metrics to identify plagiarism [1], provide suggestions to students' faced with low level programming problems [2] and finding trajectories of student solutions [3]. Though the similarity techniques that we use are rooted in previous work, the application of similarity to map out a full, massive class is novel.

2 ML Class by the numbers

When the ML Class opened in October 2011 over 120,000 students registered. Of those students 25,839 submitted at least one assignment, and 10,405 submitted solutions to all 8 homework assignments (each assignment had multiple parts



(a)

(b)

(c)

Fig. 2. (a) Number of submitting users for each problem; (b) Number of submissions per user for each problem; (c) Histogram over the 42 problems of average submission line counts.

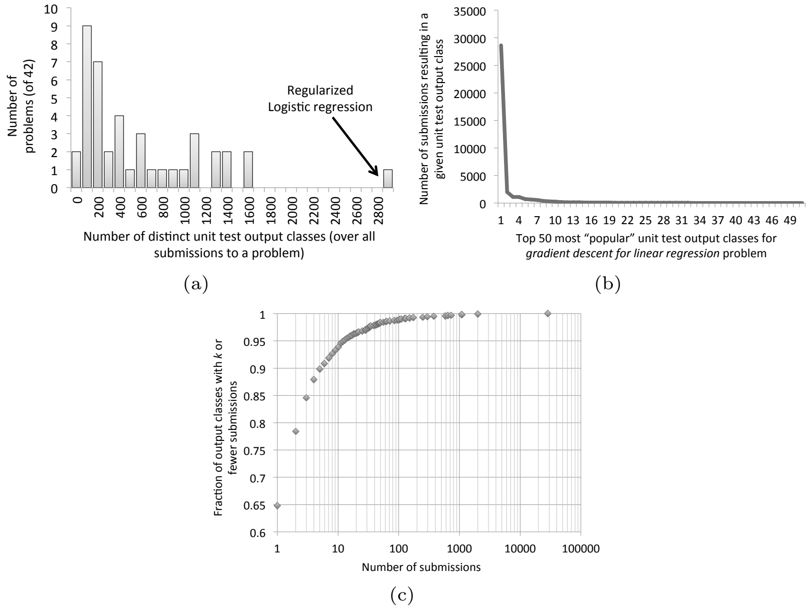
which combined for a total of 42 coding based problems) in which students were asked to program a short matlab/octave function. These homeworks covered topics such as regression, neural networks, support vector machines, among other topics. Submissions were assessed via a battery of unit tests where the student programs were run with standard input and assessed on whether they produced the correct output. The course website provided immediate confirmation as to whether a submission was correct or not and users were able to optionally resubmit after a short time window.

Figure 2(a) plots the number of users who submitted code for each of the 42 coding problems. Similarly, Figure 2(b) plots the average number of submissions per student on each problem and reflects to some degree its difficulty.

In total there were 1,008,764 code submissions with typical submissions being quite short — on average a submission was 16.44 lines long (after removing comments and other unnecessary whitespace). Figure 2(c) plots a histogram of the average line count for each of the 42 assignments. There were three longer problems — all relating to the backpropagation algorithm for neural networks.

3 Functional variability of code submissions

First, we examine the collection of unit test outputs for each submitted assignment (which we use as a proxy for *functional variability*). In the ML Class, the



(a)

(b)

(c)

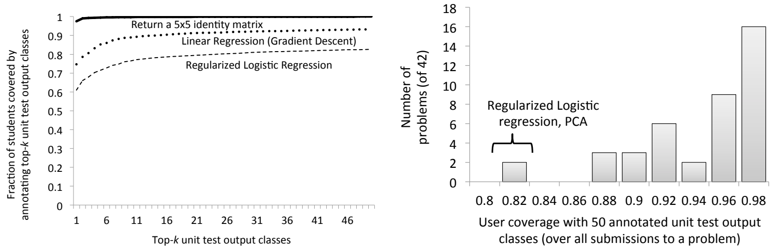
Fig. 3. (a) Histogram over the 42 problems of the number of distinct unit test outputs; (b) Number of submissions to each of the 50 most common unit test outputs for the "gradient descent for linear regression" problem; (c) Fraction of distinct unit test outputs with k or fewer submissions. For example, about 95% of unit test outputs owned fewer than 10 submissions.

unit test outputs for each program are a set of real numbers, and we consider two programs to be functionally equal if their unit test output vectors are equal.[1]

Not surprisingly in a class with tens of thousands of participants, the range of the outputs over all of the homework submissions can be quite high even in the simplest programming assignment. Figure 3(a) histograms the 42 assigned problems with respect to the number of distinct unit test outputs submitted by all students. On the low end, we observe that the 32,876 submissions to the simple problem of constructing a 5×5 identity matrix resulted in 218 distinct unit test output vectors. In some sense, the students came up with 217 wrong ways to approach the identity matrix problem. The median number of distinct outputs over all 42 problems was 423, but at the high end, we observe that the 39,421 submissions to a regularized logistic regression problem produced 2,992 distinct unit test outputs!

But were there *truly* nearly 3,000 distinct wrong ways to approach regularized logistic regression? Or were there only a handful of "typical" ways to be wrong and a large number of submissions which were each wrong in their own unique way? In the following, we say that a unit test output vector v *owns* a submission

[1] The analysis in Section 4 captures variability of programs at a more nuanced level of detail



(a) (b)

Fig. 4. (a) Number of students covered by the 50 most common unit test outputs for several representative problems; (b) Histogram over the 42 problems of number of students covered by the top 50 unit test outputs for each problem. Observe that for most problems, 50 unit test outcomes is sufficient for covering over 90% of students.

if that submission produced v when run against the given unit tests. We are interested in common or "popular" outputs vectors which own many submissions.

Figure 3(b) visualizes the popularity of the 50 unit class output vectors which owned the most submissions for the gradient descent for linear regression problem. As with all problems, the correct answer was the most popular, and in the case of linear regression, there were 28,605 submissions which passed all unit tests. Furthermore, there were only 15 additional unit test vectors which were the result of 100 submissions or more, giving some support to the idea that we can "cover" a majority of submissions simply by providing feedback based on a handful of the most popular unit test output vectors. On the other hand, if we provide feedback for only a few tens of the most popular unit test outputs, we are still orphaning in some cases thousands of submissions. Figure 3(c) plots the fraction of output vectors for the linear regression problem again which own less than k submissions (varying k on a logarithmic scale). The plot shows, for example, that approximately 95% of unit test output vectors (over $1,000$ in this case) owned 10 or fewer submissions. It would have been highly difficult to provide feedback for this 95% using the vanilla output-based feedback strategy.

To better quantify the efficacy of output-based feedback, we explore the notion of *coverage* — we want to know how many students in a MOOC we can "cover" (or provide output-based feedback for) given a fixed amount of work for the teaching staff. To study this, consider a problem P for which unit test output vectors $S = \{s_1, \ldots, s_k\}$ have been manually annotated by an instructor. This could be as simple as "good job!", to "make sure that your for-loop covers special case X". We say that a student is covered by S if every submitted solution by that student for problem P produces unit test outputs which lie in S. Figure 4(a) plots the number of students which are covered by the 50 most common unit test output vectors for several representative problems. By and large, we find that annotating the top 50 output vectors yields coverage of 90% of students or more in almost all problems (see Figure 4(b) for histogrammed output coverage over the 42 problems). However, we note that in a few cases, the top 50 output vectors might only cover slightly over 80% of students, and that even at

90% coverage, typically between 1000-2000 students are *not* covered, showing limitations of this “vanilla” approach to output-based feedback.

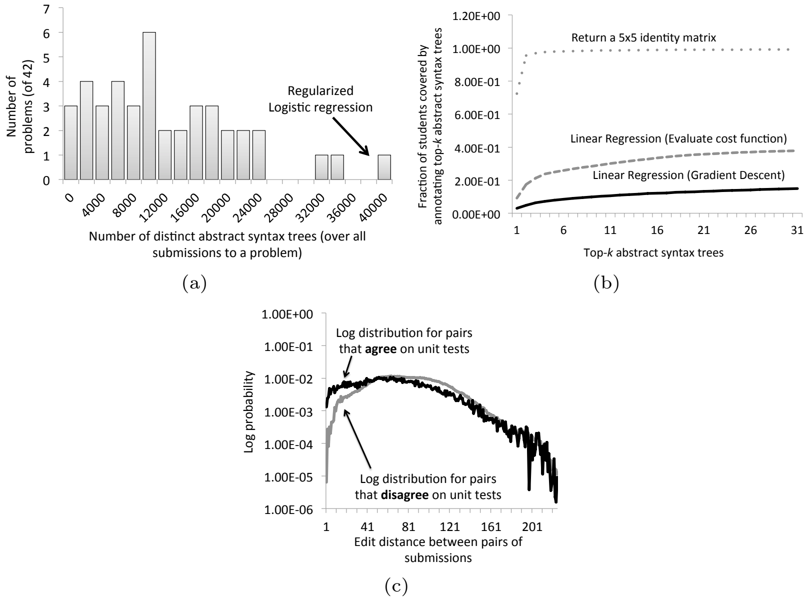
Thus, while output-based feedback provides us with a useful start, the vanilla approach has some limitations. More importantly however, output based feedback can often be too much of an oversimplification. For example, output-based feedback does not capture the fact that multiple output vectors can result from similar misconceptions and conversely that different misconceptions can result in the same unit test outputs. Success of output-based feedback depends greatly on a well designed battery of unit tests. Moreover, coding style which is a critical component of programming cannot be captured at all by unit test based approaches to providing feedback. In the next sections, we discuss a deeper analysis which delves further into program structure and is capable of distinguishing the more stylistic elements of a submission.

4 Syntactic variability of code submissions

In addition to providing feedback on the functional output of a student’s program, we also investigate our ability to give feedback on programming style. The syntax of code submission in its raw form is a string of characters. While this representation is compact, it does not emphasize the meaning of the code. To more accurately capture the structure of a programming assignment, we compare the corresponding Abstract Syntax Tree (AST) representation.

This task is far more difficult due to the open ended nature of programming assignments which allows for a large space of programs. There were over half a million unique ASTs in our dataset. Figure 5(b) shows that homework assignments had substantially higher syntactic variability than functional variability. Even if a human labeled the thirty most common syntax trees for the Gradient Descent part of the Linear Regression homework, the teacher annotations would cover under 16% of the students. However, syntactic similarity goes beyond binary labels of “same” or “different”. Instead, by calculating the *tree edit distance* between two ASTs we can measure the degree to which two code submissions are similar. Though it is computationally expensive to calculate the similarity between all pairs of solutions in a massive class, the task is feasible given the dynamic programming edit distance algorithm presented by Shasha et al [5] . While the algorithm is quartic in the worst case, it is quadratic in practice for student submission. By exploiting the [5] algorithm and using a computing cluster, we are able to match submissions at MOOC scales.

By examining the network of solutions within a cutoff edit distance of 5, we observe a smaller, more manageable number of common solutions. Figure 1 visualizes this network or landscape of solutions for the linear regression (with gradient descent) problem, with node representing a distinct AST and node sizes scaling logarithmically with respect to the number of submissions owned by that AST. By organizing the space of solutions via this network, we are able to see clusters of submissions that are syntactically similar, and feedback for one AST could potentially be propagated to other ASTs within the same cluster.



(a)

(b)

(c)

Fig. 5. (a) Histogram of the number of distinct abstract syntax trees (ASTs) submitted to each problem.; (b) Number of students covered by the 30 most common ASTs for several representative problems; (c) (Log) distribution over distances between pairs of submissions for pairs who agree on unit test outputs, and pairs who disagree. For very small edit distances (<10 edits), we see that the corresponding submissions are typically also functionally similar (i.e., agree on unit test outputs).

Figure 1 also encodes the unit test outputs for each node using colors to distinguish between distinct unit test outcomes.[2] Note that visually, submissions belonging to the same cluster typically also behave similarly in a functional sense, but not always. We quantify this interaction between functional and syntactic similarity in Figure 5(c) which visualizes (log) distributions over edit distances between pairs of submissions who *agree* on unit test outcomes and pairs of submissions who *disagree* on unit test outcomes. Figure 5(c) shows that when two ASTs are within approximately 10 edits from each other, there is a high probability that they are also functionally similar. Beyond this point, the two distributions are not significantly different, bearing witness to the fact that programs that behave similarly can be implemented in significantly different ways.

5 Discussion and ongoing work

The feedback algorithm outlined in this paper lightly touches on the potential for finding patterns that can be utilized to force multiply teacher feedback. One

[2] Edge colors are set to be the average color of the two endpoints.

clear path forward is to propagate feedback, not just for entire programs, but also for program parts. If two programs are different yet share a substantial portion in common we should be able to leverage that partial similarity.

Though we focused our research on creating an algorithm to semi-automate teacher feedback in a MOOC environment, learning the underlying organization of assignment solutions for an entire class has benefits that go beyond those initial objectives. Knowing the space of solutions and how students are distributed over that space is valuable to teaching staff who could benefit from a more nuanced understanding of the state of their class. Moreover, though this study is framed in the context of MOOCs, the ability to find patterns in student submissions should be applicable to any class with a large enough corpus of student solutions, for example, brick and mortar classes which give the same homeworks over multiple offerings, or Advanced Placement exams where thousands of students answer the same problem.

