Understanding Learning and Learning Design in MOOCs: A Measurement-Based Interpretation

Sandra Milligan

Science of Learning Research Centre, Assessment Research Centre, Learning Analytics Research Group, University of Melbourne, Australia
s.milligan@unimelb.edu.au

Patrick Griffin

Emeritus Professor of Education, University of Melbourne, Australia

ABSTRACT: The paper describes empirical investigations of how participants in a MOOC learn, and the implications for MOOC design. A learner capability to generate higher order learning in MOOCs — called crowd-sourced learning (C-SL) capability — was defined from learning science literature. The capability comprised a complex yet interrelated array of attitudes, beliefs, and understandings about learning that participants bring to a MOOC and which shape their behaviour and explain why individuals differ in their ability to generate higher order learning. The capability was formulated as a developmental progression describing behaviours associated with five levels, from novice to expert, charting the degree to which learners regulate their own learning, effectively exploit the scale and diversity of MOOCs, and harness opportunities for distributed teaching. Item response theory was applied to log stream data in two MOOCs to construct empirically validated measures of this capability, enabling each MOOC learner to be assessed for learning capability on a scale from novice to expert. The majority of participants did not behave in ways conducive to the generation of higher order learning, but the C-SL progression suggested principles to guide MOOC design to make them more efficacious, which, when empirically investigated, were found to be efficacious.

Keywords: Measurement, IRT, Rasch, learning design, crowd-sourced learning, MOOCs, capability, 21st century skills

1   INTRODUCTION

This paper reports findings of empirical investigations into the quality and character of learning in MOOCs. The investigations originated in curiosity about whether or not MOOCs are able to support learners to generate higher order learning outcomes of the kind generally expected from university courses. Learning in a MOOC is unlike learning on campus or in traditional digitally mediated courses. In MOOCs, the familiar core of higher education is missing: there is no relatively homogenous class of motivated, like-minded students known to each other and who form stable working groups over a period of time. Teachers do not guide individuals' learning or monitor their induction into the attitudes, understandings, mindsets, tools, and techniques that make up a discipline or profession. A MOOC learner's progress, or lack of it, or even his or her very existence, can pass entirely unremarked by

anyone. Most participants do not complete (Breslow et al., 2013; DeBoer, Ho, Stump, & Breslow, 2014; University of Melbourne, 2014). Despite some indication that the experience of participants can be positive and educationally rewarding (Kop, 2011; Milligan & Griffin, 2015; Milligan, Littlejohn, & Margaryan, 2013; Veletsianos, 2013), critical commentary about the shortcomings of MOOCs as generators of higher order learning is common (Creelman, 2013; Daniel, 2012; Gillani, Yasseri, Eynon, & Hjorth, 2014).

Four questions arise. Given the distinctive environment of a MOOC, what learning processes would be effective in generating higher order learning? Can learning analytics be used to provide better insight into effective learning in MOOCs? Are principles of good pedagogy in a MOOC the same as those applying to on-campus courses or their e-learning counterparts? Can a better understanding of learning in a MOOC direct design improvements? Empirical investigations that address these questions are reported in this paper.

2   EFFECTIVE LEARNING PROCESSES IN MOOCS

The first question to consider is what learning processes are likely to be effective in generating higher order learning in a MOOC. The learning sciences have long examined attitudes and behaviours likely to distinguish successful from unsuccessful learners in higher education settings (Biggs & Tang, 2011). As a first step for the investigations reported here, a literature review was conducted to distill what from this literature could reasonably be applied to learners in MOOCs.

The body of literature specifically focused on MOOCs suggests that patterns of learning can be explained by the existence of a set of learning skills — the 21st century learning skills — required of learners. Stewart (2010, 2013) described digital media literacies that enable a learner to engage, to be confident, and to learn in a MOOC. These literacies included print and visual literacy, information literacy, critical thinking, ability to use hypertext, and mastery of complex etiquette. Kop (2011) took a similar view, citing a list of 21st century meta-literacies from the National Council of Teaching of English. The list included proficiency with tools of technology; building relationships with others to solve problems collaboratively and cross-culturally; managing, analyzing, and synthesizing multiple simultaneous streams of information; creating, critiquing, analyzing, and evaluating multimedia texts; and attending to ethical responsibilities. Siemens's (2004) theory of connectivism explored how learning occurs in networked environments. He argued that learners require the ability to traverse and construct knowledge networks, aggregate information, remix, repurpose, and share with others. Others have also suggested that distinctive learner capabilities are required (Ahn, Butler, Alam, & Webster, 2013; Fournier & Kop, 2010; Milligan et al., 2013; Littlejohn, Beetham, & McGill, 2012; Yeager, Hurley-Dasgupta, & Bliss, 2013). Similar discussion in non-MOOC contexts supports the existence of complex 21st century learning skills (Deakin Crick, Stringher, & Ren, 2014; Griffin, McGaw, & Care, 2012). For example, a construct "learning power" has been defined (Buckingham Shum & Deakin Crick, 2012; Deakin Crick, Broadfoot, & Claxton, 2004) as a complex set of skills or predispositions that students need

to move from the world of schooling where learning objectives are clear and articulated to one in which knowledge requirements can no longer be predetermined with confidence.

Another field of literature that had explanatory power for understanding the behaviour of learners in MOOCs is the study of self-regulation of learning, and the closely associated study in education of the use of feedback by learners. Butler and Winne (1995) identified the defining characteristics of successful self-regulated learning as the generation and use of feedback to guide learning. Feedback answers the three existential questions a self-regulated learner faces: Where am I going? How am I going there? Where to next? (Hattie & Timperley, 2007). The ability to self-regulate depends on range of cognitive, metacognitive, affective, and behavioural skills and beliefs, such as ability to set goals, capacity to motivate oneself, and ability to evaluate one's own performance (Bernacki, Aguilar, & Byrnes, 2011; Fournier & Kop, 2010; Kop, 2011; Zimmerman, 2002).

The extensive literature on the social construction of learning (e.g., Biggs & Tang, 2011) also generated relevant insights. For example, learners require skills to harness the learning potential of dialogue and peer feedback effectively (Prins, Sluijsmans, Kirschner, & Strijbos, 2005; Price, Handley, & Millar, 2011), including skills in critical reflection, and in calibrating performance standards for peer-based learning (Sadler, 2010). However, understandings from this literature required further consideration to reflect the distinctiveness of peer relationships in networked, scaled environments. Surowiecki's (2004) exposition on "the wisdom of the crowd" pointed out that the Internet has provided a powerful tool to capture individuals' willingness to collaborate. Jenkins (2009) identified these newer forms of organization as examples of "the participatory culture" of the Internet, arguing that Internet-based organizations are qualitatively distinct from earlier forms. Comparisons between Wikipedia and the Encyclopedia Britannica exemplify this point. MOOCs can operate as an example of participatory culture (Stewart, 2013) because they can harness scale and diversity to operate as networks of distributed participants supporting each other's learning, together providing a teaching resource potentially superior in many respects to a teacher. Participants can organize themselves to mutual learning benefit through forums, social media, and collaborative knowledge-building applications such as wikis, blogs, and aggregation services. But, as Jenkins points out, this demands distinctive skills and abilities.

These different strands of literature, with their diverse range in terminologies and perspectives, generated a surprisingly consistent understanding of the characteristics that individuals possess to a greater or lesser degree, and which determine their success in leaning. A construct was defined as the capability[1] a learner would require to generate higher order learning in the distinctive environment of a MOOC. Notably, the importance of technical or computer-related skills was de-emphasized, as were the ideas that what matters are immutable dispositions, or prior experience with social media, or Internet culture such as gaming. Rather, the capabilities are best described as a complex constellation of

---

[1] In this study, the term "capability" is used to refer to a learnable ability to meet complex demands by drawing on and mobilizing internal resources (including knowledge, understandings, skills, attitudes, and values) in a particular context.

attitudes, beliefs, values, knowledge, and understandings *about learning* that shape how a learner perceives the task of learning, and thus how they behave as learners in a MOOC.

Four inter-related themes were identified to capture this constellation, which together define a construct for the capability. Broadly, the four themes are:

1. *Epistemic standpoint*, which captures the differences between learners in the attitudes, values, and beliefs about the nature of knowledge that they aspire to learn. Learners may privilege abstract, generalized, universal, stable, expert understandings in a domain. Others give equal weight to learning based on practical wisdom, regarding knowledge as changing, highly contextual, socially defined, and widely distributed.

2. *Learning orientation*, which captures expectations about the process of learning and the intellectual, attitudinal, and emotional challenges inherent in it. Learners who regard learning as an individual act of consumption of knowledge, transferred from experts, will differ in approach from those who view learning as a messy, extended process of co-production, based on dialogue and collaboration with others, and involving emotional engagement, risk taking, confusion, and persistence.

3. *Orientation to teaching,* which captures the degree to which individuals embrace the idea that it is possible to learn from many sources, that teaching can be a distributed function, and that each learner can learn from and teach others. Those who believe that they can learn best from direct contact with expert teachers will behave differently from those who regard the Internet as an opportunity to exploit reciprocal teaching, shared experience, the wisdom of the crowd, and automated teaching agents.