References

1. D. Gitchell and N. Tran. Sim: a utility for detecting similarity in computer programs. In *ACM SIGCSE Bulletin*, volume 31, pages 266–270. ACM, 1999.
2. B. Hartmann, D. MacDougall, J. Brandt, and S. R. Klemmer. What would other programmers do: suggesting solutions to error messages. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1019–1028. ACM, 2010.
3. C. Piech, M. Sahami, D. Koller, S. Cooper, and P. Blikstein. Modeling how students learn to program. In *Proceedings of the 43rd ACM technical symposium on Computer Science Education*, pages 153–160. ACM, 2012.
4. P. M. Sadler and E. Good. The impact of self-and peer-grading on student learning. *Educational assessment*, 11(1):1–31, 2006.
5. D. Shasha, J. T.-L. Wang, K. Zhang, and F. Y. Shih. Exact and approximate algorithms for unordered tree matching. *IEEE Transactions on Systems, Man, and Cybernetics*, 24(4):668–678, 1994.

Revisiting and Extending the Item Difficulty Effect Model

Sarah Schultz and Trenton Tabor
Worcester Polytechnic Institute
100 Institute Rd, Worcester, MA
{seschultz, tstabor}@wpi.edu

Abstract: Data collected by learning environments and online courses contains many potentially useful features, but traditionally many of these are ignored when modeling students. One feature that could use further examination is item difficulty. In their KT-IDEM model, Pardos and Heffernan proposed the use of question templates to differentiate guess and slip rates in knowledge tracing based on the difficulty of the template- here, we examine extensions and variations of that model. We propose two new models that differentiate based on template- one in which the learn rate is differentiated and another in which learn, guess, and slip parameters all depend on template. We compare these two new models to knowledge tracing and KT-IDEM. We also propose a generalization of IDEM in which, rather than individual templates, we differentiate between multiple choice and short answer questions and compare this model to traditional knowledge tracing and IDEM. We test these models using data from ASSISTments, an open online learning environment used in many middle and high school classrooms throughout the United States.

Keywords: Knowledge tracing, student modeling, item difficulty, Bayesian networks, educational data mining

1. Introduction

Traditionally, knowledge tracing (KT), does not take into account much of the data collected by tutoring system. Some work has been done on leveraging hint and attempt counts in KT [8], [9], and in individualizing based on student [6], but one area that merits more exploration is the use of item difficulty to more accurately model students. Pardos and Heffernan proposed a model to do just that [5], but explored only one such possible model. We created two variations on this model and a generalization of it in order to determine which of these models is the best predictor of student knowledge. Our goal is to discover how item difficulty really affects students' knowledge and performance.

2. Models

2.1 Knowledge Tracing

In classic knowledge tracing [1], the goal is to predict whether a student will answer the next question correctly based upon the current estimate of their knowledge. In the Bayesian network, the responses are the observed nodes, and the student's knowledge at each time-step are the latent nodes. Using Expectation Maximization (EM) or another algorithm, we learn values for the probability of initial knowledge, P(L_0); the probability of learning the skill from one time step to the next, P(T); the probability of guessing correctly when the skill is in the unlearned state, P(G); and the probability of slipping, or answering incorrectly when the skill is in the learned state, P(S) (Figure 1).

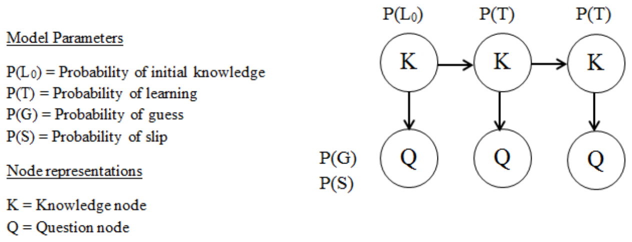

Model Parameters
P(L₀) = Probability of initial knowledge
P(T) = Probability of learning
P(G) = Probability of guess
P(S) = Probability of slip
Node representations
K = Knowledge node
Q = Question node
P(L₀)
P(T)
P(T)
K
K
K
P(G)
P(S)
Q
Q
Q

Fig. 1- Standard Knowledge Tracing

2.2 KT-IDEM

In 2011, Pardos and Heffernan proposed the Knowledge Tracing- Item Difficulty Effect Model (KT-IDEM), which adds difficulty to the traditional KT model by adding an item difficulty node affecting the question node. This model learns a separate guess and slip rate for each item, and therefore has N*2+2 parameters, where N is the number of unique items, in comparison to KT's four [5]. Figure 2 illustrates the KT-IDEM model.

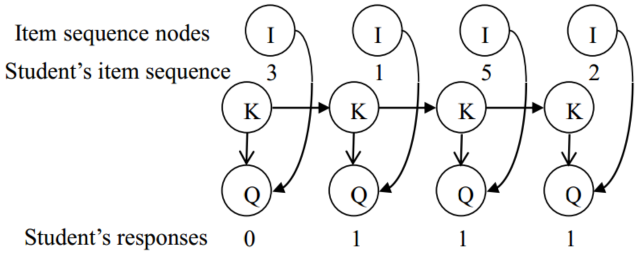


Fig. 2- Knowledge Tracing- Item Difficulty Effect Model

2.3 Extensions to IDEM

We believe that question difficulty not only affects performance, but will also have an effect on learning. By answering questions of different difficulties and receiving feedback on whether or not the answer is correct, students could learn differing amounts. We therefore propose two new variations on KT-IDEM. The first individualizes learn rates by item difficulty, but keeps guess and slip consistent. The second individualizes learn, guess, and slip rates based on item difficulty. In a ten item dataset, KT would have four parameters, KT-IDEM would have 22, the first of our models, Item Difficulty Effect on Learning (IDEL), would have 12, and the second, Item Difficulty Effect All (IDEA), would have 32. It is possible that certain datasets will be over-parameterized in some of these models if there are not enough data points per item, but as Pardos and Heffernan pointed out in their original KT-IDEM paper, “there has been a trend of evidence that suggests models that have equal or even more parameters than data points can still be effective” [5]. These models are illustrated below (Figures 3 and 4).

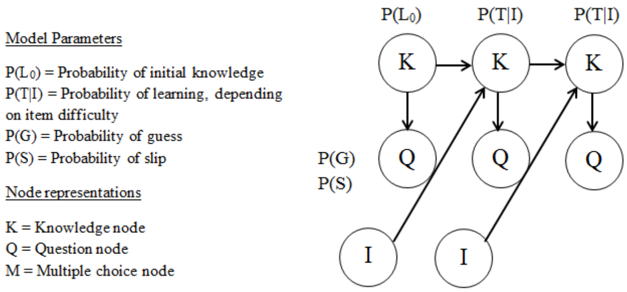

Model Parameters
P(L0) = Probability of initial knowledge
P(T|I) = Probability of learning, depending on item difficulty
P(G) = Probability of guess
P(S) = Probability of slip
Node representations
K = Knowledge node
Q = Question node
M = Multiple choice node
P(L0)
P(T|I)
P(T|I)
K
K
K
P(G)
P(S)
Q
Q
Q
I
I

Fig. 3- Item Difficulty Effect on Learning

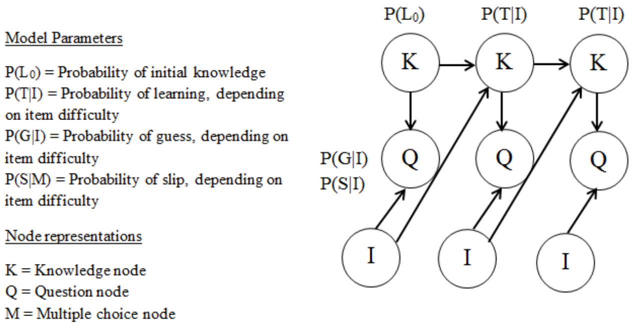

Model Parameters
P(L0) = Probability of initial knowledge
P(T|I) = Probability of learning, depending on item difficulty
P(G|I) = Probability of guess, depending on item difficulty
P(S|M) = Probability of slip, depending on item difficulty
Node representations
K = Knowledge node
Q = Question node
M = Multiple choice node
P(L0)
P(T|I)
P(T|I)
K
K
K
P(G|I)
P(S|I)
Q
Q
Q
I
I
I

Fig. 4- Item Difficulty Effect All

2.4 MC

The final model we implemented is a generalization of KT-IDEM, which adds a multiple choice node to KT at each time step, indicating whether the particular question is multiple choice or not, rather than an item difficulty node. We now learn two different guess and slip rates, one each for multiple choice questions and for non-multiple choice questions. As is standard in KT and all other models explored in this paper, we assume that students do not forget. The multiple choice model (MCKT) is illustrated in Figure 5.

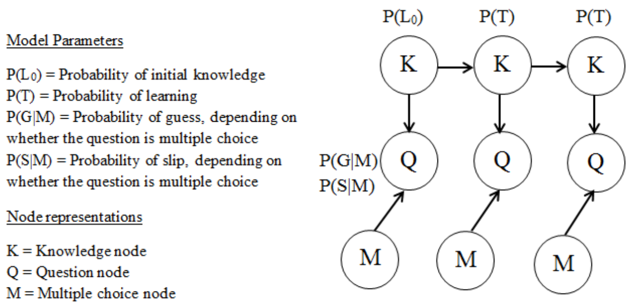

Model Parameters
P(L0) = Probability of initial knowledge
P(T) = Probability of learning
P(G|M) = Probability of guess, depending on whether the question is multiple choice
P(S|M) = Probability of slip, depending on whether the question is multiple choice
Node representations
K = Knowledge node
Q = Question node
M = Multiple choice node
P(L0)
P(T)
P(T)
K
K
K
P(G|M)
P(S|M)
Q
Q
Q
M
M
M

Fig. 5- Multiple Choice Model

We expected that the guess rate for multiple choice questions would be higher than the guess rate for non-multiple choice questions, since there are a finite number of options presented as opposed to an open response where it is possible to enter almost anything. We also expected that the slip rate would be lower for multiple choice questions, as recognizing the correct answer is generally easier than recalling it [3].

3. Dataset

3.1 The ASSISTments Tutoring System

The data used in this work is from ASSISTments, a freely available online mathematics tutoring system for grades 4 to 10 [2]. This system is used in classrooms across the country, and while it is not currently in itself a course, it is certainly an open, large-scale online learning tool.

In ASSISTments, multiple items can be built using the same template, where the only difference is the actual numbers in the problem. We consider problems generated from the same template to be the same item when working with the models that consider item difficulty.

We used six skills from the dataset, all of which came from skill builder data. In ASSISTments, skill builders are sessions where a student practices a certain skill until s/he gets three questions correct in a row, at which point it is considered to be learned. Within each skill, there are different sequences of templates that a student could encounter. In order to be sure that all students in our dataset were seeing the same templates, we used one sequence from each skill, except for Ordering Integers, from which we sampled two sequences separately. Table 1 shows information about the sequences we used in our experiments.

Table 1- Sequences used to test the models

Skill Name	**Percent correct**	**Number of Templates**	**Percent Multiple Choice**
Pythagorean Theorem	34	8	70
Ordering Integers (1)	88	3	34
Ordering Integers (2)	84	3	65
Square Root	89	2	38
Ordering Positive Decimals	74	3	100
Percent	33	13	67
Pattern Finding	48	5	45

4. Methods

Using Kevin Murphy's Bayes Net Toolbox for Matlab [4], we built each of our proposed models. We performed a 5-fold cross-validation on each of the seven sequences from the ASSISTments dataset using all five models, where four folds were used for training and the fifth for testing. The data was partitioned into folds randomly such that each student within a skill was in only one fold and the same folds were used for every model to guarantee a fair comparison. To avoid over-fitting the models to any student who practiced a skill a large number of times, only the first five opportunities of the skill for each student were used. We used expectation maximization to learn the parameters for each of our models.

5. Results

In order to compare models, we calculated mean absolute error (MAE), root mean square error (RMSE), and area under the curve (AUC) of each model's predictions compared to the actual data. We performed a paired t-test of each of these measures using the runs from each fold and found that RMSE was the most consistently reliable measure, so we use that to determine which model is best. Table 2 shows an example of all metrics, obtained from the skill "Percent," which has 13 templates. From this data, it appears that KT has the worst MAE and AUC of all the models, but KT-IDEL has a worse RMSE.

Table 2- Results for “Percent”

	Knowledge Tracing	KT-IDEM	KT-IDEL	KT-IDEA	MCKT
MAE	0.433231	0.350409	0.433039	0.352525	0.352107
AUC	0.531074	0.762205	0.56607	0.706951	0.754057
RMSE	0.472552	0.449915	0.481702	0.441461	0.462738

Comparing the template-based models to KT, we found that for this skill, the MAE was reliably better for KT-IDEM than KT or KT-IDEL and the AUC of KT-IDEM was reliably better than KT and both other template models. On the other hand, KT-IDEA had a reliably better RMSE than KT-IDEM for this skill.

Taking the data from all seven sequences, we unfortunately did not find a conclusive answer to the question of which template-based model performs best. For the skill “Pattern Finding,” we found that KT-IDEM did best in all three measures, whereas for the first sequence of “Ordering Integers,” KT-IDEL outperformed the other two template-based models, but was not significantly different from KT. (A few additional results tables can be found in the appendix of this paper.)

Our next question, was whether the multiple choice model would perform better than KT or KT-IDEM. While theoretically, the multiple choice model should be the same as KT when all problems are of one type, when we ran the models over a sequence that was all multiple choice, the models learned different parameters. This is probably because the multiple choice nodes must always have two values in their CPT tables. We therefore exclude this sequence from analysis of the multiple choice model. On the other hand, we did test a sequence that was all one template, and all template models behaved the same, since the number of values in the template nodes’ CPT tables is the same as the number of templates. Out of the six remaining sequences in which we can compare MCKT, each with three metrics, for a total of 18 comparisons, we found that MCKT was reliably better than KT six times, and reliably better than KT-IDEM four times. Out of these, only two instances showed MCKT better than both of the other models. Out of the remaining nine comparisons, four showed that MCKT was better than the others, but not reliably so, in one case KT-IDEM outperforms MCKT, which is marginally better than KT, and in six cases the both of the other models performed better than MCKT. Since MCKT is at least marginally better than KT a majority of the time, and significantly better in 6 out of 18 cases, it looks like it could be a promising model, although more research is needed.