4. *Self-regulation,* which captures differences between learners in where they locate control of their learning. Externally regulated learners are likely to believe that the decisions on content and processes of learning are best made by teaching staff, whom they trust to establish requisite standards, to judge performance, and advise on learning activity. Self-regulated learners, by contrast, control their own learning, set their own goals, calibrate their own and others' performance against internalized standards, and continually adjust activity to improve learning outcomes.

The literatures outlined above were also used to identify a number of behavioural correlatives explained by these attitudes, beliefs, and values about learning. Ten observable *behavioural elements* were identified: the breadth or narrowness of attention to different learning sources; the breadth or narrowness of perspective-taking behaviour; the degree of systematicacy, orderliness, and persistence applied in learning activity; the degree of dialogic and reciprocal interaction with peers; the balance of consumption versus production in learning activity; the degree of recursiveness applied in use of

feedback; the willingness to exercise critical consumption; the degree of risk-taking behaviour; and the level of participation in peer- and self-evaluation activities.

These four themes and their 10 behavioural correlatives are distinguishable from each other but should not be thought of as separable. The learning sciences literature shows that in some circumstances complex metacognitive constructs can usefully be represented as a single linear progression or continuum of capability that different individuals have to different degrees. (Cronbach & Meehl, 1955). Such a progression can comprise capacities that act together to exhibit a simple developmental integrity. An individual can be assessed to indicate whether they have more or less of that capability. Wilson and Scalise (2012) produced a developmental progression for the capability to create knowledge in an ICT-rich environment, based on a constellation of constituent skills, and abilities. Hesse, Care, Buder, Sassenberg, and Griffin (2015) proposed a developmental progression for the capacity to undertake collaborative problem solving, comprising a range of social and cognitive skills. The work of Dreyfus and Dreyfus (1980) was particularly influential in this study. The Dreyfus framework, and subsequent elaborations (Dreyfus, 2002, 2008; Luntley, 2009), articulated stages of competence (novice, beginner, proficient, competent, expert, mastery) that can be applied to the development of any complex skill, such as learning to be a good pilot or becoming an accomplished jazz musician, or, in the case of this study, becoming a learner capable of generating higher order learning in the distinctive environment of a MOOC.

The combination of the Dreyfus framework and the four themes and their behavioural correlatives identified above is outlined in Table 1, which represents an hypothesized developmental progression with five levels of capability of learners to generate higher order learning from MOOCs. For convenience, this capability is referred to in the rest of this paper as crowd-sourced learning (or C-SL) capability.

3   CONSTRUCTING A MEASURE OF CAPABILITY FROM LOG STREAM DATA

The definition of a construct for C-SL capability provided the basis for exploring the second question addressed in this study: Can learning analytics be used to provide better insight into effective learning in MOOCs? To empirically explore this question, it was decided to investigate the use of log stream data in MOOCs to develop measures of individuals' C-SL capability. Log stream data is the extensive, detailed, time-stamped, digital record of each action that each participant makes using the keyboard, mouse, track-pad, or touch screen while working on a MOOC platform.

The task of measuring individuals' C-SL capability is, in principle, well within the scope of methodologies of measurement science in education. It is now common for reliable, valid, scaled assessments of complex capabilities to be generated for diverse, global cohorts across different contexts. High-stakes educational assessment programs such as PISA (OECD, 2014) provide examples. Further, one of the prerequisites for such methodologies — a theoretically derived developmental progression that

articulates the nature of the underlying construct and the associated behaviours that distinguish higher from lower performing individuals — was available (as shown in Table 1).

Use of digital or log stream data in conjunction with measurement methodologies (rather than more traditional test items or assessment inventories) is in its early stages, but this, too, is not unknown. Related methods have recently been applied to the measurement of complex generic skills using log stream data generated from especially designed online tasks or online games (Griffin & Care, 2015; Shute & Ventura, 2013). The key challenge in the study reported here was not just to measure the capability of learners using digital data from digital tasks, but to do it using the natural log stream, which is a by-product of MOOC platforms and not designed for any particular purpose.

Log stream data does not provide a complete picture of all learner behaviours; for example, it excludes those forms of engagement with learning that occurs off-line or may not leave a digital trace, such as in some forms of emotional engagement (Fournier, Kop, & Sitlia, 2011; Veletsianos, Colleer, & Schneider, 2015). The data are rarely designed for research purposes, and have been found variously to be too granular or not granular enough for some purposes (Kizilcec, Piech, & Schneider, 2013), or too patchy or incomplete, or too inaccurate, confounded, or corrupted with data unrelated to learning (Dringus, 2012; Greller & Draschler, 2012; Siemens & Long, 2011). However, for this study, the real challenge was to work with it, adopting methods and approaches to overcome its shortcomings.

The methodology selected to test this was *constructing and validating measures using item response theory* (IRT) (Messick, 1995; Wilson, 2005; Wright & Masters, 1982). It was applied to data derived from two Coursera MOOCs from the University of Melbourne: Assessment and Teaching of 21st Century Skills (ATC21S) and Introduction to Macroeconomics (Macro MOOC). These two MOOCs were selected because their curriculum, pedagogy, and cohort characteristics were as different as it is possible to get within the University of Melbourne MOOC program, thus providing a test of generalizability of results across MOOCs. Macro MOOC is a very large, undergraduate-level, quantitative MOOC (more than 60,000 registrants in 2013), attracting relatively young, male participants, many from developing countries. ATC21S MOOC is relatively small (18,000 registrants in 2014) targeting professional development for practicing teachers, with an older, predominantly female cohort (University of Melbourne, 2015; Milligan & Griffin, 2015).

The methodology to construct valid measures from the log stream data, building on the definition of the construct and identification of behavioural elements, progressed through six steps.

Step 1. Selection of indicators: The log stream was first explored to clean it for measurement purposes, and to identify any indicators suitable for distinguishing expert from novice behaviours identified in the construct. For these MOOCs, data included digital time-stamped records of every activity of each learner, including patterns of access to videos, readings and resources, quizzes, or polls. It contained drafts, submissions, and evaluations of performance of peers and self on each assessment. The

frequency and location of forum viewing, posting, voting, and thread creation was captured, along with the content of every post, comment, or assessment.

The selection of indicators from the array available was an iterative, interpretive exercise guided by the literature, and by experience. For example, expert learners and novice learners are both likely to watch videos early in the course, and that data did not discriminate between them. Experts but not novices were likely to post heavily, although experts were not always the most prolific viewers or the most prolific posters (Huang, Dasgupta, Ghosh, Manning, & Sanders, 2014), so those data required careful interpretation. Hundreds of potential indicators were identified, examined, and, if selected, mapped to one or more behavioural elements to establish a range of indicators for each element.

Step 2. Definition of variables: Selected indicators were used to define variables that reflect the range of behaviours of individuals. For example, reposting in a thread was selected as an indicator of dialogic behaviour, and this was used to construct the variable "the number of weeks in which an individual posted more than once in a thread." Values for individuals on this variable ranged from zero weeks to the full number of weeks of the course duration.

Step 3. Coding to generate thresholds: Variables were then coded to generate scoring thresholds. For instance, for the variable "number of weeks in which an individual posted more than once in a thread," a value of three weeks was selected as a threshold. Those individuals who reposted in a thread in three or more weeks were coded "1," and others coded "0." This process generated a dichotomous (or sometimes polytomous) response for each person on each variable, and the frequency count in each threshold category provided a quasi-measure of the relative difficulty of that threshold.

Step 4. Scoring individuals: Raw scores of individuals on the capability measure were constructed by summing their coded scores on each threshold.

Step 5. Measurement modelling, calibration, and equating: Measurement modelling and calibration were undertaken to equate measures across the two MOOCs, and to check technical robustness. This step generated a standardized score for each individual (expressed in units of a logit), which was used to infer their position on the C-SL capability scale. The Partial Credit Measurement Model (Masters, 1982) was selected, as expressed in the software program Conquest (Wu, Adams, & Wilson, 1998). Fit to the model requires that responses to the thresholds have the technical prerequisites of good measurement. The thresholds must form a scale for which there is an underlying linear magnitude, and individuals must be able to be ordered conjointly on the same scale. A standard unit of measurement (the logit) must operate without variation over the range of the scale to quantify what "more" or "less" means (Wright & Masters, 1982). The set of items should be distributed across the range of abilities.

Table 1: A theoretically derived developmental progression for C-SL capability with construct themes and their behavioural correlates.[2]

NOVICE	BEGINNER	EMERGENT	COMPETENT	EXPERT
Believes that the goal of learning is mastery of stable, objective, generalizable knowledge and understandings, defined by experts		**THEME 1: EPISTEMIC STANDPOINT**		*Sees learning as growth in mastery in a domain; values practical wisdom and experience, including knowhow, attitudes, beliefs, values, ethics, and conventions; believes that knowledge changes, is contextual, is widely distributed around networks, and is socially defined.*
VALUES ABSTRACT KNOWLEDGE **Breadth of attention:** focuses on content from authoritative sources; aims to cover course content	**VALUES APPLIED ABSTRACT KNOWLEDGE** **Breadth of attention:** focuses on range of inputs from authoritative texts and sources	**VALUES APPLIED, CONTEXTUALIZED KNOWLEDGE** **Breadth of attention**: scans the range of MOOC features **Systematic and persistent:** in relation to authoritative texts and teacher-based feedback features	**VALUES PRACTICAL WISDOM IN OWN CONTEXT** **Breadth of attention**: scans the full range of MOOC features **Systematicity, orderliness, persistence**: persistent and systematic; engages strongly with authoritative teacher-supplied documents and feedback, and sometimes forums	**VALUES BROADLY DISTRIBUTED PRACTICAL WISDOM** **Breadth of attention**: eclectic in use of the range of MOOC elements, including texts, case studies, exercises such as quizzes, in-class responses; seeks inputs of experienced and expert peers through forums and other social media **Systematicity, orderliness, persistence**: engages with the range of elements of the course over the full duration of the course, persistently, frequently, and systematically **Perspective taking**: seeks out and explores diverse perspectives; curious; trusts value of learning from contexts unlike their own
Sees learning as a process of individual consumption of expert knowledge; sees expert teachers as responsible for resources, processes, assessments, and standards		**THEMES 2 & 3: ORIENTATION TO TEACHING AND LEARNING**		*Sees learning as a messy, effortful, emotionally demanding co-production; regards teaching services as being diverse and distributed; believes learners in networks have the capacity and responsibility for teaching and supporting the learning of others*
INDEPENDENT CONSUMER of EXPERT KNOWLEDGE **Production**: completes graded exercises only	**INDEPENDENT CONSTRUCTOR OF LEARNING** **Production**: understands learning as the organized consumption of content from authoritative sources, and as a process of incorporation of knowledge and understanding; accesses teacher texts, and exercises and completes graded exercises only **Recursiveness**: focuses on teaching texts	**PARTICIPATIVE CONSTRUCTOR OF OWN LEARNING** **Production**: sees learning as involving both consumption and production. Tries out own understanding and knowledge, through accessing automated response features like quizzes **Recursiveness**: focuses on teaching texts and automated exercises and feedback **Dialogic activity, reciprocity, critical consumption, risk taking**: sees teachers as responsible for sourcing resources, content, ideas, assessments; experienced peers might assist in interpretation	**COLLABORATIVE CONSTRUCTOR OF OWN LEARNING** **Production**: sees learning as involving both consumption and production; tries out ideas, attitudes, theories; practices skills; generates "performances" **Recursiveness**: recursively interrogates available automated feedback from sources such as quizzes, automated feedback, or reflective processes or peer comment **Dialogic activity**: open to using opportunities for dialogue and collaboration with others; interested in observing others' views, especially if in contexts similar to own **Reciprocal learning/teaching**: recognizes that others might provide resources, ideas, or experience of value to own learning; open to sharing	**RECIPROCAL LEARNER/TEACHER, CONSTRUCTING OWN AND OTHERS' LEARNING** **Production**: sees learning as involving both consumption and production; actively tries out ideas, attitudes, theories; practices skills; actively and frequently generates "performances" with posts, essays, blogs, images; argues new positions; articulation of new processes; explores gaps **Recursiveness**: recursively interrogates available feedback on own performance from diverse sources such as quizzes, automated feedback, or reflective processes or peer comment, until value is exhausted **Dialogic activity**: creates opportunities for and engages in extended dialogue with others through viewing, posting and voting, and use of social media **Reciprocal learning/teaching**: recognized and acknowledged by peers for leadership in opinion, advice; takes responsibility for collective learning; values reciprocity in learning/teaching; uses crowd-sourcing; open to learning from diverse sources and working with diverse others; contributes to the learning of others **Critical consumption**: independent-minded, consumes critically, makes independent judgments of the relevance and value of inputs and contributions to learning; trusts own judgment of quality of input and acts on it **Risk taking**: open to reputational risk; expresses opinion; may generate negative response or express non-conformist views; open to risk of failure, trying new things, being confused, and emotionally involved
Regulated by course structure; relies on teacher/expert judgment to gauge success		**THEME 4: REGULATION OF LEARNING**		*Self-regulated; internalizes and reflects on standards; sets own learning goals and monitors, explores, supports, and evaluates own and others' learning, and adjusts learning accordingly*
EXTERNALLY REGULATED **Monitoring/evaluation**: Sees standards as fixed, external to self; relies on grading assessments	**EXTERNALLY REGULATED** **Monitoring/evaluation**: engages with grade-related assessment; sees standards as fixed, external to self; trusts performance guidance and judgments on performance from authoritative sources	**EXTERNALLY REGULATED** **Monitoring/evaluation**: engages with some formative assessment and feedback features of the MOOC; applies performance standards set by teachers to evaluate own and others performance for grading	**SELF-REGULATED** **Monitoring/evaluation**: engages with the range of assessment and feedback features of the MOOC; generates feedback on own performance on tasks; interested in other's performance **Peer evaluation**: applies performance standards set by teachers to evaluate own and others performance; conscientious in peer evaluation.	**SELF-REGULATED AND CO-REGULATING** **Monitoring/evaluation**: self-reflective; engages with the range of formal and informal assessment and feedback features of the MOOC; seeks to generate feedback and advice on own performance and the performance of others in tasks and at holistic level; seeks to reconcile conflicting feedback **Peer evaluation:** actively seeks out opportunities to collaborate, share, express, and share opinions about performance standards; interprets performance of others in a range of contexts on tasks, informal and formal, and at holistic level; seeks to provides helpful learning feedback and teacherly advice to peers

---

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

95

The construct map for C-SL capability — showing the details of the relationship between the construct, the 10 behavioural elements, 28 selected indicators, 60 variables, and their 80 thresholds — is shown in Figure 1. The construct map also shows the difficulty level (one for each threshold) expressed in logits, and the Infit mean square error (one for each item formed from a variable) as explained in technical appendix derived from measurement modelling.

In each MOOC, calibration and fit of the scale used a sample of MOOC participants who had the most complete records (3320 students in Macro MOOC and 4438 in ATC21S MOOC). In both MOOCs, the model was found to fit, with model parameters within an acceptable range. In addition, the item functioning in each MOOC was found to be comparable: the Spearman rank correlation coefficient for item difficulties calculated independently in the two MOOCs was 0.96. Calibration of the items using a conjoined calibration sample, including both Macro and ATC21S participants, also supported the assumption that the measures work equally well across MOOCs. A technical appendix summarizes the indicators of fit to the partial credit model in the two MOOCs and in the conjoined sample. The satisfactory level of fit enabled the participants to be arranged in order of their level of inferred capability on a common scale applicable in both MOOCs.

It was also possible at this stage to verify the construct definition empirically. By arraying the various thresholds on the basis of their logit scores, it was possible to distinguish successive categories of expert, competent, emergent, beginning, and novice learners according to their behaviours. Table 2 summarizes the resultant empirical expression of the progression, showing how real-world behaviour patterns evidenced in the log stream were expressed.

Step 6. Validity investigations: A range of empirical and theoretical investigations were designed to build confidence that there were no flaws that make implausible the interpretation and usefulness of the metric (Messick, 1995; Kane, 2013; Wolfe & Smith, 2007). Of particular interest were investigations of bias attributable to demographic characteristics. Demographic information was available for 1,198 self-selected participants who completed a pre-course survey in ATC21S MOOC. Investigations of differential item functioning (Wilson, 2005) found no significant item response bias in relation to gender, age, previous experience in the domain, previous experience with MOOCs, level of previous educational attainment, or fluency with English. Spearman Rank correlations between the item parameters calculated independently for each of these sub-groups in the ATC21S MOOC were all above 0.98.

Perhaps the key finding from the array of validation investigations was that the C-SL scores significantly predicted the independent measure of final grades in both MOOCs. This relationship is shown for all 8,468 active participants in the ATC21S MOOC (Figure 2) as well as for the 2,728 persistent, diligent participants who did all the assessments and who were thus candidates for certification (Figure 3).

THE DOMAIN	CAPABILITIES	BEHAVIOURAL INDICATORS	DERIVED VARIABLE	ITEM CATEGORIES	DIFFICULTY	MSQ INFIT
PART 1 CAPABILITY TO LEARN IN SCALED ENVIRONMENTS	BREADTH OF ATTENTION	Extent of forum access	36. No. forum sessions	Viewed formums once	-5.27	0.93
				Viewed forums twice or more	-3.74	
		Extent of voting participation	16. No. voting sessions	Voted in forums	0.55	1.08
		Extent of syllabus guide engagement	37. Proportion of syllabus docs accessed	Viewed >25% of syllabus documents	-3.41	1.00
		Practice quiz engagement	49. Proportion of practice quizzes accessed	Particpated in >75% quizzes on offer	-1.35	1.20
		Engagement with embedded video quiz	54. Proportion of video quizzes accessed	Participated in up to 75% of vide quizzes	-2.46	1.37
				Participated in >75% of video quizzes	-0.94	
	DIALOGICNESS/ RECIPROCITY	Pattern of revisiting threads	13. No. sessions including a revisit	Revisited a thread	-1.93	0.76
			15. No. threads re-visited	One thread revisited	-1.6	1.02
				Five threads revisited	-0.28	
				Six or more threads revisited	1.22	
			14. No. threads with two or more revisits	1 thread with more tthan 1 revisit	0.62	1.02
				More than 1 thread with more than 1 revisit	2.23	
			12. No. weeks of thread revisited	Revisted threads in at least one week	-1.47	0.91
				Revistied threads in two or more weeks	0.62	
				Revisted threads in three or more weeks	0.61	
		Pattern of re-posting to threads	21. No. of threads with mulitple posts	Re-posted in 1 or more threads	1.67	0.83
			19. No. of threads with re-posts in a week	Re-posted in a thread within a week	1.74	0.83
			25. No. of re-posts to threads per week	Re-posted twice or more in a week	1.06	0.82
			29. No. weeks with re posts to threads	Re-posted in 3 or more weeks	3.92	0.80
			27. Max. no. posts in any thread	Posted 3 or more times in a thread	3.62	0.94
	CRITICAL CONSUMPTION	Breadth of voting locations	18. No. re-votes	Re-voted in a thread	1.84	1.04
			20. No. threads with re-votes	Multiple votes in at least one thread	1.82	1.04
			26. Max. no. of votes per thread	Votes three or more times in one thread	3.44	1.00
			23. No. of weeks in which re-voted in threads	Re-voted in threads in two or more weeks	3.26	0.85
		Timing of access to study guides	44. No. of study guides accessed	Study guides accessed	-4.07	1.01
			47. No. weeks of study guides accessed	Study guides accessed in >1 week	-1.20	1.00