6. Contributions and Future Work

In this work, we proposed three new models; IDEL, IDEA, and MCKT. We compared these models to traditional KT and to KT-IDEM and found that different models worked best for different sequences. Our findings are not in agreement with [5], which states that IDEM works better than KT in ASSISTments skill builder data, and our observations also seem to indicate that other item difficulty models could work better than KT-IDEM. The interesting contribution here is that this means question difficulty does, in fact, appear to affect learning, possibly more than performance on the current item.

We used only six sequences (and had to exclude one from analysis), all from the same system, in this preliminary look at these models and would like to, in the future, try using more sequences and data from other tutors to see be sure that findings hold true in other scenarios and are not useful only in ASSISTments. Although, even if the latter is the case, having a better student modeling technique for this system would be very useful in developing ways to make it better.

One clear next step is to implement the same extensions made to the IDEM model to the multiple choice model in order to determine how the different types of questions- multiple choice and short answer- effect student knowledge and performance.

Acknowledgements

This work was supported by the WPI Computer Science department and Robotics Engineering program. We would like to thank all of the people associated with creating ASSISTments, as well as the schools and teachers using ASSISTments. We thank Professor Neil Heffernan for helping us to conceptualize this work and for his ideas and Yutao Wang for the code base for knowledge tracing. We would also like to thank Zach Pardos for his advice.

Appendix

Table 3- Results for “Pythagorean Theorem”

	KT	KT-IDEM	KT-IDEL	KT-IDEA	MCKT
MAE	0.480245	0.448852	0.478075	0.431558	0.472431
AUC	0.610767	0.630755	0.661355	0.671785	0.587751
RMSE	0.491635	0.517432	0.487239	0.511354	0.530694

Table 4- Results for “Ordering Positive Decimals” (MCKT excluded)

	KT	KT-IDEM	KT-IDEL	KT-IDEA
MAE	0.352754	0.434477	0.362968	0.451735
AUC	0.58984	0.549476	0.61913	0.577328
RMSE	0.422419	0.474215	0.418596	0.492843

Table 5- Results for “Ordering Positive Integers (1)”

	KT	KT-IDEM	KT-IDEL	KT-IDEA	MCKT
MAE	0.223823	0.268527	0.223668	0.270949	0.2948
AUC	0.545965	0.351229	0.560837	0.36537	0.38731
RMSE	0.333427	0.365692	0.335122	0.394251	0.3903

References

1. Corbett, A.T. and Anderson, J.R., 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User Adapted Interaction, 4(4), pp.253–278.
2. Heffernan, N.T. ASSISTments. http://teacherwiki.assistment.org/wiki/About www.assistments.org
3. Moreno, R., 2010. Education Psychology, John Wiley & Sons, Inc.
4. Murphy, K. 2007. Bayes Net Toolbox for Matlab.
5. Pardos, Z.A. and Heffernan, N.T., 2011. KT-IDEM: Introducing Item Difficulty to the Knowledge Tracing Model. In J. A. Konstan, R. Conejo, J. L. Marzo, & N. Oliver, eds. User Modeling, Adaption and Personalization. Springer Berlin Heidelberg, pp. 243–254.
6. Pardos, Z.A. and Heffernan, N.T., 2010. Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing. User Modeling Adaptation and Personalization, In press, pp.255–266.
7. Qiu, Y., Qi, Y., Lu, H., Pardos, Z. and Heffernan, N., 2011. Does time matter modeling the effect of time in Bayesian knowledge tracing. In M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, & J. Stamper, eds. Proceedings of the 4th International Conference on Educational Data Mining. pp. 139–148.
8. Wang, Y. and Heffernan, N.T., 2010. Leveraging First Response Time into the Knowledge Tracing Model.
9. Wang, Y. and Heffernan, N.T., 2011. The “Assistance” model: Leveraging how many hints and attempts a student needs. In 24th International Florida Artificial Intelligence Research Society FLAIRS 24 May 18 2011 May 20 2011. AAAI Press, pp. 549–554.

Using Argument Diagramming to Improve Peer Grading of Writing Assignments

Mohammad H. Falakmasir [1], Kevin D. Ashley
Christian D. Schunn

Intelligent Systems Program, Learning Research and Development Center,
University of Pittsburgh
{mhf11, ashley, schunn}@pitt.edu

Abstract. One of the major components of MOOCs is the weekly assignment. Most of the assignments are multiple choice, short answer or programming assignments and can be graded automatically by the system. Since assignments that include argumentation or scientific writing are difficult to grade automatically, MOOCs often use a crowd-sourced evaluation of the writing assignments in the form of peer grading. Studies show that this peer-grading scheme faces some reliability issues due to widespread variation in the course participants' motivation and preparation. In this paper we present a process of computer-supported argumentation diagramming and essay writing that facilitates the peer grading of the writing assignments. The process has not been implemented in a MOOC context but all the supporting tools are web-based and can be easily applied to MOOC settings.

Keywords: Computer Supported Argumentation, Argument Diagramming, Peer Review and Grading

1 Introduction

MOOCs in general and Coursera, in particular, started with courses in the area of Computer Science. These courses offered a variety of homework including multiple choice, short answer, and programming assignments that can be graded automatically by the system. However, recently, many MOOCs have started offering courses in social sciences, humanities, and law subjects whose assignments naturally involve more writing and argumentation. Automatic grading of those kinds of assignments is more challenging given the current state of natural language processing technologies. Coursera and most of the other current systems use a peer-grading mechanism in order to address this issue. However, because of the open access nature of the MOOCs, a massive number of people with different educational backgrounds and language skills from all around the world participate in these courses and this heterogeneity in prior preparation negatively affects the validity and reliability of

[1] Corresponding Author

peer-grades. Researchers have investigated this issue (Duneier, 2012) and some steps have been taken to address it. Coursera, for example, flags students who give inaccurate grades and assigns their assessments less weight, but this method does not directly address the diversity of knowledge and writing skills among the students. In this paper, we recommend an approach to this issue that combines computer-supported argument diagramming and writing with scaffolded peer-review and grading. With support of the National Science Foundation,[2] our ArgumentPeer process combines two web-based tools (SWoRD and LASAD) that have been used in several university settings and courses, and applies them to support argumentation and writing assignments in science and law. The process enables the instructional team to carefully define and monitor the writing assignment and revision procedure and involves several machine learning and natural language processing components.

2 Background

Writing and argumentation are fundamental skills that support learning in many topics. Being able to understand the relationships among abstract ideas, to apply them in solving concrete problems, and to articulate the implications of different findings for studies and theories are essential for students in all areas of science, engineering, and social studies. However, inculcating these skills, or compensating for the lack of them, is especially difficult in MOOC setting where students have such diverse preparations and motivations.

Our approach to tackle this problem involves breaking down the process of writing into multiple measurable steps and guiding the student through the steps with careful support and feedback. The first step of the process, computer-supported argument planning, engages the students with a graphical representation for constructing arguments and provides them with feedback and intelligent support. We use LASAD[3] as our argument-diagramming tool (cf. Scheuer et al., 2010). LASAD is a web-based argumentation support system to help students learn argumentation in different domains. It supports flexible argument diagramming by enabling instructors to define a pre-structured palette of argumentation elements (Argument Ontology) along with a set of help system rules in order to give instant feedback to students while working on their diagrams.

The massive number of students in MOOC settings makes it impossible for the instructional team to provide reflective feedback on each individual student's argument. We handle this issue with computer-supported peer-review and grading using SWoRD[4] (Cho & Schunn, 2007). In general, peer review is consistent with learning theories that promote active learning. Furthermore, the peer-review of writing has some learning benefits for the reviewer, especially when the students provide constructive feedback (Wooley, Was, Schunn, & Dalton, 2008), and put effort into the process (Cho & Schunn, 2010). Moreover, studies have shown that

[2] Award No.1122504 DIP: Teaching Writing and Argumentation with AI-Supported Diagramming and Peer Review

[3] http://cscwlab.in.tu-clausthal.de/lasad/

[4] https://sites.google.com/site/swordlrdc/

feedback from a group of peers can be at least as useful as that of teachers (Cho & Schunn, 2007), especially when good rubrics and incentives for reviewing are included. Most relevant here, studies have shown that even students with lower levels of knowledge in the topic can provide feedback that is useful to the ones with higher levels (Patchan & Schunn, 2010; Patchan, 2011).

3 The Process

The ArgumentPeer process includes two main phases: 1) Argument Planning, and 2) Argument Writing. Fig. 1 shows an overview of the process and its underlying components and steps.

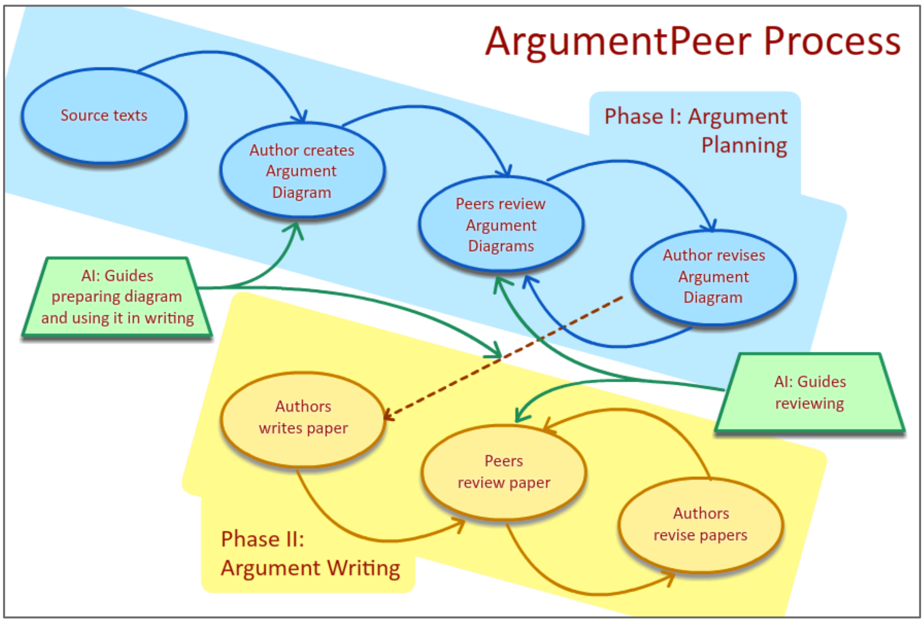
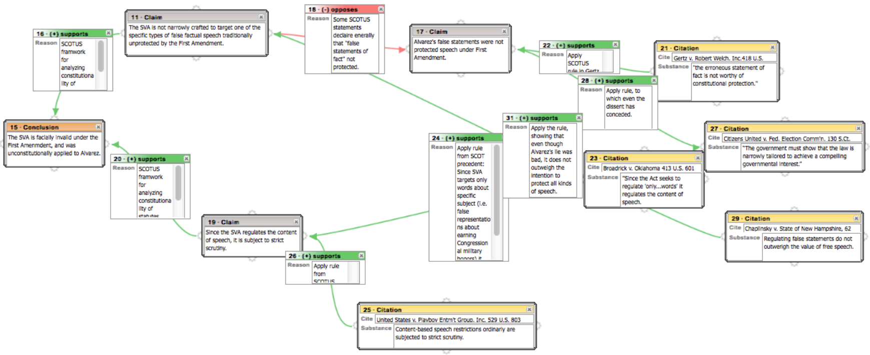

ArgumentPeer Process
Source texts
Author creates Argument Diagram
Peers review Argument Diagrams
Author revises Argument Diagram
Phase I: Argument Planning
AI: Guides preparing diagram and using it in writing
AI: Guides reviewing
Authors writes paper
Peers review paper
Authors revise papers
Phase II: Argument Writing

Fig. 1: ArgumentPeer Process

3.1 Phase I: Argument Diagramming

This phase includes studying the assigned resources and creating the argument diagram. As an example, students in a legal writing course used LASAD in order to prepare textual brief on appeal to the U.S. Supreme Court in the case of *United States v. Alvarez* (Lynch et al., 2012). The system had been introduced to them in a 45-minutes lecture session (that could easily be made a video) and students were directed toward a recommended stepwise format for written legal argumentation as set forth in a noted authority (Neumann 2005). Figure 2 shows an example diagram in this study.



16 · (+) supports
Reason: SCOTUS framwork for analyzing constitutionality of

11 · Claim
The SVA is not narrowly crafted to target one of the specific types of false factual speech traditionally unprotected by the First Amendment.

18 · (-) opposes
Reason: Some SCOTUS statements declaire enerally that "false statements of fact" not protected.

17 · Claim
Alvarez's false statements were not protected speech under First Amendment.

22 · (+) supports
Reason: Apply SCOTUS rule in Gertz

21 · Citation
Cite: Gertz v. Robert Welch. Inc.418 U.S.
Substance: "the erroneous statement of fact is not worthy of constitutional protection."

28 · (+) supports
Reason: Apply rule, to which even the dissent has conceded.

15 · Conclusion
The SVA is facially invalid under the First Amenmdent, and was unconstitutionally applied to Alvarez.

31 · (+) supports
Reason: Apply the rule, showing that even though Alvarez's lie was bad, it does not outweigh the intention to protect all kinds of speech.

24 · (+) supports
Reason: Apply rule from SCOT precedent: Since SVA targets only words about specific subject (i.e. false representations about earning Congressional military honors) it

27 · Citation
Cite: Citizens United v. Fed. Election Comm'n. 130 S.Ct.
Substance: "The government must show that the law is narrowly tailored to achieve a compelling governmental interest."