Figure 1 (part 1): Construct map for C-SL capability as expressed in MOOC log stream data

THE DOMAIN	CAPABILITIES	BEHAVIOURAL INDICATORS	DERIVED VARIABLE	ITEM CATEGORIES	DIFFICULTY	MSQ INFIT
PART 2: CAPABILITY TO LEARN IN SCALED ENVIRONMENTS	PERSISTENCE/ SYSTEMATICACY	Timing of forum access	1. Viewed forums in Quarter (Q) 2 Week (W) 1	Viewed forums in Q2 W1	-0.69	1.14
			2. Viewed forums in Q2 W2	Viewed forums in Q2 W2	0.29	1.14
			3. Viewed forums in Q3 W1	Viewed forums in Q3 W1	0.44	1.08
			4. Viewed forums in Q3 W2	Viewed forums in Q3 W2	0.53	1.04
			5. Viewed forums in Q4 W1	Viewed forums in Q4 W1	0.51	1.03
			6. Viewed forums in Q4 W2	Viewed forums in Q4 W2	0.20	1.11
			7. Viewed forums post close	Viewed forums post close	0.35	1.10
		Consistency of forum viewing	11. No. weeks forum viewing	Viewed forums in 3 or more weeks	-0.45	0.85
		Timing of voting	22. No. weeks voting	Voted in two or more weeks	2.17	0.89
		Consistency of access to video quizzes	38. No. video quizzes submitted Q2 W1	Active in Q2 W1 video quizzes	-2.19	1.12
			39. No. video quizzes submitted Q2 W2	Active in Q2 W2 video quizzes	-1.28	1.01
			40. No. video quizzes submitted Q3 W1	Active in Q3 W1 video quizzes	-0.45	1.02
			41. No. video quizzes submitted Q3 W2	Active in Q3 W2 video quizzes	0.18	1.09
			42. No. video quizzes submitted Q4	Active in final quarter video quizzes	0.3	1.11
	PERSPECTIVE TAKING	Breadth of thread coverage	8. No. threads viewed	Viewed > 3 threads	-2.12	0.89
				Viewed > 30 threads	0.01	
				Viewed > 40 threads	2.41	
		Frequency of forum viewing	9. No. of forum viewing sessions	More than 3 viewing sessions	-2.43	0.72
				More than 10 viewing sessions	-0.45	
				More than 40 viewing sessions	1.66	
		Intensity of forum viewing	10. No. of threads viewed in one week	Viewed up to two threads max. in a week	-5.74	1.02
				Viewed up to 5 threads max. in a week	-0.05	
				Viewed up to 15 max. threads in a week	1.57	
				Viewed >15 threads max. in a week	4.13	
	PRODUCTION ORIENTATION	Extent of posting and commenting	17. No. of posts made	Posted up to three times	-0.43	0.97
				Posted more than three times	1.33	
			24. No. of weeks of posting	Posted in two or more weeks	1.43	0.72
		Depth of video quiz engagemnt	53. No video quizzes submitted	At least one video quiz submitted	-3.46	1.17
			56. No. of video quizzes submissions	Average submission up to 0.75 per quiz	-2.89	1.45
				Average submissions > 0.75 per quiz	-1.9	
		Word counts in posts and comments	58. No. posts with expository length	Posted >2 times at expository length (>400)	2.28	0.81
			59. No. of words posted	Posted up to 400 words total	-0.15	0.97
				Posted over 400 words total	1.67	
			60. No. of posts longer than tweets	Posted >2 posts of >120 words	0.93	0.75

Figure 1 (part 2): Construct map for C-SL capability as expressed in MOOC log stream data

THE DOMAIN	CAPABILITIES	BEHAVIOURAL INDICATORS	DERIVED VARIABLE	ITEM CATEGORIES	DIFFICULTY	MSQ INFIT
PART 3: CAPABILITY TO LEARN IN SCALED ENVIRONMENTS	RECURSIVENESS/ FOCUS	Extent of video quiz resubmission	55. No. video quiz resubmissions	Resubmitted at least one video quiz	-3.38	1.18
			57. Proportion of video quizzes resubmitted	Resubmitted in < 50 % video. quizzes	-2.18	1.48
				Resubmitted in ≥ 50% video quizzes	-0.15	
		Patterns of re-use of syllabus guides	46. No. re accessions of any study guides	Re accessed study guide documents	-5.73	1.18
				Re acessed documents ≥10 times	-3.08	
				Re accessed documents > 20 times	-2.83	
		Pattern of re-voting in threads	28. No. threads with re-voting	Re-voted in 2 or more threads	3.06	0.90
		Practice quiz resubmission	50. No. practice quizzes with resubmission	Resubmitted at least one quiz	-1.39	1.34
	RISK TAKING	Peer response to threads created	30. No. of views of created threads	Attracted >10 views of a thread created	1.97	1.03
			32. No. of votes on created threads	Attracted votes to threads createed	2.93	1.08
		Peer response to posts created	31. No. of views of posts created	Attracted >10 views of posts created	2.02	1.03
			34. No. of votes on posts created	Attracted votes on posts	0.81	0.86
			35. No. negative votes of posts created	Attracted negative votes to posts	2.99	0.97
		Coursera reputation points	33. Coursera reption points	Awarded Coursera reputation points	0.87	0.87
	PEER EVALUATION	Exploration of peer assessment	43. No.times accessed peer assessment guide	Accessed peer asssessment guide	-2.68	1.14
			45. No. re accessions peer asseszment guide	Re-accessed peer assessment guide	-2.44	1.24
	SELF-MONITORING/ EVALUATION	Pattern of practice quiz usaage	48. No. submissions of practice quizzes	Submitted one or more practice quizzes	-5.46	1.15
			51. Average resubmissions per quiz	Average on practice quiz submissions ≥1	-0.97	1.28
			52. No. of practice quiz resubmissions	Re-submissions >2 in 2nd/ later prac quizzes	1.26	1.47

SCORE

Figure 1 (part 3): Construct map for C-SL capability as expressed in MOOC log stream data.

99

Table 2: Empirically determined thresholds of C-SL progression in MOOCs.

LEVEL	THRESHOLD DESCRIPTION	LOGIT SCORE
EXPERT	Maximum thread views in any week is more than 15	4.13
	Re-posted in a thread in three or more weeks	3.92
	Posted in at least one thread three or more times	3.62
	Voted in at least one thread three times	3.44
	Re-voted in a thread in two or more weeks	3.26
	Re-voted in two or more threads	3.06
	Attracted negative votes on posts	2.99
	Attracted votes to threads created	2.93
	Viewed more than 40 threads	2.41
	Posted three or more expository-length posts	2.28
	Triple-visited more than one thread	2.23
	Voted in two or more weeks	2.17
	Attracted more than 10 views to a created post	2.02
	Attracted more than 10 views to a created thread	1.97
	Re-voted in a thread	1.84
	Votes more than once in the forums	1.82
COMPETENT	Reposted in a thread within a week	1.74
	Posted dialogically in threads	1.67
	Posted 400 or more words	1.67
	Viewed forums on more than 40 occasions	1.66
	Maximum thread views in any week is 15 or less	1.57
	Posted in at least two weeks of the course	1.43
	Posted or commented more than three times	1.33
	Resubmitted three or more times in second and subsequent practice quizzes	1.26
	Six or more threads revisited	1.22
	Reposted more than once in any week	1.06
	Posted three or more posts of length >twitter (90 words)	0.93
	Scored reputation points	0.87
	Received votes on posts	0.81
	Revisited threads in at least two weeks	0.62
	Triple-visited at least one thread	0.62
	Revisited threads in three or more weeks	0.61
	Voted in forums	0.55
	Viewed forums in second week of third quarter of the course	0.53
	Active in final quarter video quizzes	0.30
	Active in third quarter week 2 in video quizzes	0.18
	Viewed more than 10 threads	0.01
EMERGENT	Maximum thread views in any week is five or less	-0.05
	Posted 90 to 399 words	-0.15
	Re-accessed more than half the video quizzes	-0.15
	Up to five threads revisited	-0.28

LEVEL	THRESHOLD DESCRIPTION	LOGIT SCORE
	Posted or commented up to three times	-0.43
	Viewed threads in three or more weeks of the course	-0.45
	Active in third quarter week 1 in video quizzes	-0.45
	Viewed forums on more than 10 occasions	-0.45
	Viewed forums in quarter 2 week 1 of the course	-0.69
	Accessed more than half the 75% of quizzes	-0.94
	Average of one or more submissions per quiz	-0.97
	Accessed syllabus guides in more than one week	-1.20
	Active in second quarter week 2 in video quizzes	-1.28
	Participated in 75% of more of or more practice quizzes	-1.35
	Resubmitted at least one practice quiz	-1.39
	Revisited threads in at least one week	-1.47
	One thread revisited	-1.60
BEGINNER	Average video quizzes submissions 75% of no. of quizzes	-1.90
	Revisited threads	-1.93
	Viewed more than three threads	-2.12
	Re-accessed fewer than half the video quizzes	-2.18
	Active in second quarter week 1 in video quizzes	-2.19
	Viewed forums on three or more occasions	-2.43
	Re-accessed peer-assessment guides	-2.44
	Accessed fewer than half 75% of the video quizzes	-2.46
	Accessed peer-assessment guides	-2.68
	Re-accessed syllabus documents more than 20 times	-2.83
	Up to 75% of video quizzes submitted	-2.89
	Re-accessed syllabus documents more than 10 times	-3.08
	Re-accessed a video quiz	-3.38
	Viewed more than a quarter of syllabus documents	-3.41
	Participated in video quizzes	-3.46
	Viewed forums two or more times	-3.74
	Accessed study guide	-4.07
	Viewed forums once	-5.27
	Participated in at least one practice quiz	-5.46
	Re-accessed study guide documents	-5.73
	Maximum forum views in any week is two or less	-5.74
NOVICE	Visited the course, watched videos No use of forums, practice quizzes, video quizzes, or syllabus guides	
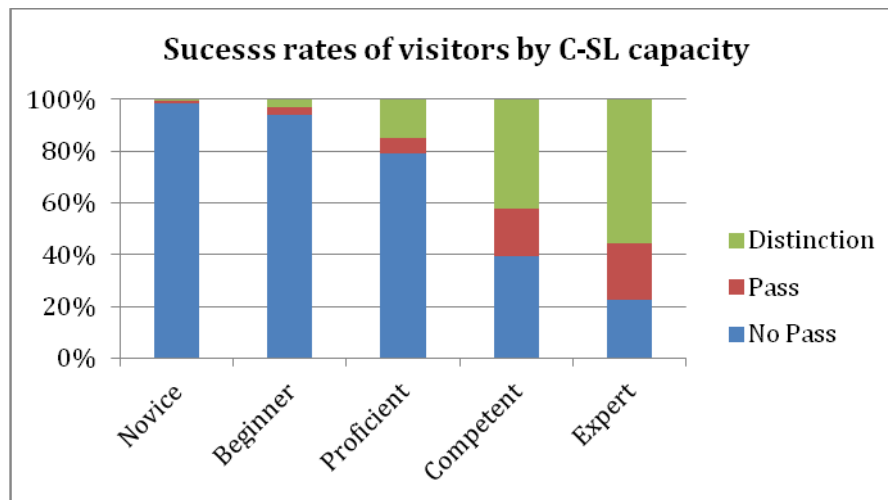
Figure 2: Relationship between course grade and C-SL measures; all visitors to the ATC21S MOOC.
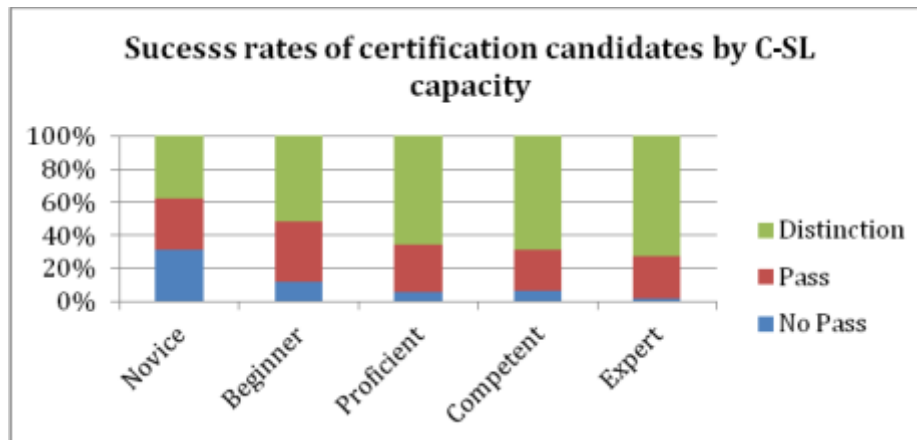
Figure 3: Relationship between course grade and C-SL measures; certificate candidates in ATC21S MOOC.

Overall, it was judged that the validity investigations provided no evidence to disrupt the core presumption that the metric provides a reasonable (although not perfect) basis for inference of the level of C-SL capability of learners, as described in the progression. A range of qualifications, interpretations, and nuances were identified during validation investigations that should be considered in any proposed use of the metric or the construct (Milligan, 2015). Discussion of these is beyond the scope of this paper.

4 IMPLICATIONS OF C-SL CAPABILITY FOR MOOC DESIGN

The third question for this paper was: Are principles of good pedagogy in a MOOC the same as those applying to on-campus courses or their e-learning counterparts?

Of note is that few participants in the two MOOCs performed above the emergent level of the C-SL capability scale. In Macro MOOC, even among those 2,464 persistent and serious participants who undertook a sufficient number of graded assessments to pass, only 6% performed at the expert level. A

further 23% performed at the competent level, 40% at emergent, 13% at beginner, and 18% at novice levels. Comparable results were found in ATC21S MOOC. In short, 30% of even serious and persistent participants exhibited behaviour that is in any sense self-regulating. Only 30% made any use of the crowd to support their learning, and fewer than 1% of all participants and perhaps as few as 6% of serious and persistent participants engaged in reciprocal teaching. It is possible to conclude that MOOCs, in their current, early form are used by the majority of participants in ways not conducive to the generation of higher order learning. To the extent that the empirically verified developmental progression for C-SL capability provides insight into how learners can generate higher order learning in MOOCs, it should also provide assistance to MOOC designers wishing to maximize the efficacy of MOOCs in this regard. To explore this idea, four design principles for shaping pedagogy (including assessment) were derived from C-SL capability.

A first principle was to *maximize scale and diversity of the participant cohort* in a course. Scale and diversity are prerequisites for harnessing the crowd for effective reciprocal, distributed teaching. Yeager et al. (2013) reported differences in learner experience between a MOOC that had more than 515 participants and another that had 76 participants. By their reckoning, the larger of their MOOCs had sufficient scale but the second did not. Waite, Mackness, Roberts, and Lovegrove (2013) suggested that 150 active participants are required, the number derived by the English anthropologist Robin Dunbar and arising from his work on the primate brain. Dunbar estimated that it is possible to maintain only 100 to 230 "normal" social relationships without cognitive overload or social disruption. The logic of this as it is applied to MOOCs is that "normal" learning interactions are suitable in unscaled environments. Beyond that, different patterns of interaction emerge, and crowd wisdom and distributed teaching can result (Dron & Anderson, 2014).

A second principle was to scaffold activities to generate and support self-regulation, crowd-sourced learning, reciprocal teaching, and the use of automated teaching agents. A key challenge for learning designers is to exploit the distinctive teaching features of MOOCs. Self-regulation thrives when there is plentiful automated machine feedback of different kinds for each learner on their own performances, covering the full range of learning outcomes. As digital technologies and artificial intelligence develop, the efficacy of automated feedback is likely to evolve rapidly. Equally important is promoting the emergence of crowd-based reciprocal teaching by supporting dialogue and collaboration between peers as reciprocal teachers. Early evidence has emerged as to the capacity of learners to provide reciprocal teaching services, such as identifying quality material, guiding selection of what to read, identifying quality posts, and evaluating peers (Gunnarsson & Alterman, 2013, 2014). The work of others (Sadler, 2010) has suggested that engagement with exemplars and in dialogues about holistic quality can assist learners, over time, to develop the tacit knowledge of unseen, unarticulated knowhow that marks out an expert from an inexpert appraiser, and the teacher from the taught.

A third principle was to encourage participants to take a broad epistemic standpoint. The progression suggests that expert and competent learners value learning that increases their expertise or practical wisdom in a domain or profession, not just their grasp of generalized transferable knowledge. This suggests that learners are best supported when specific cognitive or learning objectives are situated in

the wider frame of expertise in the domain or profession. Development by learners of their own practice can then be explored with others, each bringing different backgrounds and experiences to the task.

A fourth principle was to support participants' metacognition about how to learn in a MOOC. Self-regulating skills and other metacognitive skills are learnable, responsive to instruction and modelling (Zimmerman, 2002; Butler & Winnie, 1995). The capability to use the open-scaled environments for learning purposes can similarly be taught. The C-SL capability progression provides a template suggesting areas of development by learners. At a minimum, for example, it might be used to make explicit for participants the skills and attitudes required to learn effectively in a MOOC.

These principles are consistent with formulations developed by other MOOC scholars who emphasize self-regulation (Littlejohn & Milligan, 2015; Milligan, Littlejohn, & Ukadike 2015). Their principles for design of MOOCs were derived from their extensive practice in professional learning and are broadly resonant with the theoretically and empirically derived schema proposed here.