20 · (+) supports
Reason: SCOTUS framwork for analyzing constitutionality of statutes

23 · Citation
Cite: Broadrick v. Oklahoma 413 U.S. 601
Substance: "Since the Act seeks to regulate 'only...words' it regulates the content of speech.

19 · Claim
Since the SVA regulates the content of speech, it is subject to strict scrutiny.

29 · Citation
Cite: Chaplinsky v. State of New Hampshire, 62
Substance: Regulating false statements do not outwerigh the value of free speech.

26 · (+) supports
Reason: Apply rule from SCOTUS

25 · Citation
Cite: United States v. Playboy Entm't Group. Inc. 529 U.S. 803
Substance: Content-based speech restrictions ordinarly are subjected to strict scrutiny.

Fig. 2: Example Argument Diagram in Legal Writing Course

The instructional team tailored the argument ontology to support the recommended argumentation format; the nodes were basically legal "claim" and "conclusion" nodes that are connected together via "supporting" and "opposing" links providing reasons for and against. The development of a suitable ontology is a critical aspect in the design of an argumentation system and might involve iterative refinement based on observed problems and weaknesses (Buckingham et al., 2002). Specifically, ontologies affect the style of argumentation (Suthers, et al., 2001) and the level of details expected for students to provide. LASAD provides an authoring tool that enables the instructional team to carefully design the argumentation ontology.

After creating the argument diagrams, the students submit their diagrams to the SWoRD system for revision. As noted, SWoRD lets instructors provide a detailed rubric with which peers should assess the diagram. Moreover, it has a natural language processing (NLP) component that pushes reviewers to provide useful feedback that is not ambiguous or vague (more details in section 3.3). After receiving the reviews, the author will revise his/her argument diagram and get ready to write the first draft of the writing assignment in phase 2. To support this transition to a written argument, a system component creates a textual outline based on a depth-first traversal of the argumentation diagram and informed by the argument ontology. In this way, students are encouraged to create a well-annotated argumentation diagram because the diagram text is easily transferred directly to the written draft.

3.2 Phase II: Writing

In this phase, students write their first drafts using the outlines generated from the argument diagrams and submit them to SWoRD. After that, the system automatically assigns the draft to *n* reviewers based on the instructors' policy. The instructor can also assign the individual or groups of peers for the revision using various methods. For example, in the Legal Writing course, the instructor divided the students into two groups, one, writing for the majority and the other writing for the dissenting judge in the 9th Circuit U.S. Court of Appeals and assigned the peers in a way such that there is at least one peer from the other group among the reviewers.

In the next step, the instructor carefully designs the paper reviewing criteria (rubric) for the peers and then starts the reviewing process. The key feature of SWoRD is the ease with which instructors can define rubrics to guide peer reviewers in rating and commenting upon authors' work. The instructor-provided rubrics, which may include both general domain writing and content-specific criteria (Goldin & Ashley, 2012), should help to focus peer feedback and compensate for the wide diversity of peer-reviewers' preparation and motivation.

Reviewers, then, download the paper and evaluate them based on the defined rubric and submit their reviews and ratings to SWoRD. Again, the NLP component of the system, checks the reviews for usefulness and then the system deliverers the reviews back to the author. SWoRD automatically determines the accuracy of each reviewer's numerical ratings using a measure of consistency applied across all of the writing dimensions (Cho & Schunn, 2007). Finally, the author submits the second draft to the system and the final draft can either be grader by peers or the instructional team, although of course in a MOOC context peers would grade it again.

3.3 AI Guides Student Authors and Reviewers in Both Phases

As mentioned, the LASAD Authoring tool and its flexible ontology structure enable instructors to specify the level of detail on which they want the students to focus. Instructors can also use the Feedback Authoring tool to define help system rules that guide the students through the argumentation diagramming process. The instant feedback component of LASAD is an expert system that uses logical rules to analyze students' developing argument diagrams and to provide feedback on making more complete and correct diagrams. The hints can be as simple as telling the student to fill in a text field for an element, or as complex as telling the student to include opposing, as well as supporting, citations for a finding. Using this in-depth intervention, instructors can focus students on their intended pedagogical goals. For example, in the legal writing course, a help system rule asks students to include at least one opposing "citation" in their diagrams to anticipate possible important counterarguments that a court would expect an advocate to have addressed in his or her brief.

The NLP component of SWoRD helps the students improve their reviews by detecting the presence or absence of key feedback features like the location of the problem and the presence of an explicit solution. This feature has been implemented for review comments on both argumentation diagrams and the written drafts. The details of the computational linguistic algorithm that detects the feedback issues are described in (Xiong et al., 2012; Nguyen & Litman, in press). The interface provides reviewers with advice like: "Say where this issue happened." "Make sure that for every comment below, you explain where in the paper it applies." In addition, it provides examples of the kind of good feedback likely to result in an effective revision: "For example, on page [x] paragraph [y], …. Suggest how to fix this problem." "For example, when you talk about [x], you can go into more detail using quotes from the reading resource [y]." The system tries to be as helpful as possible, but in order to prevent frustration, it allows the reviewers to ignore the suggestions and submit the review as is. However, SWoRD considers these reviewers as less accurate and gives lower weight to their ratings.

4 Assessment and Grading

After submitting the final draft, the papers are assigned automatically or by the instructors to the same or another group of peers (or members of the instructional team in non-MOOC contexts) for grading. The same rubric can be used for the second round of review but it is also possible to define new criteria particularly for grading purposes.

According to (Cho, Schunn, & Wilson, 2006; Patchan, Charney, & Schunn, 2009) the aggregate ratings of at least 4 peers on a piece of writing in this setting are more highly reliable and just as valid as a single instructor's ratings. However, some studies (e.g., Chang et al., 2011) note that there can be systematic differences between peer and instructor assessment in a web-based portfolio setting. We believe that by breaking down the argument planning and writing process into multiple guided steps, each subject to review according to instructor-designed peer-review criteria, we move toward a more reliable peer-grading scheme that can be especially useful in a MOOC context.

5 Discussion

Grading writing assignments requires considerable effort, especially when the class size increases. Peer-review and grading is one way to deal with this problem but many instructors are hesitant to use it in their classrooms. The main concern is whether the students are actually capable of grading the papers accurately and responsively. Studies have shown that peer rating alone can be reliable and valid in a large-scale classroom under appropriate circumstances and well-chosen review criteria (Cho, Schunn, & Wilson, 2006; Patchan, Charney, & Schunn, 2009). The ArgumentPeer project not only enables the instructor to design the rubric but also makes it salient for the reviewer to see the deep structure of the argumentation by viewing the argumentation diagram. This positive synergy between diagramming and peer-review makes it easier for the reviewer to see the argument structure in the diagram and its reflection in the writing.

Regarding scalability and the possibility of being used in a MOOC setting, both SWoRD and LASAD are web-based projects developed using Java 2 Platform, Enterprise Edition (J2EE) architecture. LASAD uses automated load balancing in order to support a large number of students. The rich graphical interface of LASAD along with flexible structure of the ontologies helps students gain an understanding of the topic of argumentation (Loll, et al., 2010). Moreover, the collaborative nature of LASAD can be used in order to facilitate engagement, particularly in MOOC settings that face the problem of student retention.

SWoRD, which is the main platform for peer-review and grading, has also been successfully used in classrooms with a large number of students (Cho, Schunn, & Wilson, 2006). The basic review structure in SWoRD is quite similar to the journal publication process, which makes it a familiar process among academics. In addition, publicizing students' papers to their peers can make students put more effort into writing by increasing audience awareness (Cohen & Riel, 1989).

6 Conclusion

In this paper, we presented a process of argument diagramming and reciprocal peer-review in order to facilitate the grading of writing assignments. The ArgumentPeer process and its preexisting components, SWoRD and LASAD, have been applied across different university settings in different courses with large numbers of students. We have decomposed writing assignments into separate steps of planning an argument and then writing it, support students in each step with instructor- and AI-guided peer reviewing and grading. The results of our past studies show that high reliability and validity in the peer grading can be achieved with multiple reviewers per paper. The web-based nature of the components of the ArgumentPeer process makes it relatively easy to apply in MOOC settings. We believe that its fine-grained support for authoring and reviewing could help achieve higher levels of reliability and validity in MOOCs despite their massive numbers of highly diverse participants.

References

1. Buckingham Shum, S. J., Uren, V., Li, G., Domingue, J., Motta, E., & Mancini, C. (2002). Designing representational coherence into an infrastructure for collective sense-making. Invited discussion paper presented at the 2nd International Workshop on Infrastructures for Distributed Collective Practices.
2. Chang, C. C., Tseng, K. H., & Lou, S. J. (2011). A comparative analysis of the consistency and difference among teacher-assessment, student self-assessment and peer-assessment in a Web-based portfolio assessment environment for high school students. *Computers and Education*, 58(1), 303-320.
3. Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education*, 48(3), 409–426.
4. Cho, K., & Schunn, C. D. (2010). Developing writing skills through students giving instructional explanations. In M. K. Stein & L. Kucan (Eds.), *Instructional Explanations in the Disciplines: Talk, Texts and Technology*. New York: Springer.
5. Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, 98(4), 891-901.
6. Cohen, M., & Riel, M. (1989). The effect of distant audiences on students' writing. *American Educational Research Journal*, 26, 143–159.
7. Duneier, M. (2012). Teaching to the world from central New Jersey. *Chronicle of Higher Education*, September 3.
8. Goldin, I. M. & Ashley, K. D. (2012) Eliciting Formative Assessment in Peer Review. *Journal of Writing Research* 4(2) pp. 203–237.
9. Loll, F., Scheuer, O., McLaren, B. M. & Pinkwart, N. (2010). Computer-Supported Argumentation Learning: A Survey of Teachers, Researchers, and System Developers. In M. Wolpers, P. A. Kirschner, M. Scheffel, S. Lindstaedt, & V. Dimitrova, Proceedings of the 5th European Conference on Technology Enhanced Learning (EC-TEL 2010), LNCS 6383, pp. 530-535. Springer.
10. Lynch, C., Ashley, K. D., Falakmassir, M. H., Comparing Argument Diagrams, in proceedings of The 25th Annual Conference on Legal Knowledge and Information Systems (JURIX), Amsterdam, Netherlands, December 2012, pp. 81-90.

11. Neumann, R. (2005) *Legal Reasoning and Legal Writing: Structure, Strategy, and Style*. (5th Ed.) Walters Kluwer.
12. Nguyen H., Litman D., (in press). Identifying Localization in Peer Reviews of Argument Diagrams. Accepted in the 16th International Conference on Artificial Intelligence in Education (AIED 2013), Memphis, TN.
13. Patchan, M. M., Charney, D., & Schunn, C. D. (2009). A validation study of students' end comments: Comparing comments by students, a writing instructor, and a content instructor. *Journal of Writing Research*, 1(2), 124-152.
14. Patchan, M. M., & Schunn, C. D. (2010). Impact of Diverse Abilities on Learning to Write through Peer-Review. Paper presented at the 32nd annual meeting of the Cognitive Science Society, Portland, OR.
15. Scheuer, O., Loll, F., Pinkwart, N. and McLaren, B. M. (2010). Computer-supported argumentation: A review of the state-ofthe-art. *International Journal on Computer Supported Collaborative Learning*, 5(1), 43-102. Springer.
16. Suthers, D. D., Connelly, J., Lesgold, A., Paolucci, M., Toth, E. E., Toth, J., & Weiner, A. (2001). Representational and advisory guidance for students learning scientific inquiry. In K. D. Forbus & P. J. Feltovich (Eds.), *Smart machines in education: The coming revolution in educational technology* (pp. 7–35). Menlo Park, CA: AAAI/MIT Press.
17. Wooley, R., Was, C., Schunn, C., & Dalton, D. (2008). The effects of feedback elaboration on the giver of feedback. Paper presented at the 30th Annual Meeting of the Cognitive Science Society.
18. Xiong, W., Litman, D., & Schunn, C. D. (2010). Natural Language Processing techniques for researching and improving peer feedback. *Journal of Writing Research*, 4(2), 155-176.

Improving learning in MOOCs with Cognitive Science

Joseph Jay Williams[1]
[1] University of California at Berkeley
joseph_williams@berkeley.edu

Abstract. MOOCs and other platforms for online education are having a tremendous impact on the learning of tens of thousands of students. They offer a chance to build a set of educational resources from the ground up, at a time when scientists know far more about learning and teaching than at the advent of the current education system. This paper presents practical implications of research from cognitive science, showing empirically supported and actionable strategies any designer or instructor can use to improve students' learning. These all take the form of augmenting online videos and exercises with questions and prompts for students to consider explanations: *before, during,* and *after* learning. This class of instructional strategies provides students with direction while allowing them to take charge of their learning, is technically easy to implement, and is applicable to a wide variety of video and exercise content, that ranges across multiple topics.

Keywords: learning, learning, cognitive science, MOOCs, educational software, online learning, problem based learning, explanation, self-explanation, retrieval practice, interleaving, mixing, spacing

1 Introduction

High quality pedagogy is an essential goal for MOOCs. There are few barriers to students moving between courses, and the expectations are also that online learning platforms will take advantage of their greater freedom to innovate than many education reform movements in traditional schools.

One way to complement the practical experience of quality instructors is to synthesize and apply insights from scientific research. The nature of such work is produce insights that people's direct experience is unlikely to uncover. This paper considers how research from cognitive science can improve learning in MOOCs. The following consider educational implications of cognitive science more generally. [1] is an Institute of Education Sciences practice guide that is short, available online, constructed by an expert panel, and peer-reviewed. Books include [2], which is targeted at university instructors, [3] is for a general audience and K-12 teachers, and [4] focuses on multimedia learning for both K-16 education and corporate training.