5 A FIELD-TRIAL FOR C-SL DESIGN ORIENTATION

The four design principles outlined above were subject to an empirical test on the occasion of the second offering of the ATC21S MOOC, in a form of field trial. An evaluation of the first offering of the MOOC (Milligan & Griffin, 2015) was conducted using benchmark data from other University of Melbourne MOOCs, instructor perception, formal feedback through a post-course survey provided by more than 500 self-selected participants, and comment in forums. Strengths and weaknesses were identified. On the positive side, satisfaction, completion, and pass rates were in the normal-to-high range compared with other University of Melbourne MOOCs. Forum participation rates were particularly high — 66% of active visitors compared with 50% and 48% in two other successful University of Melbourne MOOCs (Animal Behaviour and Epigenetics). ATC21S MOOC forum users posted an average of 1.8 posts each compared to averages of 0.8 and 0.9 in the other two MOOCs.

Areas for improvement were also identified. Spot checks of major assignments and a review of forum posts reinforced the impression that a proportion of participants achieved a pass without developing requisite understanding or skills. Conversely, some participants with a good grasp failed. Peer- and self-evaluations were not reliable. A small but significant 3.5% of those who submitted the final major assignment failed to undertake evaluation of their peers' work. About 5% consistently gave themselves full marks despite low rating by their peers. About 1.7% consistently gave everyone they graded full marks, regardless of quality of the performance. There was some plagiarism, probably rare. The consistency between various graders, as measured by correlations, was low. Correlations between peer-assessment and the respective self-assessment scores were 0.27 and 0.30 respectively for the two assignments. Forums were used by only 66% of active participants, and were judged by participants as difficult to navigate. Although there were more than 1,117 forum threads in ATC21S, less than 50 generated 40 or more posts, and there was a high proportion of dangling posts and sluggish threads, leading to the conclusion that few posts were part of generative extended dialogue.

In this context, the four design principles derived from the C-SL progression were applied to guide adjustments to the design of the ATC21S MOOC, within the bounds supported by the Coursera platform. The specific adjustments are outlined in Table 3. One change aimed to orient the epistemic standpoint of learners more towards development of "practical wisdom" in the domain, and not just intellectual understanding of course themes. This was undertaken by re-articulation of the course learning objectives as a set of developmental progressions relating to teaching practice. The progressions constituted a progress map for the course (https://amstandards.files.wordpress.com/2016/03/atc21s-mooc_progress-map_wk6jla.pdf). The major assignment and quizzes were targeted more closely to questions of professional practice. To encourage self-regulation, each element of the progress map was referenced to particular formative or summative assessments, and opportunities were provided for learners to self-assess against the map. To further support self-regulation, the full scope of the course was supported with various forms of automated feedback on performance (not a small task), and the quality of rubrics for peer-assessment was improved. To better support metacognitive understanding of required learning skills, hints were provided in course messaging to guide learning methods, and a separate resource site (http://crowdsourcedlearning.org/) was also built to which learners were referred to assist their evaluation of their own level of C-SL capability.

Table 3: Design adjustments between Version 1 and Version 2 of ATC21S MOOC.

Design principle 1: Maximize scale and diversity
• No change in design. Coursera MOOC platform attracts a participant base with both scale and diversity
Design principle 2: Scaffold activities to generate and support self-regulation, crowd-sourced learning, reciprocal teaching, and use of automated teaching agents
• Redesigned automated assessments as extension activities, exploring application and synthesis rather than recall and understanding of concepts covered in video material
• Doubled the number of automated quizzes and quiz questions, to cover most aspects of the course, allowing recursiveness, focus, and critical consumption
• Targeted new quiz exercises to areas of confusion identified in first running
• Trimmed videos to reduce viewing time overall to encourage time commitment to production and engagement rather than consumption
• Designed quiz exercises to clarify for participants the professional standards inherent in the major assignment, providing practice on using the rubrics provided for peer- and self-assessment
Design principle 3: Encourage broad epistemic standpoint
• Converted list of learning objectives to a progress map
• Provided visualization of growth in expertise in targeted learning outcomes (as in progress map above
• Linked assessments (formative and summative) to progress map
• Provided non-graded self-assessment quizzes to assist monitoring of progress against the progress map
• Provide benchmarks of class performance against progress map
• Redesigned peer-assessed assignment to improve focus on teaching practice
Design principle 4: Support participants' metacognition on how to learn in a MOOC
• Messaged through weekly emails about purposes of forums, encouraging dialogue and reciprocity, risk-taking and perspective taking, and production
• Provided a resource site that included description of expert behaviour and self-assessment tools.

6   FIELD TRIAL RESULTS

Like many initiatives in education aimed at improved efficacy, the re-design of ATC21S was not conducted as a controlled experiment. Course assessments were varied substantially so it was not possible to provide direct comparison of learning outcomes in the two versions of the MOOC. Nor, pending automation of the process, was it possible to compare C-SL performance between the two cohorts. However, a range of common indicators of student experience was available, enabling direct comparisons of take-up rate, engagement rate, reliability of assessments and self-assessments, and estimates of learner success and satisfaction. The latter were derived from common post-course surveys volunteered by 591 and 293 participants respectively in each version of the MOOC, constituting the most committed of the visitors in each (4.7% and 5.2% respectively).

The characteristics of the two cohorts were remarkably consistent. Both were small cohorts by University of Melbourne standards (18,000+ and 14,000+ registrants for the first and second offerings respectively). Both had a preponderance of female participants (54% and 53%). Both enrolled high proportions of participants from the USA (16% and 14%), from Australia (10% and 7%), and with higher degree qualifications (57% and 59% respectively).

A "validity argument" approach was applied to test the efficacy of the design decisions (Messick, 1995; Kane, 2013). This involved selecting a range of indicators directly comparing student experience in the two versions, and predicting which way such indictors *should* move if the design changes were efficacious. Each of these predictions therefore provided an opportunity to test the validity of the argument that the design changes worked as hypothesized. None of the tests could be used to *prove* the effect of the design, even if they were in the predicted direction; but if in the wrong direction, they could be used to *disprove* it. Taken together, they provided the basis of an argument to support, or not, efficacy of the design changes.

The discussion below, therefore, predicts and tests the direction of the change in each set of indicators from the first to the second versions, thereby providing evidence for or against an argument as to the efficacy of the design changes.

6.1   Registrations and Take-up Rates

For the purposes of evaluating the design changes in the ACT21S MOOC, take-up rates were disregarded as evidence on the grounds that the changes were aimed not at improved take-up but at improved engagement after take-up. For the record, however, the overall take-up rate of the second version (proportion who registered, entered, and became active) was 37% lower than for the first version. The two other University of Melbourne MOOCs used for benchmarking (Animal Behaviour and Epigenetics) also showed reductions in take-up rates between the first and second versions of 37% and 38% respectively.

6.2 Engagement Rates

Predictions as to the effect of design changes on participant engagement with various elements of the MOOC were made and tested as per Table 4. It was predicted that making quizzes central to the instructional process would lift engagement levels, and this prediction was supported. Engagement with forums, as measured by forum participation and number of posts per candidate, was expected to improve, although as the support provided was through messaging, the effect was expected to be small. This prediction was supported. Engagement with videos was expected to show little if any change as it was intended to reduce reliance on videos without reducing their attractiveness. In fact, engagement rates with videos marginally decreased, a finding that is interesting, perhaps supporting the idea that video presentation (a consumption activity, of lower order on the learning capability) becomes less important when there are more production opportunities. Predictions on candidature rates and pass rates were difficult to make as the aims were to increase the proportion of participants who attained high standards of performance while also improving peer-assessment capability and thus reducing the proportion of low-performance participants attaining pass grades.

Table 4: Engagement rates in various MOOC elements, versions 1 and 2, ATC21S MOOC.

Version 1	Version 2	Statistical significance		Interpretation
		z value	*p value*	
Forum engagement: Percentage of active participants using forums				
66%	69%	3.55	$p < .01$	**Increased significantly (as predicted)
Forum engagement: Posts per candidate				
8.06	8.31	na	na	Increased (as predicted)
Video pull-through rate: Percentage of week 2 video users still active in week 6				
54%	52%	2.07	$p < .04$	*Decreased marginally (not predicted)
Quiz engagement rate: Percentage of active participants using quizzes				
43%	69%	23.8	$p < .00$	**Increased very significantly (as predicted)
Candidature rate: Proportion of active participants doing assessments for grading				
15%	14%	1.46	$p < .14$	No change (no prediction made)
Pass rate of candidates (proportion of candidates passing)				
89%	86%	1.79	$p < .07$	No change (no prediction made)

*Significant difference, $p < .05$; **Significant difference, $p < .01$; na: not available.

6.3 Quality of Assessments

To examine the effects of design change on quality of peer- and self-assessments, correlations between peer- and self-assessment on the major assignments in each running of the MOOC were compared. It was predicted that the correlation would increase if the design changes were efficacious. Pearson's correlation coefficient between self-assessed and peer grades awarded in the major assignments increased from 0.25 in the first running to 0.32, suggesting that design changes were efficacious.

6.4   Satisfaction Levels

It was predicted that satisfaction levels between the first and second versions should increase overall, especially in relation to use of quizzes, forums, and videos. No design changes were made to social media, and therefore predictions were made that there would be no increase or decrease in satisfaction with these features. Table 5 shows that in every case the changes were as predicted, encouraging the presumption that the design changes were efficacious. Table 5 also shows positive response to the new elements of design, the progress map and its linked self-assessments, which rated 2.32 and 2.50 (out of 3) respectively.