This paper follows the approach taken in the reviews above in selecting practical principles from a broad review and synthesis of literature in cognitive science. This includes publications of basic research and controlled laboratory experiments, as well as studies with educational materials and K-12 and university students from K-12 and university students – which are directly relevant to lessons in current MOOCs.

The principles are selected to target key challenges in online learning, like ensuring learners remain engaged and active even without a physical community, promoting deep understanding rather than superficial memory, and supporting students in being strategic and independent learners, even without much direct feedback.

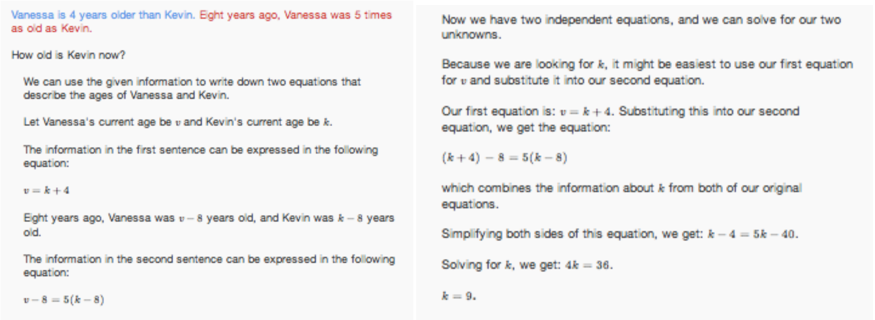
The principles specifically focus on how to appropriately prompt students to answer questions and provide explanations, *before*, *during*, and *after* watching instructional videos or engaging in exercises. It is a common intuition that students learn when they are *given* comprehensive knowledge: MOOCs deliver high-quality online videos with cogent explanations, and include practice exercises like that in Figure 1, accompanied by clear answers and solutions. However, there is substantial evidence that students can learn far more by trying to *answer* questions themselves (than by receiving the answers), or by being pushed to construct explanations (rather than provided with them), which will be discussed in the following sections.

2 Context of application: example video and exercise

Each principle for adding question prompts is targeted at the grain size of an online *module* – a short, self-contained batch of information like a video or exercise.

The principles are abstract in that they can improve learning from a range of online videos and exercises, but to provide concrete and actionable insight they are illustrated through application to specific examples of a video and exercise.

The example video is a three minute Udacity.com video from an introductory statistics course (http://tiny.cc/examplevideo): It explains what the normal distribution is, and how the area under its curve corresponds to the probability of observing certain sampled observations from a population.



Vanessa is 4 years older than Kevin. Eight years ago, Vanessa was 5 times as old as Kevin.

How old is Kevin now?

We can use the given information to write down two equations that describe the ages of Vanessa and Kevin.

Let Vanessa's current age be v and Kevin's current age be k.

The information in the first sentence can be expressed in the following equation:

$v = k + 4$

Eight years ago, Vanessa was $v - 8$ years old, and Kevin was $k - 8$ years old.

The information in the second sentence can be expressed in the following equation:

$v - 8 = 5(k - 8)$

Now we have two independent equations, and we can solve for our two unknowns.

Because we are looking for k, it might be easiest to use our first equation for v and substitute it into our second equation.

Our first equation is: $v = k + 4$. Substituting this into our second equation, we get the equation:

$(k + 4) - 8 = 5(k - 8)$

which combines the information about k from both of our original equations.

Simplifying both sides of this equation, we get: $k - 4 = 5k - 40$.

Solving for k, we get: $4k = 36$.

$k = 9$.

Fig. 1. Example math exercise from Khan Academy: http://tiny.cc/exampleexercise

The example exercise is shown in Figure 1, an algebra word problem from Khan Academy's collection of mathematics exercises at www.khanacademy.org/exercisedashboard. These share a common format. Only the problem statement is shown at first (blue & red text in Figure 1). Students can submit an answer for feedback or request a hint at any point. They only move onto the next problem when they are correct, but each hint request reveals the next step in a worked example solution – which ultimately gives the answer as its final step.

3 Adding questions before, during, and after videos & exercises

Questions or prompts to generate explanations can be added in at least three ways to online modules: *pre*-module (immediately preceding or in the very beginning of a video/exercise, preceding the presentation of content), *intra*-module (popping up in a video or emphasized as an activity by the instructor, embedded into the steps of an exercise), or *post*-module (following the student's engagement with a video/exercise).

3.1 Pre-Module: Framing Questions

Even before learners are presented with information in a video or exercise, prompting them to consider *framing questions* can make them more motivated to learn, as well as help them connect a module's content to their existing knowledge, and understand how they can apply it to future problems.

In contrast to delivering a traditional sequence of *subject-focused* videos & exercises (which touch on a succession of topics students may struggle to relate), *problem-based learning* [5] frames videos & exercises as the knowledge needed to solve particular problems and answer previously articulated questions. For example, a problem-based learning version of an introductory statistics course [6] would precede lessons with a keen emphasis on what problems the lesson would teach students how to solve, rather than a typical focus on the specific facts and concepts in each lesson.

Examples of pre-module framing questions are shown in Table 1.

Table 1. Examples of Framing Questions that could precede videos and exercises.

Udacity video on the normal distribution	Khan Academy algebra math exercise
Before a video, a page with a Framing Question can be presented: "*Explain what you already know about normal distributions*." *"What is a normal distribution useful for?"* Instructors can also introduce a fixed time delay (e.g. 10 seconds), a required text response, or a strong emphasis on a Framing Question at the start of a video.	If you are only told about the relationships between two people's ages, what kind of math is useful for figuring out actual ages? The guiding question to keep in mind for this exercise is: "How can you convert word problems into algebra expressions?"

The motivational benefit is in greater excitement to learn in order to solve a problem, rather than learn to memorize and be tested. The cognitive benefit arises in part by getting learners to activate their existing knowledge, so they connect new information to well-established ideas. Prompts to explain a fact can be largely unsuccessful, but still increase how much is learned once a lesson is presented [7]. [8]

showed that students were mostly unsuccessful when asked to solve a problem related to calculating variability, but that having tried to solve this problem changed *what* they learned from a subsequent lesson. Compared to other students who received alternative instruction without this framing question or problem, these students were better able to apply what they learned in subsequent lessons to new situations.

Developing Framing Questions. To generate framing questions for a particular resource, an instructor can ask:

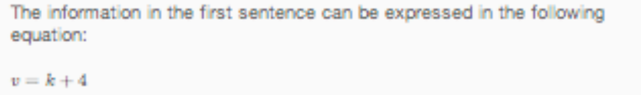
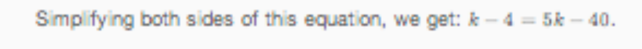
- "What questions should students be able to answer after watching this video, that they can't right now?"
- "What problems do I think they should be able to solve afterwards, that they would have struggled with before?"

3.2 Intra–Module: Reflection Questions

Typically, instruction is seen as *providing* learners with answers or *giving* them explanations. But extensive work in cognitive science, education, and intelligent tutoring has shown that giving learners the right prompts to self-generate explanations can be *more* effective than giving students explanations [9] [10]. This provides empirical insight into how and when "teaching is the best way to learn". Without changing the content of online videos and exercises, MOOCs can improve learning by appropriately embedding questions and prompts for learners to provide explanations.

Videos in MOOCs already have the functionality to pop-up short multiple choice exercises, which could be used to present questions that are more conceptual and that allow open-ended responses. Solutions to exercises can be split up into multiple lines, and have questions and prompts with text boxes to type answers embedded inline. Examples are shown in Table 2.

Table 2. Examples of how Reflection Questions could be embedded in videos and exercises.

Udacity video on normal distribution	Khan Academy algebra math exercise
Explain what the video has talked about so far. (@1:35) What are you thinking about right now? Just say it out loud. (@ 2:15)	 The information in the first sentence can be expressed in the following equation: $v = k + 4$ Do you see why this step makes sense or is justified?  Simplifying both sides of this equation, we get: $k - 4 = 5k - 40$. What step do you think is coming next?

There is substantial evidence that learners' understanding is improved by prompts to explain out loud the meaning of what they are learning or say out loud what they are thinking [9] – although studies typically ensure learners are not confused by the sudden appearance of these prompts. Asking learners to explain *why* particular facts are true or answers are correct has been shown to help them understand key principles and generalizations [11]. [12] shows that anticipating next steps in a solution and making predictions about what will be discussed next leads to a better understanding of how and where to use what they are learning about, and provides implicit feedback as the video continues or solution is revealed.

Developing Reflection Questions. In addition to examining the methods of the studies cited above, the Institute of Education Sciences practice guide [1] provides a reference of effective question stems: E.g., why, why-not, how, what-if, how does X compare to Y, what is the evidence for X?

An instructor can use a list of these stems to generate and insert question or explanation prompts throughout an instructional video or an exercise's solution.

3.3 Post–Module Memory Practice Questions

Questions that target information from a past video or exercise are common in MOOCs, but often do not realize their potential for *Memory Practice*. One reason is that they are often designed to *assess* learning without attention towards *improving* it. [13] shows that simply asking students to recall what they read in a science passage (an open ended prompt that is not common in testing, but encourages Memory Practice) greatly improved memory a week later – outperforming students who read the passage *three more times*, or made elaborate concept maps. Post-module prompts for this paper's current examples might include "Write down the main points from that video." or "Explain the method you used to solve these exercises."

In fact, MOOCs often do include post-module questions designed to help students revisit content – such as review questions or practice exercises. However, these may not successfully produce Memory Practice if they occur so soon after a module that a learner can answer using rote memory. [14] provides an extensive review of how to ensure post-module questions are beneficial, so that Memory Practice helps learners generate the meaningful cues and connections to other concepts that are needed to remember over the long-term.

For example, simply *spacing* practice exercises improves long-term retention (although benefits are deceptively absent in the *short-term*), and learning is even further improved by *interleaving* or *mixing* problems and concepts that students frequently confuse [15]. For example, a typical practice sequence might be 12 problems of type A, then 12 of type B, and 12 of type C. But it can be better for deep, lasting learning to practice [6 A, 4 B, 2 C], [4 A, 6 B, 2 C], and [2 A, 4 B, 6 C]. Often, however, students and instructors may assume that the more challenging learning in the *mixed* condition means that it is a poorer strategy and abandon it – even though it produces larger and lasting benefits *without any increase* in the number of problems [15]. Ironically, the same studies that empirically show the advantages of Memory Practice also find that students expect typical study strategies to help more [13] [14].

Conclusion

This paper considered how to improve learning in MOOCs by adding question & explanation prompts before, during, and after online videos and exercises. This is not to say that MOOCs *never* incorporate questions into instruction as advised – this is unlikely given the diversity of online instruction. Scientific principles for learning can be used to design novel instruction *or* to support *benchmarking* – to identify which of the vast set of instructional strategies are supported by cognitive science. Moreover, consulting and working with cognitive scientists (to embed practical experiments and design measures of learning) allows MOOCs to maximize learning by tailoring general learning principles to specific courses and lessons. Collaborations like these between instructions and scientists can provide the best outcomes for students.

References

1. Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., Metcalfe, J.: Organizing Instruction and Study to Improve Student Learning (NCER 2007-2004). Washington, DC: Institute of Education Sciences, U.S. Department of Education (2007)
2. Ambrose, S. A., Bridges, M. W., DiPietro, M., Lovett, M. C., Norman, M. K.: How learning works: Seven research-based principles for smart teaching. Jossey-Bass (2010)
3. Willingham, D. T.: Why Don't Students Like School. Jossey-Bass (2010)
4. Clark, R. C., & Mayer, R. E.: E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning. Pfeiffer (2004)
5. Hmelo-Silver, C. E.: Problem-based learning: What and how do students learn? Educational Psychology Review, 16(3), 235-266 (2004)
6. Boyle, C. R.: A problem-based learning approach to teaching biostatistics. Journal of Statistics Education, 7(1) (1999)
7. Needham, D. R., Begg, I. M.: Problem-oriented training promotes spontaneous analogical transfer: Memory-oriented training promotes memory for training. Memory & Cognition 19, 543–557 (1991)
8. Schwartz, D.L., Martin, T.: Inventing to Prepare for Future Learning: The Hidden Efficiency of Encouraging Original Student Production in Statistics Instruction. Cognition and Instruction, 22(2), 129-184 (2004)
9. Fonseca, B. Chi, M.T.H.: The self-explanation effect: A constructive learning activity. In: Mayer, R. & Alexander, P. (Eds.), The Handbook of Research on Learning and Instruction (pp. 270-321). New York, USA: Routledge Press (2011)
10. Mazur, E.: Farewell, Lecture? Science (2009)
11. Williams, J. J., Walker, C. M., Lombrozo, T.: Explaining increases belief revision in the face of (many) anomalies. In: N. Miyake, D. Peebles, & R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 1149-1154). Austin, TX: Cognitive Science Society (2012)
12. Renkl, A.: Learning from Worked-Out Examples: A Study on Individual Differences. Cognitive Science, 21(1), 1–29 (1997)
13. Karpicke, J. D., Blunt, J. R.: Retrieval practice produces more learning than elaborative studying with concept mapping. Science, 331(6018), 772-775 (2011)
14. Roediger, H. L., Putnam, A. L., Smith, M. A.: Ten benefits of testing and their applications to educational practice. In: J. Mestre & B. Ross (Eds.), Psychology of learning and motivation: Cognition in education (pp. 1-36). Oxford: Elsevier (2011)
15. Rohrer, D., Pashler, H.: Increasing retention without increasing study time. Current Directions in Psychological Science, 16, 183-186 (2007)

Measurably Increasing Motivation in MOOCs

Joseph Jay Williams[1], Dave Paunesku[2], Benjamin Haley[3], Jascha Sohl-Dickstein[2,4]

[1] U.C. Berkeley, [2]Stanford University, [3]Northwestern University, [4]Khan Academy
United States
joseph_williams@berkeley.edu, paunesku@stanford.edu, benjamin.haley@gmail.com,
jascha@khanacademy.org

Abstract. A key challenge in online learning is keeping students motivated. We report an experiment that added motivational messages to students solving mathematics problems on the KhanAcademy.org platform. By simply adding sentences above the text of a math problem, students attempted (successfully) a greater number of problems, were more likely to acquire exercise proficiencies, and even solved a larger proportion of attempted problems correctly. The key feature for producing these measurably improved outcomes was in using messages that emphasized that intelligence is malleable – e.g., "Remember, the more you practice the smarter you become!". Control conditions that provided neutral science facts or even positive messages – e.g., "This might be a tough problem, but we know you can do it." – were not as effective. There are many pedagogical strategies that instructors of online courses might hypothesize will increase motivation; these findings underscore the value in empirically testing such predictions, using the unique data that is now available in MOOCs.