This evidence suggested that strategies based on the four design orientations had the effect of improving the experience of learners in the MOOC. The strategies appeared to have been efficacious, improving engagement with most elements of the course, improving the quality of peer- and self-assessment, and improving satisfaction, while retaining candidacy and pass rates. On this basis it can be argued that the efficacy of the design changes was supported, pending further investigation and evidence. Clearly, a more thorough test of the design principles would employ more and better indicators of success. In particular, in any future re-runs of the MOOC, changes should allow direct comparisons of the quality of learning outcomes in the domain (as represented in course assessments), and to allow exploration of the distribution patterns of C-SL capability.

Table 5: Comparison of mean satisfaction ratings, versions 1 and 2, ATC21S MOOC.

Rating	Version 1 $N = 591$	Version 2 $N = 293$	Significance *t value p value*		Interpretation
Rating scale 1,2,3,4,5 where 5 = very satisfied					
Overall satisfaction	4.25	4.40	2.42	P < .02*	Improved (as predicted)
Rating scale 1,2,3 where 3 = satisfied					
Videos	2.47	2.77	6.59	$p < .01$**	Improved (as predicted)
Quizzes	2.48	2.75	6.11	$p < .01$**	Improved (as predicted)
Assignments	2.49	2.67	3.51	$p < .01$**	Improved (as predicted)
Extra resources	2.56	2.60	4.56	$p < .01$**	Improved (unexpectedly)
Self-assessments	na	2.50			*new design feature*
Reflective questions	2.39	2.43	0.88	$p = .38$	No change (as predicted)
Progress map	na	2.32			*new design feature*
Forums	2.09	2.19	1.99	$p < .05$*	Marginal improvement (as predicted)
Social media	1.70	1.75	1.37	$p = 0.17$	No change: remained dissatisfied (as predicted)

*Significant difference, $p < .05$; ** Significant difference, $p < .01$.

7   CONCLUSION

It seems likely that there is explanatory value in viewing MOOC learners as having different levels of capability to generate higher order learning. The empirically verified developmental progression for C-SL capability suggested a continuum of expertise from novice to expert along which MOOC learners can be ranged. Each stage is distinguishable by a complex but coherent constellation of attitudes, values, knowledge, skills, and beliefs about learning. Learners at different levels differ in the epistemic standpoint they bring to learning, their understanding of the nature of learning and the source of teaching, and who has responsibility for regulation and management of learning. Those deemed to be expert have a broad conception of what constitutes valued knowledge, are equipped to self-regulate, and have the skills to harness the wisdom of the crowd and to assume reciprocal teaching responsibilities.

A further conclusion is that measurement models can be applied to log stream data in MOOCs to assess this capability for each learner. The detailed traces of activity left in log streams of MOOCs by each participant as they interacted with videos, quizzes, resources, and peer- and self-assessments were able to be used to position a learner on a developmental continuum of learning expertise, as outlined in Figure 1. A process of coding and mapping empirically verified behaviours onto the theoretical progression provided the basis for an interpretable, reliable, and arguably valid assessment of each learner's capability to generate higher order learning in a MOOC.

A third conclusion was that few MOOC learners demonstrate expert learning behaviour; most perform at a level at which the potential of scaled networked learning is scarcely tapped.

Fourth, the C-SL developmental continuum appeared to provide a systematic framework for improving decisions about the design of MOOCs. The four design principles aimed at enhancing opportunities for self-regulation, better scaffolding the adoption of a broad epistemic standpoint by learners, and better supporting the metacognitive understandings of learners about how to harness the crowd and exploit reciprocal teaching opportunities. Preliminary evidence suggests some efficacy of this approach, at least in one MOOC.

Further research and development of C-SL measures will focus on better definition of the nature of learner skills in peer- and self-evaluation, and on automation of assessment of C-SL capability in MOOCs. Work will continue to explore how to adjust pedagogical practices to improve the efficacy of MOOCs, which may well suggest changes in attitudes and practices appropriate for mainstream learning environments as well.

REFERENCES

Ahn, J., Butler, B. S., Alam, A., & Webster, S. A. (2013). Learner participation and engagement in open online courses: Insights from the Peer 2 Peer University. *Journal of Online Learning & Teaching, 9*(2), 160–171. Retrieved from http://jolt.merlot.org/vol9no2/ahn_0613.htm

Bernacki, M. L., Aguilar, A. C., & Byrnes, J. P. (2011). Self-regulated learning and technology-enhanced learning environments: An opportunity–propensity analysis. In G. Dettori & D. Persico (Eds.), *Fostering self-regulated learning through ICT* (pp. 1–26). Hershey, PA: IGI Global.

Biggs, J. B., & Tang, C. (2011). *Teaching for quality learning at university* (4th ed.). Maidenhead, UK: Open University Press.

Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research and Practice in Assessment, 8*, 13–15. Retrieved from http://www.rpajournal.com/studying-learning-in-the-worldwide-classroom-research-into-edxs-first-mooc/

Buckingham Shum, S., & Deakin Crick, R. D. (2012). Learning dispositions and transferable competences: Pedagogy, modelling and learning analytics. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (LAK '12), 92–101. http://dx.doi.org/10.1145/2330601.2330629

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281. http://dx.doi.org/10.3102/00346543065003245

Creelman, A. (2013, September). The silent majority: Why are MOOC forums so counterproductive? [Web log post]. Retrieved from http://acreelman.blogspot.com.au/2013/09/the-silent-majority-why-are-mooc-forums.html

Cronbach, L., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302. http://dx.doi.org/10.1037/h0040957

Daniel, J. (2012). Making sense of MOOCs: Musings in a maze of myth, paradox and possibility. *.Journal of Interactive Media in Education, 2012(3)*. http://dx.doi.org/10.5334/2012-18

Deakin Crick, R., Broadfoot, P., & Claxton, G. (2004). Developing an effective lifelong learning inventory: The ELLI project. *Assessment in Education: Principles, Policy & Practice, 11*(3), 247–272. http://dx.doi.org/10.1080/0969594042000304582

Deakin Crick, R., Stringher, C., & Ren, K. (2014). *Learning to learn: International perspectives from theory and practice*. London/New York: Routledge.

DeBoer, J., Ho, A. D., Stump, G. S., & Breslow, L. (2014). Changing "course": Reconceptualising educational variables for massive open online courses. *Educational Researcher, 43*(2), 74–84. http://dx.doi.org/10.3102/0013189X14523038

Dreyfus, H. L. (2002). Anonymity versus commitment: The dangers of education on the Internet. *Educational Philosophy and Theory, 34*(4), 369–378. http://dx.doi.org/10.1023/A:1010010325208

Dreyfus, H. L. (2008). *On the Internet* (2nd ed.). Hoboken, NJ: Taylor & Francis.

Dreyfus, S. E., & Dreyfus, H. L. (1980). A five stage model of the mental activities involved in directed skill acquisition. Berkley, CA: Operations Research Centre, University of California.

Dringus, L. P. (2012). Learning analytics considered harmful. *Journal of Asynchronous Learning Networks, 16*(3), 87–100.

Dron, J., & Anderson, T. (2014). *Teaching crowds: Learning and social media*. Edmonton, AB: Athabasca University Press.

Fournier, H., & Kop, R. (2010). New dimensions of self-directed learning in an open networked learning environment. *International Journal of Self-Directed Learning, 7*(2), 35–55.

Fournier, H., Kop, R., & Sitlia, H. (2011). The value of learning analytics to networked learning on a personal learning environment. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (LAK '11), 104–109. http://dx.doi.org/10.1145/2090116.2090131

Gillani, N., Yasseri, T., Eynon, R., & Hjorth, I. (2014). Structural limitations of learning in a crowd: Communication vulnerability and information diffusion in MOOCs. *Scientific Reports, 4,* [Article 6447]. http://dx.doi.org/10.1038/srep06447

Greller, W., & Draschler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Educational Technology and Society, 15*(3), 42–47.

Griffin, P., McGaw, B., & Care, E. (Eds.). (2012). *Assessment and Teaching of 21st Century Skills* (Vol. 1) Dordrecht, The Netherlands: Springer.

Griffin, P., & Care, E. (2015). *Assessment and teaching of 21st century skills: Methods and approaches* (Vol. 2). Dordrecht, The Netherlands: Springer.

Gunnarsson, B. L., & Alterman, R. (2013). Understanding promotions in a case study of student blogging. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (LAK '13), 57–65. http://dx.doi.org/10.1145/2460296.2460309

Gunnarsson, B. L., & Alterman, R. (2014). Peer promotions as a method to identify quality content. *Journal of Learning Analytics, 1*(2), 126–150. Retrieved from https://epress.lib.uts.edu.au/journals/index.php/JLA/article/view/3539

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. http://dx.doi.org/10.3102/003465430298487

Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In P. Griffin & E. Care (Eds.), *Assessment and teaching for 21st century skills: Educational assessment in the information age* (Vol. 1, pp. 37–56). Dordrecht, The Netherlands: Springer.

Huang, J., Dasgupta, A., Ghosh, A., Manning, J., & Sanders, M. (2014). Superposter behaviour in MOOC forums. *Proceedings of the 1st ACM conference on Learning@Scale* (L@S 2014), 117–126. http://dx.doi.org/10.1145/2556325.2566249

Jenkins, H. (2009). *Confronting the challenge of participatory culture*. Cambridge, MA: MIT Press.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. http://dx.doi.org/10.1111/jedm.12000

Kizilcec, R., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (LAK '13), 170–179. http://dx.doi.org/10.1145/2460296.2460330

Kop, R. (2011). The challenges to connectivist learning on open online networks: Learning experiences during a massive open online course. *International Review of Research in Open and Distance Learning, 12*(3), 19–38. Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/882

Littlejohn, A., Beetham, H., & McGill, L. (2012). Learning at the digital frontier: A review of digital literacies in theory and practice. *Journal of Computer Assisted Learning, 28*, 547–566. http://dx.doi.org/10.1111/j.1365-2729.2011.00474.x

Littlejohn, A., & Milligan, C. (2015, June). Designing MOOCs for professional learners: Tools and patterns to encourage self-regulated learning. *eLearning Papers, 42*. Retrieved from http://www.openeducationeuropa.eu/sites/default/files/asset/Design_Patterns_for_Open_Onlin e_Teaching_and_Learning_Design_Paper_42_4.pdf

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transaction, 12(*2), 878. Retrieved from http://www.rasch.org/rmt/rmt162f.htm

Luntley, M. (2009). Understanding expertise. *Journal of Applied Philosophy, 26*(4), 356–370. http://dx.doi.org/10.1111/j.1468-5930.2009.00468.x

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrica, 47*, 149–174. http://dx.doi.org/10.1007/BF02296272

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, Winter*, 5–8. http://dx.doi.org/10.1111/j.1745-3992.1995.tb00881.x

Milligan, C., Littlejohn, A., & Ukadike, O. (2014). *Professional learning in massive open online courses: Design recommendations.* In S. Bayne, C. Jones, M. de Laat, T. Ryberg, C. Sinclair (Eds.), *Proceedings of the 9the International Conference on Networked Learning (*pp.368— 371*).* Edinburgh, Scotland. Retrieved from http://www.lancaster.ac.uk/fss/organisations/netlc/past/nlc2014/abstracts/pdf/milligan.pdf

Milligan, C., Littlejohn, A., & Margaryan, A. (2013). Patterns of engagement in connectivist MOOCs. *Journal of Online Learning & Teaching, 9*(2), 149–159. Retrieved from http://jolt.merlot.org/vol9no2/milligan_0613.htm

Milligan, S. K. (2015). Crowd-sourced learning in MOOCs: Learning analytics meets measurement theory. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 151–155. http://dx.doi.org/10.1145/2723576.2723596

Milligan, S. K., & Griffin, P. (2015). Mining a MOOC: What our MOOC taught us about professional learning, teaching and assessment. In E. McKay & J. Lenarcic (Eds.), *Macro-level learning through Massive Open, Online Courses: Strategies and predictions for the future.* Hershey, PA: IGI Global.

OECD. (2014). *PISA 2012 Technical Report*. Retrieved from https://www.oecd.org/pisa/pisaproducts/pisa2012technicalreport.htm

Price, M., Handley, K., & Millar, J. (2011). Feedback: Focusing attention on engagement. *Studies in Higher Education, 36*(8), 879–896. http://dx.doi.org/10.1080/03075079.2010.483513

Prins, F. J., Sluijsmans, D. M. A., Kirschner, P. A., & Strijbos, J.-W. (2005). Formative peer assessment in a CSCL environment: A case study. *Assessment and Evaluation in Higher Education, 30*(4), 417–444. http://dx.doi.org/10.1080/02602930500099219

Sadler, D. R. (2010). Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education, 35*(5), 535–550. http://dx.doi.org/10.1080/02602930903541015

Shute, V., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. Cambridge, MA: The MIT Press.

Siemens, G. (2004, December 12). *Connectivism: A learning theory for the digital age*. [Web log post]. Retrieved from http://www.elearnspace.org/Articles/connectivism.htm

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review, 46*(5), 30–32.

Stewart, B. (2010). *Social media literacies and perceptions of value in open online courses*. Unpublished manuscript. Retrieved from http://portfolio.cribchronicles.com/social-media-literacies-and-perceptions-of-value-in-open-online-courses/

Stewart, B. (2013). Massiveness + openness = new literacies of participation? *MERLOT Journal of Online Learning and Teaching, 9*(2), 10. Retrieved from http://jolt.merlot.org/vol9no2/stewart_bonnie_0613.htm

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations.* New York: Doubleday.

University of Melbourne. (2014.) *Massive open online courses data snapshots 2013–2014 (*Report). Retrieved from https://le.unimelb.edu.au/wp-content/uploads/2015/03/UOM-MOOC-report.pdf

Veletsianos, G. (2013). *Learner Experiences with MOOCs and Open Online Learning*. Hybrid Pedagogy. Retrieved from http://learnerexperiences.hybridpedagogy.com/

Veletsianos, G., Collier, A., & Schneider, E. (2015). Digging deeper into learner's experiences in MOOCs: Participation in social networks outside of MOOCs, notetaking and contexts surrounding consumption. *British Journal of Educational Technology, 46*(3), 570–587. http://dx.doi.org/10.1111/bjet.12297

Waite, M., Mackness, J., Roberts, G., & Lovegrove, E. (2013). Liminal participants and skilled orienteers: Learner participation in a MOOC for new lecturers. *MERLOT Journal of Online Learning and Teaching, 9*(2), 200–213. Retrieved from http://jolt.merlot.org/vol9no2/waite_0613

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. New York: Taylor & Francis.

Wilson, M., & Scalise, K. (2012). Assessment of learning in digital networks. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 37–56). Dordrecht, The Netherlands: Springer.

Wolfe, E. W., & Smith, E. V. J. (2007). Instrument development tools and activities for measure validation using Rasch models: Part 1 & Part 2. In E. V. Smith, Jr. & R. M. Smith (Eds.), *Rasch measurement: Advanced and specialized* (pp. 203–273). Maple Grove, MN: JAM Press.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: Mesa Press.

Wu, M., Adams, R., & Wilson, M. (1998). *ACER conquest: Generalised item response modelling.* Melbourne, Australia: ACER Press.

Yeager, C., Hurley-Dasgupta, B., & Bliss, C. A. (2013). cMOOCs and global learning: An authentic alternative. *Journal of Asynchronous Learning Networks, 17*(2), 133–147.

Zimmerman, B. (2002). Becoming a self-regulated learner. *Theory into Practice, 4*(2), 6. http://dx.doi.org/10.1207/s15430421tip4102_2

TECHNICAL APPENDIX:

FIT OF C-SL MEASURE TO A PARTIAL CREDIT MODEL IN TWO MOOCS

The measurement model used in this study was the partial credit model. Separate tests of fit to the model were conducted for each of two groups of participants, one in the Macro MOOC and one the ATC21S MOOC, who participated in most elements of each MOOC. Then modelling was repeated on a conjoined sample constituted from both groups. The indicators of fit are summarized below. In this table:

Infit MSQ refers to the mean square of differences between the estimated and observed difficulties of the items (Wolfe & Smith, 2007) weighted to take more account of responses closer to the mean. A MSQ of 1.3 suggests that the item has 30% more "noise" than expected by the model. Linacre (2002) suggested that an MSQ up to 1.5 is productive for measurement, while between 1.5 and 2 is unproductive but not degrading.

- Item separation reliability is essentially a correlation coefficient of the variance of the estimated measure and the observed measures
- Item-total correlations should be greater than zero
- Person ability refers to the position of a person on the scale, expressed in the standard unit of the logit
- Item difficulty refers to the position of the threshold on the same scale

Type of fit	Fit indicators MACRO MOOC 60 items, 80 categories *n= 3,320*	Fit indictors ATC21S MOOC 60 items, 80 categories *n= 4,438*	Fit indictors in both MOOCs conjoined 60 items, 80 categories *n= 7,758*
Item parameter Infit MSQ	Mean 0.98, variance 0.03. Each of the 36 relevant items and 54 item*step parameters MSQs in the range 0.7 to 1.3	Mean 1.02, variance .02. Each of the 32 relevant items and 51 item*step parameters MSQs in the range 0.7 to 1.3	Mean 1.02; variance: 0.03; Each of the 32 relevant items and 51 item*step parameters MSQs in the range up to 1.3, except for four items with MSQ of 1.34, 1.37, 1.48 and 1.45
Reliability measures	Item Separation reliability: 0.999	Item Separation reliability: 0.999	Item Separation reliability: 1.0
Item-total correlations	Range between 0.41 to 0.81	Range between 0.28 to 0.81	Range between: 0.24 to 0.84
Item vs. person parameters	Mean item difficulty was 0 logits, variance 1.38; Mean person ability was -3.58 logits, variance 4.29 logits, maximum 5.5 logits, minimum -6.22 logits	Mean item difficulty was 0 logits, variance 2.42; Mean person ability was -2.11 logits, variance 3.4 logits; maximum 5.6 logits, minimum -7.1 logits	Mean item difficulty was 0 logits; variance 4.69 logits; Mean person ability: -1.79 logits, variance 6.13, logits; max: 6.7 logits; min: -7.95 logits