Keywords: motivation, MOOCs, learning, mindset

Controlled experiments on millions of students to personalize learning

Eliana Feasley[1,2], Chris Klaiber[1], James Irwin[1], Jace Kohlmeier[1], Jascha Sohl-Dickstein[1,3]
[1]Khan Academy, [2]UT Austin, [3]Stanford
United States
{eliana, chris, james, jace, jascha}@khanacademy.org

Abstract. Khan Academy is a personalized learning resource that enables students to watch educational videos and answer questions across a variety of levels of mathematics and other subjects. With over one billion problems solved, Khan Academy has a massive dataset from which to draw evidence and make inferences about student learning behaviors. Our goal is to use this unprecedented quantity of data to learn what content each student will benefit the most from seeing, and to present it to them. Towards this goal, we have run more than one hundred massive controlled experiments, evaluating hypotheses about learning.

We focus here on personalizing the learning experience by using student responses to assessment items to adaptively suggest new content. We discuss the metrics by which we measure student improvement and the tradeoffs that occur when increased exercise difficulty reduces student engagement. We further discuss personalizing content such as exercise or video suggestions, and measuring student responses to such interventions. Leveraging massive data to personalize learning is one of the greatest promises of online education, and this work represents first steps towards fulfilling that promise for millions of users worldwide.

Keywords: personalized learning, data mining, machine learning, massive data, Khan Academy

Analysis of video use in edX courses

Daniel T. Seaton, Albert J. Rodenius, Cody A. Coleman, David E. Pritchard, Isaac Chuang
Massachusetts Institute of Technology
United States
{dseaton, albertr, colemanc, dpritch, ichuang}@mit.edu

Abstract. In Massive Open Online Courses (MOOCs), online videos serve as the equivalent of lectures found in their traditional on-campus courses. Across a number of courses offered by edX in the Fall of 2012, the number of unique videos watched shows bimodal student engagement similar to ``attendance" of large-lecture on-campus courses; only half the participants are watching the majority of course videos. The overall scale of MOOC populations still allows for meaningful measurements of video activity, while also providing a tremendous opportunity to experiment with methods of improving engagement of those participants showing low video use. We present preliminary analyses of the nature of video engagement through both the fraction of videos viewed over the course and the detection of convergent activity (``hot spots") in the collective pause and play interactions within each video. We discuss our results in the context of improving video content, as well as a new video annotation tool being integrated into assessment items.

Keywords: MOOC, Video, Online, Analytics

Exploring Possible Reasons behind Low Student Retention Rates of Massive Online Open Courses: A Comparative Case Study from a Social Cognitive Perspective

Yuan Wang
Columbia University
United States
elle.wang@columbia.edu

Abstract. Massive Open Online Courses (MOOCs) have been widely lauded by the press since its fairly recent inception. Besides its wide popularity among learners worldwide, the majority of MOOCs still present challenges with steep dropout rates in spite of their promising enrollment numbers. While enjoying various benefits MOOCs brings along, learners apparently face new challenges. This paper intends to explore possible reasons behind this phenomenon from a social cognitive perspective by analyzing and comparing the same subject content taught in both the traditional face-to-face setting and on a MOOC-based platform.

Based on past research and theories including both the larger distance learning fields as well as recent MOOC-specific ones, three areas, namely, the lack of self-efficacy, self-regulation, and self-motivators are identified to help present an exploratory framework in interpreting findings of this study. Although far from all encompassing, this exploratory framework attempts to enhance our understanding of distinct challenges MOOC learners as well as MOOC designers face.

Keywords: MOOCs, Distance Learning, Student Retention Rate, Sustainability of Learning.

Using EEG to Improve Massive Open Online Courses Feedback Interaction

Haohan Wang, Yiwei Li, Xiaobo Hu, Yucong Yang, Zhu Meng, Kai-min Chang

Language Technologies Institute
School of Computer Science
Carnegie Mellon University

Abstract. Unlike classroom education, immediate feedback from the student is less accessible in Massive Open Online Courses (MOOC). A new type of sensor for detecting students' mental states is a single-channel EEG headset simple enough to use in MOOC. Using its signal from adults watching MOOC video clips in a pilot study, we trained and tested classifiers to detect when the student is confused while watching the course material. We found weak but above-chance performance for using EEG to distinguish when a student is confused or not. The classifier performed comparably to the human observers who monitored student body language and rated the students' confusion levels. This pilot study shows promise for MOOC-deployable EEG devices being able to capture tutor relevant information.

Keywords: MOOC, EEG, confuse, feedback, machine learning

1 Introduction

In recent years, there is an increasing trend towards the use of Massive Open Online Courses (MOOC), and it is likely to continue [1]. MOOC can serve millions of students at the same time, but it has its own shortcomings. In [2], Thompson studied post-secondary students who had negative attitudes toward correspondence-based distance education programs. The results indicate that lack of immediate feedback and interaction are two problems with long-distance education. Current MOOC can offer interactive forums and feedback quizzes to help improve the communication between students and professors, but the impact of the absence of a classroom is still being hotly debated. As also discussed in [3], indicates the lack of feedback is one of the main problems for student-teacher long distance communication.

There are many gaps between online education and in-class education [4] and we will focus on one of them: detecting students' confusion level. Unlike in-class education, where a teacher can judge if the students understand the materials by verbal inquiries or noticing their body language (e.g., furrowed brow, head scratching, etc.), immediate feedback from the student is less accessible in long distance education. We address this limitation by using electroencephalography (EEG) input from a commercially available device as evidence of students' mental states.

adfa, p. 1, 2011.

The EEG signal is a voltage signal that can be measured on the surface of the scalp, arising from large areas of coordinated neural activity manifested as synchronization (groups of neurons firing at the same rate) [5]. This neural activity varies as a function of development, mental state, and cognitive activity, and the EEG signal can measurably detect such variation. Rhythmic fluctuations in the EEG signal occur within several particular frequency bands, and the relative level of activity within each frequency band has been associated with brain states such as focused attentional processing, engagement, and frustration [6-8], which in turn are important for and predictive of learning [9].

The recent availability of simple, low-cost, portable EEG monitoring devices now makes it feasible to take this technology from the lab into schools. The NeuroSky "MindSet," for example, is an audio headset equipped with a single-channel EEG sensor [10]. It measures the voltage between an electrode that rests on the forehead and electrodes in contact with the ear. Unlike the multi-channel electrode nets worn in labs, the sensor requires no gel or saline for recording and therefore requires much less expertise to position. Even with the limitations of recording from only a single sensor and working with untrained users, a previous study [11] found that the MindSet distinguished two fairly similar mental states (neutral and attentive) with 86% accuracy. MindSet has been used to detect reading difficulty [12] and human emotional responses [13] in the domain of intelligent tutoring systems.

A single-channel EEG device headset currently costs around $99-149 USD, which would be a cost deterant to the free service of MOOC. We suggest that MOOC providers (e.g., Coursera, edX) supply EEG devices to a select group of students. In return, MOOC providers would get feedback on students' EEG brain activity or confusion levels while students watch the course materials. These objective EEG brain activities can be aggregated and augment subjective rating of course materials to provide a simulation of real world classroom responses, such as when a teacher is given feedback from an entire class. Then teachers can improve video clips based on these impressions. Moreover, even though an EEG headset is a luxury device at the moment, the increasing popularity of consumer-friendly EEG devices may one day make it a house-hold accessory like audio headsets, keyboards and mice. Thus, we are hopeful of seeing our proposed solution come to fruition as the market for MOOC grows and the importance of course quality and student feedback increases.

To assess the feasibility of collecting useful information about cognitive processing and mental states using a portable EEG monitoring device, we conducted a pilot study with college students watching MOOC video clips. We wanted to know if EEG data can help distinguish among mental states relevant to confusion. If we can do so by better than chance, then these data may contain relevant information that can be decoded more accurately in the future. Thus, we address two questions:

1. Can EEG detect confusion?
2. Can EEG detect confusion better than human observers?

The rest of this paper is organized as follows. Section 2 describes the experiment design. Section 3 and 4 answers the two research questions, respectively. Finally, Section 5 concludes and suggests future work.

2 Experiment Design

In a pilot study, we collected EEG signal data from college students while they watched MOOC video clips. We extracted online education videos that are assumed not to be confusing for college students, such as videos of introduction of basic algebra or geometry. We also prepare videos that are assumed to confuse a normal college student if a student is not familiar with the video topics like Quantum Mechanics, and Stem Cell Research[1]. We prepared 20 videos, 10 in each category. Each video was about 2 minutes long. We chopped the two-minute clip in the middle of a topic to make the videos more confusing.

We collected data from 10 students. One student was removed because of missing data due to technical difficulties. An experiment with a student consisted of 10 sessions. We randomly picked five videos of each category and randomized the presentation sequence so that the student could not guess the predefined confusion level. In each session, the student was first instructed to relax their mind for 30 seconds. Then, a video clip was shown to the student where he/she was instructed to try to learn as much as possible from the video. After each session, the student rated his/her confusion level on a scale of 1-7, where 1 corresponded to the least confusing and 7 corresponded to the most confusing. Additionally, there were three student observers watching the body-language of the student. Each observer rated the confusion level of the student in each session on a scale of 1-7. The conventional scale of 1-7 was used. Four observers were asked to observe 1-8 students each, so that there was not an effect of observers just studying one student.

The students wore a wireless single-channel MindSet that measured activity over the frontal lobe. The MindSet measures the voltage between an electrode resting on the forehead and two electrodes (one ground and one reference) each in contact with an ear. More precisely, the position on the forehead is Fp_1 (somewhere between left eye brow and the hairline), as defined by the International 10-20 system [14]. We used NeuroSky's API to collect the EEG data.

3 Can EEG detect confusion?

We trained Gaussian Naïve Bayes classifiers to estimate, based on EEG data, the probability that a given session was confusing rather than not confusing. We chose this method (rather than, say, logistic regression) because it is generally best for problems with sparse (and noisy) training data [15].

To characterize the overall values of the EEG signals while the students watch the 2 minute video, we computed their means over the interval. To characterize the temporal profile of the EEG signal, we computed several features, some of them typically used to measure the shape of statistical distributions rather than of time series: minimum, maximum, variance, skewness, and kurtosis. However, due to the small number of data points (100 data points for 10 subjects, each watching 10 videos), inclusion of

[1] http://open.163.com/

those features tends to overfit the training data and results in poor classifier performance. As a result, we used the means as the classifier features for the main analysis. **Table 1** shows the classifier features.

Table 1. Classifier features

Features	**Description**	**Sampling rate**	**Statistic**
Attention	Proprietary measure of mental focus	1 Hz	Mean
Meditation	Proprietary measure of calmness	1 Hz	Mean
Raw	Raw EEG signal	512 Hz	Mean
Delta	1-3 Hz of power spectrum	8 Hz	Mean
Theta	4-7 Hz of power spectrum	8 Hz	Mean
Alpha1	Lower 8-11 Hz of power spectrum	8 Hz	Mean
Alpha 2	Higher 8-11 Hz of power spectrum	8 Hz	Mean
Beta1	Lower 12-29 Hz of power spectrum	8 Hz	Mean
Beta 2	Higher 12-29 Hz of power spectrum	8 Hz	Mean
Gamma1	Lower 30-100 Hz of power spectrum	8 Hz	Mean
Gamma2	Higher 30-100 Hz of power spectrum	8 Hz	Mean

To avoid overfitting, we used cross validation to evaluate classifier performance. We trained student-*specific* classifiers on a single student's data from all but one stimulus block (e.g., one video), tested on the held-out block (e.g., all other videos), performed this procedure for each block, and averaged the results to cross-validate accuracy within reader. We trained *student-independent* classifiers on the data from all but one student, tested on the held-out student, performed this procedure for each student, and averaged the resulting accuracies to cross-validate across students.

We use two ways to label the mental states we wish to predict. One way is the *pre-defined* confusion level according to the experiment design. Another way is the *user-defined* confusion level according to each user's subjective rating.

Detect pre-defined confusion level. We trained and tested classifiers for pre-defined confusion. Student-specific classifiers achieve a classification accuracy of 67% and a kappa statistic of 0.34, whereas student-independent classifiers achieve a classification accuracy of 57% and a kappa statistic of 0.15. Both classifier performances were statistically significant better than a chance level of 0.5 ($p < 0.05$). **Fig. 1a)** plots the classifier accuracy for each student. **Fig. 1a)** shows that both student-specific classifiers and student-independent classifiers performed significantly above chance in 6 out of 9 students.

Detect user-defined confusion level. We also trained and tested classifiers for student-defined confusion. Since students have different sense of confusing, we mapped the seven scale self-rated confusion level into a binary label, with roughly equal number of cases in the two classes. A middle split is accomplished by mapping scores less than or equal to the median to "not confusing" and the scores greater than the median are mapped to "confusing". Furthermore, we used random undersampling of the larger class(es) to balance the classes in the training data. We performed the sampling 10 times to limit the influence of particularly good or bad runs and obtain a stable measure of classifier performance.

Student-specific classifiers achieve a classification accuracy of 57% and a kappa statistic of 0.13, whereas student-independent classifiers achieve a classification accuracy of 51% and a kappa statistic of -0.04. The student-specific classifier performance was statistically significant and better than a chance level of 0.5 ($p < 0.05$), but not the student-independent classifier. **Fig. 1b)** plots the accuracy for each student. **Fig. 1b)** shows that the student-specific classifier performed significantly above chance for 5 out of 9 students and student-independent classifier performed significantly above chance for 2 out of 9 students.

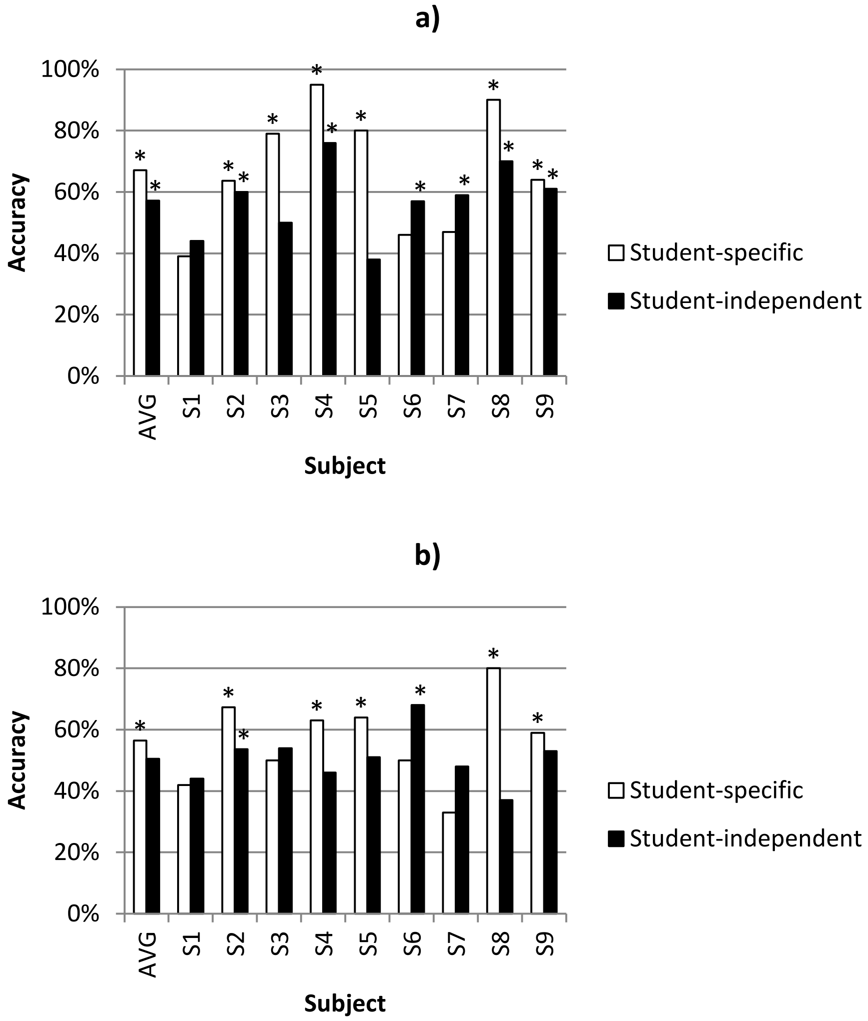


Fig. 1. Detect a) predefined, and b) user-defined confusion level

4 Can EEG detect confusion better than human observers?

To determine if EEG can detect confusion better than human observers of body language, we compared the scores from the observers, the classifier, and the students, with the label of videos. For each student, we used the average scores of the observers as the 'observer rating'. We used the classifier trained in Section 3 to predict predefined confusion level and linearly mapped the classifier's estimate of class probability (0-100%) to a scale of 1-7 and labeled it as the 'classifier rating'.

Fig. 2 shows the scatter plot of a) student vs. observer rating, and b) student vs. classifier rating. The classifier rating had a low, but positive correlation (0.17) with the students' rating, while the observer rating had a low, but positive correlation of (0.17) with the students' rating. This shows that the classifier performed comparably to the human observers who monitored student body language and rated the students' confusion levels.

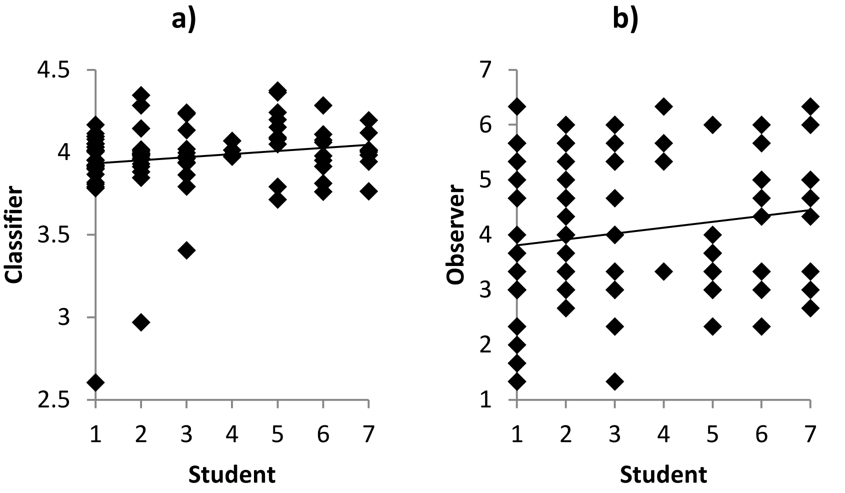


Fig. 2. Scatter plot of a) classifier vs. student rating, and b) observer vs. student rating

5 Conclusions and Future Work

In this paper, we described a pilot study, where we collected students' EEG brain activity while they watched MOOC video clips. We trained and tested classifiers to detect when a student was confused. We found weak but above-chance performance for using EEG to distinguish whether a student is confused. The classifier performed comparably to the human observers who monitored student body language and rated the students' confusion levels.

Since the experiment was based on a class project run by a group of graduate students, there were many limitations to the experiment. We now discuss the major limitations and how we plan to address them in future work.

One of the most critical limitations is the definition of experimental construct. Specifically, our pre-defined “confusing” videos could be confounded. For example, a student may not find a video clip on Stem Cell to be confusing when the instructor clearly explains the topic. Also, the predefined confusion level may be confounded with increased mental effort / concentration. To explore this issue, we examined the relationship between the predefined confusion level and the subjective user-defined confusion level. The students’ subjective evaluation of the confusion level and our predefined label has a modest correlation of 0.30. Next, we performed a feature selection experiment among all combinations of 11 features; we used cross validation through all the experiments and sorted the combinations according to accuracy. Then we found that the user-specific model Theta signal played an important role in all the leading combinations. Theta signal corresponds to errors, correct responses and feedback, suggesting the experimental construct is indeed related to confusion.

Another limitation is due to the lack of psychological professionalism. For example, the observers in our experiment were not formally trained. As a result, the current scheme allowed each observer to interpret a student’s confusion level based on his/her own interpretation. A precise labeling scheme would yield more details that could be compared among raters and, thereby, improve our rating procedure.

Another limitation is the scale of our experiment as we only performed the experiments with 10 students, and each student only watched 10 two-minute video clips. The limited amount of data points prevents us from drawing any strong conclusions about the study. We hope to scale up the experiment and collect more data.

Finally, this pilot study shows positive, but weak classifier performance in detecting confusion. The weak classifier performance may have many false-alarms and thereby frustrate a student. In addition, a student may not be willing to share their brain activity data due to privacy concerns. We are hopeful that the classifier accuracy can be improved once we conduct a more rigorous experiment, by increasing the study size, and improve the classifier with better feature selection and by applying denoising techniques to improve signal-to-noise ratio. Lastly, the classifiers are supposed to help students, so the students should be able to choose not to use EEG if they think the device is hindering.

Acknowledgments. This work was supported by the National Science Foundation under Cyberlearning Grant IIS1124240. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute, or the National Science Foundation. We thank Jessica Nelson for help with experiment design, Donna Gates for preparation of the manuscript, and the students, educators, and LISTENers who helped generate, collect, and analyze our data.

Reference

1. Allen, I.E., Seaman, J., *Going the Distance: Online Education in the United States, 2011*, 2011.
2. Thompson, G., *How Can Correspondence-Based Distance Education be Improved?: A Survey of Attitudes of Students Who Are Not Well Disposed toward Correspondence Study*. The Journal of Distance Education, 1990. **5**(1): p. 53-65.
3. Shute, V., et al. *Assessment and learning in intelligent educational systems: A peek into the future*. in *Proceedings of the 14th International Conference on Artificial Intelligence in Education Workshop on Intelligent Educational Games*. 2009. Brighton, UK.
4. Vardi, M.Y., *Will MOOCs destroy academia?*, in *Communications of the ACM*2012. p. 5.
5. Niedermeyer, E., Fernando H. Lopes da Silva, F. H., *Electroencephalography: basic principles, clinical applications, and related fields*2005: Lippincott Williams & Wilkins.
6. Marosi, E., et al., *Narrow-band spectral measurements of EEG during emotional tasks.* International Journal of Neuroscience, 2002. **112**(7): p. 871-891.
7. Lutsyuk, N.V., E.V. Éismont, and V.B. Pavlenko, *Correlation of the characteristics of EEG potentials with the indices of attention in 12- to 13-year-old children.* Neurophysiology, 2006. **38**(3): p. 209-216.
8. Berka, C., et al., *EEG correlates of task engagement and mental workload in vigilance, learning , and memory tasks.* Aviation, Space, and Environmental Medicine, 2007. **78 (Supp 1)**: p. B231-244.
9. Baker, R., et al., *Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments.* International Journal of Human-Computer Studies, 2010. **68**(4): p. 223-241.
10. NeuroSky, *Brain wave signal (EEG)*, 2009, Neurosky, Inc.
11. NeuroSky, *NeuroSky's eSense™ meters and dtection of mntal sate*, 2009, Neurosky, Inc.
12. Mostow, J., K.M. Chang, and J. Nelson. *Toward exploiting EEG input in a Reading Tutor*. in *15th International Conference on Artificial Intelligence in Education*. 2011. Auckland, New Zealand: Lecture Notes in Computer Science.
13. Crowley, K., et al., *Evaluating a brain-computer interface to categorise human emotional response* in *10th IEEE International Conference on Advanced Learning Technologies*2010: Sousse, Tunisia. p. 276-278.
14. Jasper, H.H., *The ten-twenty electrode system of the International Federation.* Electroencephalography and Clinical Neurophysiology, 1958. **10**: p. 371-375.
15. Ng, A.Y. and M.I. Jordan. *On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes* in *Advances in Neural Information Processing Systems* 2002. MIT Press.

Collaborative Learning in Geographically Distributed and In-person Groups

René F. Kizilcec

Department of Communication, Stanford University, Stanford CA 94305
kizilcec@stanford.edu

Abstract. Open online courses attract a diverse global audience of learners, many of whom might not be self-directed autodidacts with the necessary web competencies to reap the full benefits of such courses. Most of these learners would benefit from increased guidance on how to use MOOCs to enhance their learning. One potential area for guidance is in group collaboration where learners form teams to collaboratively work on assignments. Despite the global scope of these courses, a large proportion of learners live within relatively close proximity of each other, such that in-person collaboration is a feasible option. However, geographically distributed groups of learners are more likely to bring diverse viewpoints to the discussion than learners who live close to each other. Research suggests that the diversity of viewpoints in a group positively affects the quality of collaboration and outcomes. This paper reviews the literature on the feasibility of assigning local groups for collaboration and proposes concrete research directions.

1 Introduction

An increasing number of educators use online, asynchronous computer-mediated communication tools to create massive open online courses (MOOCs). These virtual classrooms attract a global audience of learners (Fig. 1) who join these courses for various reasons, including earning a certificate for completing the course or personal enrichment. The global and massive scale of these courses make them a melting pot for diverse ideas and perspectives: the learner population varies considerably in demographics, cultural background, language skills, personality, motivation, and prior knowledge.

Potentially the most important scholarly question in the midst of the rapid proliferation of open online courses is how learning can be enhanced with MOOCs. No simple answer can suffice, but it is clear that understanding the learner population is critical for developing strategies to foster learning. Borrowing a term from Lévi-Strauss [1], the online learner can be understood as a *bricoleur*–a handy-man or jack-of-all-trades–who cobbles together ways to learn from the plethora of online learning resources. The danger with this notion of the learner is that it is probably over-optimistic, given that many learners are not autodidacts or not "MOOC-ready" in other ways, e.g. not technologically adept. Hence, to ensure equal opportunities to learn, we need to provide guidance to learners to become skilled *bricoleurs* and continuously support them in their *bricolage* learning endeavor.

2 Collaborative Learning

Small group collaboration in and around MOOCs is a particularly fertile ground for increased guidance. The literature on computer-supported collaborative learning can provide theoretically and empirically grounded advice on how to support group collaboration. In addition, the rapid development of the online learning space is providing opportunities for empirical research, unprecedented in scale, to test existing recommendations and investigate novel approaches to guiding group collaboration in a variety of contexts.

Many contemporary MOOCs involve group projects as part of the course, providing learners with the opportunity to collaborate with a diverse set of people and to engage in a process of knowledge building. Group characteristics affect a group's performance, satisfaction, and processes of collaborative learning.

Group formation can follow one of two philosophies: laissez-faire (self-formed) or interventionist (assigned randomly or based on certain criteria). Both approaches raise questions of how groups are selected and the kind of guidance that should be provided from the MOOC interface or other sources.

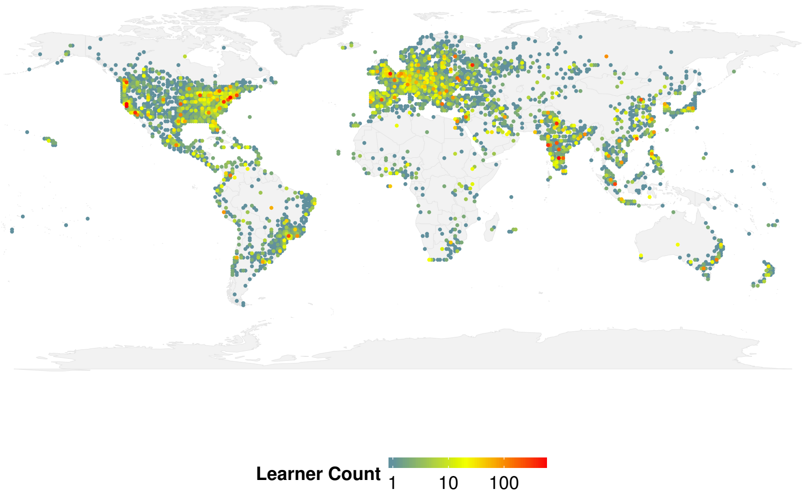
How should one form groups and guide them to encourage effective and fruitful collaboration? The remainder of the paper addresses this question. Section 3 motivates the distinction between geographically distributed and in-person groups, and presents evidence for the feasibility of assigning local groups. Section 4 reviews relevant literature on small group collaboration that can inform group assignment and guidance strategies. Section 5 proposes concrete research directions to empirically investigate strategies for group assignment and guidance, and proposes a collaboration model that combines geographical diversity and in-person collaboration. Section 6 presents concluding remarks.

3 Geographically Distributed or In-person?

Geographically distributed groups in MOOCs rely on computer-mediated communication (CMC) to work collaboratively on their project. These learners use video conferencing, and synchronous as well as asynchronous textual interfaces, such as email, instant messaging, and word processing applications with real-time collaboration. In contrast, geographic proximity can permit face-to-face (FtF) interaction. Of the two models, FtF collaboration has been associated with a significantly better learning experience in terms of the quality of group discussion and interactions compared to collaboration via asynchronous CMC [2]. This is not surprising given that FtF communication is a considerably more expressive medium than CMC.[1] However, no significant differences in learning measured by pre-post tests and self-report were found [2, 3].

[1] Interactions in immersive virtual reality are potentially more expressive than face-to-face, but the technology is not yet publicly accessible.

Fig. 1. Geographical location of active (interacted with learning materials) learners averaged over 21 MOOCs with colors representing geographical density of learners in the region. In green, yellow, and red regions, the learner population is sufficiently dense to support in-person collaboration.



In-person groups tend to be self-formed groups of friends, as geographic proximity is positively related to friendship. Such self-formed groups are subject to people's natural tendency to engage with people who are similar to themselves (homophily) [4]. The combination of homophily and the correlation between geography and demographic and other characteristics tends to make these groups even more homogeneous relative to, for instance, randomly-assigned groups. This can be a problem because collaborative learning in heterogeneous groups can be more effective than in homogeneous ones, as the wealth of alternative perspectives sparks innovative ideas [5, 6]. The research on the relationship between group members' friendship and outcomes remains split on whether collaborating with friends is beneficial [7].

The kind of guidance provided to learners partially depends on whether collaboration is in-person or computer-mediated. However, there has been no conclusive evidence that assigning groups to facilitate in-person collaboration in MOOCs is possible at a large scale. While a single MOOC attracts hundreds of thousands of learners, the feasibility of in-person collaboration relies on how many learners live close enough to fellow learners. To investigate the feasibility of in-person collaboration, geographical location data from 21 MOOCs on various topics was aggregated to produce two figures. Conclusions drawn from these data are very likely to be generalizable across MOOCs offered around the same time (late 2011 to early 2013) on MOOC platforms built around weekly video lectures and assignments.

Figure 1 illustrates the density of the active learner population on a world map.[2] Green, yellow, and red regions indicate geographical locations with sufficiently many learners to support in-person collaboration.[3]

Figure 2 illustrates the geographical density of active learners by the number of learners in the same region. At least three (five) learners live in 52% (37%) of the regions (dotted line). Moreover, due to the high learner density in a few big cities, 92% (85%) of learners live in regions with at least four (nine) other learners taking the same course (solid line). These data suggest that the distribution of learners in most parts of the world would support group assignments that facilitate in-person collaboration.

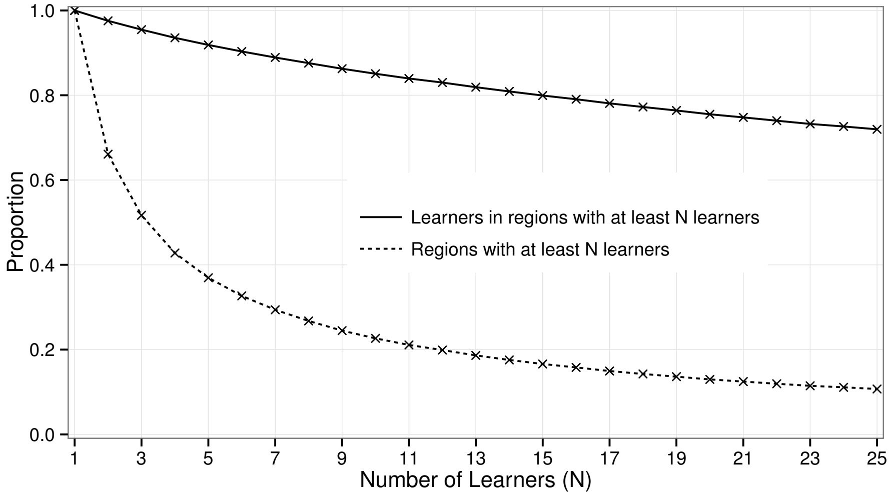
4 Relevant Literature

Scott Page's [8] work on group collaboration indicates that the diversity of viewpoints within a group is more important than the excellence of its individual members. It is reasonable to assume that people's diversity of viewpoints increases with the geographical distance between them, which would suggest that

[2] Active learners, a small subset of the enrolled learners, are defined to have used the learning materials at least once.

[3] Geographical location was determined based on users' IP address. A region is defined by all equivalent latitude/longitude coordinates rounded to zero decimal places. This definition of a region is not ideal, because the area within regions varies depending on geographical location, but it provides a rough estimate.

Fig. 2. Geographical topology of active learners (interacted with learning materials) averaged over 21 MOOCs. For 1 to 25 learners (N), the solid line illustrates the proportion of learners in regions with at least N learners and the dotted line illustrates the proportion of regions with at least N learners.



groups should be assigned with greater geographical diversity. However, there is potentially enough cultural diversity present in most major cities to assign groups with diverse viewpoints, while maintaining the geographical proximity to facilitate in-person collaboration.

Related to Page's research, Woolley and colleagues [9] report evidence for a collective intelligence in groups that has little association with the average or maximum individual intelligence of group members, but is highly correlated with the proportion of females in the group and the distribution of conversational turn-taking. While the gender distribution can be addressed by specific assignment of groups, the conversational dynamics within the group can only be influenced indirectly, for instance, by guiding group interactions technologically or with written guidelines on turn-taking. Online video conferencing tools could include timers for each participant, similar to chess clocks, to encourage balanced participation and turn-taking.

Barron's [10] findings provide further evidence that emphasizes the importance of nuanced process indicators in collaborative learning. She found indicators such as listening to proposals in group collaboration to be predictive of collaboration success, while less process-oriented measures such as group members prior achievements and how well they generated correct ideas were not correlated with positive problem-solving outcomes. Research on collaborative learning suggests that it is most effective when group members engage in rich interactions, like discussing conceptual explanations rather than providing specific answers. Thus, rich interactions can be encouraged by guiding the collaborative process

[11], for example, by providing note-taking templates that encourage certain behaviors, such as discussing conceptual explanations.

The collaboration process and how it should be guided depends on the communication medium used for collaboration. The expressiveness of the communication medium is a likely moderator of the richness of interactions [12], with FtF enabling more expressive interactions than CMC. However, advances in the learning sciences on collaborative learning with video [13] suggest that augmented CMC (augmented with tools to foster mutual awareness) can yield higher collaboration quality and learning gains than unaugmented CMC. Guidance to learners on the use of such tools, such as when and how to use them effectively, is necessary to maximize their potential benefit to learners. For instance, groups with geographically diverse members should receive guidance on several online collaboration tools, including the types of tasks that each is most suitable for and examples of how to use them effectively.

5 Research Directions

MOOCs provide researchers with a powerful platform for conducting experiments to address questions around collaborative learning in this novel context. The massive scale of these courses combined with randomized controlled field experiments can provide insights into the features of the learning environment and the kinds of guidance that can significantly enhance learning.

The effectiveness of geographically distributed compared to in-person collaboration with different models of guidance could be investigated by assigning half the project groups to maximize group members' geographic distance from each other and the other half to groups close enough to facilitate in-person collaboration. Groups could be randomly assigned to receive different guidance on collaboration strategies and technologies. Outcome measures should capture group performance (project grades and perceived learning), collaboration quality (e.g. Meier et al.'s [14] rating scheme), members' experience, and whether in-person collaboration took place for locally assigned groups. Moreover, a measure of perceived social and cultural group diversity could provide insights into the association between geographic distance and subjective group diversity, potentially an important mediator of the above outcome measures.

Beyond the question of how groups are actually assigned, the psychological implications of what learners are told about how their group members were chosen might influence their perception of the group and collaboration experience (framing effect). For example, telling learners that their collaborators were carefully chosen based on their personality and previous experience to promote productive collaboration and original ideas sets positive expectations compared to telling them that groups are randomly chosen.

An implementation that reaps the benefits of geographically distributed and in-person collaboration could be to facilitate collaboration in two steps: locally assigned groups could first collaborate in-person before connecting with a few other groups from around the world to form a larger, more distributed group that discusses the preliminary ideas and continues the collaboration online. This model of collaboration could be tested and adjusted through iterative improvement to optimize the collaboration experience.

6 Conclusion

Providing online learners with guidance, especially those who are not self-directed autodidacts, is necessary to ensure equal opportunity to learn. Group collaboration, where peers collectively solve a task or discuss an issue, is a potentially fruitful setting for increased guidance. Learning from and with peers to complement learning from the instructor is becoming increasingly important in online learning due to rapidly growing student-to-teacher ratios. It is therefore critical that collaborative learning is enhanced by providing learners with appropriate guidance.

What kind of guidance to provide will partly depend on the type of learner interaction. This paper argues that there is an important distinction between groups that have the potential for face-to-face communication and those who do not, especially as education moves out of brick-and-mortar institutions where students are all geographically accessible.

7 Acknowledgments

The author would like to thank Clifford Nass and Roy Pea for their feedback on an earlier draft of this paper and Stanford's Office of the Vice Provost for Online Learning for supporting this research.

References

1. Lévi-Strauss, C.: The savage mind. New York: Free Press. (1966)
2. Ocker, R., Yaverbaum, G.: Asynchronous computer-mediated communication versus face-to-face collaboration: Results on student learning, quality and satisfaction. Group Decision and Negotiation **8** (1999) 427–440
3. Francescato, D., Porcelli, R., Mebane, M., Cuddetta, M., Klobas, J., Renzi, P.: Evaluation of the efficacy of collaborative learning in face-to-face and computer-supported university contexts. Computers in Human Behavior **22**(2) (March 2006) 163–176
4. McPherson, M., Smith-Lovin, L., Cook, J.: Birds of a feather: Homophily in social networks. Annual Review of Sociology **27**(2001) (2001) 415–444
5. Webb, N.: Task-related verbal interaction and mathematics learning in small groups. Journal for Research in Mathematics Education **22**(5) (1991) 366–389
6. Nemeth, C.J.: Differential contributions of majority and minority influence. Psychological Review **93**(1) (1986) 23–32
7. Maldonado, H., Klemmer, S., Pea, R.: When is collaborating with friends a good idea? Insights from design education. In: Proceedings of CSCL-09 (Computer-Supported Collaborative Learning), Rhodes, Greece 227–231

8. Page, S.E.: The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies. Princeton University Press (2008)
9. Woolley, A.W., Chabris, C.F., Pentland, A., Hashmi, N., Malone, T.W.: Evidence for a collective intelligence factor in the performance of human groups. Science (New York, N.Y.) **330**(6004) (October 2010) 686–8
10. Barron, B.: When smart groups fail. The journal of the learning sciences **12**(3) (2003) 307–359
11. Dillenbourg, P., Schneider, D., Synteta, P.: Virtual Learning Environments. In Dimitracopoulou, A., ed.: 3rd Hellenic Conference ”Information & Communication Technologies in Education”, Rhodes, Greece (2002) 3–18
12. Daft, R., Lengel, R.: Organizational information requirements, media richness and structural design. Management Science **32**(5) (1986) 554–571
13. Goldman, R., Pea, R.D., Barron, B., Derry, S., eds.: Video research in the learning sciences. Lawrence Erlbaum Associates, Mahwah, NJ (2007)
14. Meier, A., Spada, H., Rummel, N.: A rating scheme for assessing the quality of computer-supported collaboration processes. International Journal of Computer-Supported Collaborative Learning **2**(1) (February 2007) 63–86