11

CHAPTER

Educational Assessment in a MOOC: The Case of Statistics One

Andrew R. A. Conway
Princeton University

David Moreau
Princeton University

LaTasha Holden
Princeton University

INTRODUCTION

Consistent with the theme of this book, we start from the perspective of Badrul Khan's eight-dimensional e-learning framework. In this chapter we focus on the Evaluation dimension, which has multiple components, for example, assessment of student learning, assessment of overall course quality, assessment of instructor ability, and more. As well, a wide range of assessments can be found among the rapidly expanding library of Massive Open Online Courses (MOOCs). It is therefore difficult to adopt a broad approach to Evaluation when it comes to the Evaluation of MOOCs.

Instead, the current chapter will focus on assessment of student learning, and more specifically, learning in an introductory statistics course. The lead author is the instructor of *Statistics One*, which is a MOOC offered on *Coursera*. *Statistics One* is designed to be a basic introduction to statistics. Based on retention, student activity, and feedback in the discussion forums, it is fair to conclude that *Statistics One* is a successful MOOC. The production quality of the course is exceptional, the content is applicable to a wide audience, lectures are sequenced in a meaningful fashion, and weekly assignments reinforce concepts presented in lecture.

That being said, one component of *Statistics One* that could be better is assessment of student learning. Because we believe this is an aspect of a course that can always be improved, we analyzed the reliability and validity of the assessments offered in *Statistics One*. Analyses of these data are reported later in the chapter. First, we begin with a detailed description of the course and our personal

experience as instructors; next we provide an overview of Psychometrics and how to assess the reliability and validity of educational assessment tools. We then provide some data from our own course. Finally, we consider limitations of this iteration of the course and consider potential ways to improve the course for the next iteration, as well as some general advice for anyone entertaining the idea of teaching a MOOC.

At the outset we should note that our philosophical approach to assessment is objective, empirical, and multi-faceted. That is, our assessments, such as assignments and exams consist of objective questions that have clear correct and incorrect answers. As well, our approach to the evaluation of the assessments is empirical and rooted in measurement theory, which we describe below. Finally, we used multiple assessment tools because we argue that no one approach is free from measurement error and/or bias.

It is also important to note here two features of the course that impact the way in which assessments may be interpreted. First, *Statistics One* does not offer a *Certificate of Completion* or a *Statement of Accomplishment* to students. This is not the case for most *Coursera* courses but it is the policy for all Princeton University courses offered on *Coursera* (as of summer, 2014). What makes this unique, and important, in the context of assessment is that from a student's perspective assessments are criteria-referenced but from the instructor's perspective they can be viewed as norm-referenced. In other words, the student's approach to the assessments, and to the course in general, is to master the material, that is, to reach a certain criteria. However, as instructors, we have the option to analyze student performance on a continuous distribution and, theoretically, assign grades to students in a normative fashion.

COURSE DESCRIPTION

Statistics One was offered for the second time on *Coursera* in the Fall of 2013, one year after its first iteration. It was designed to be a comprehensive introduction to fundamental concepts in statistics, providing a solid foundation for students planning to pursue more advanced courses in statistics. The course assumed very little background knowledge in statistics and a working knowledge of basic algebra. In addition, the course provided an introduction to the R programming language, with examples and assignments that involved interpreting R output and writing code in R (R Development Core Team, 2013). Various incentives justified choosing R as a support statistical language, including the fact that it can be downloaded for free, and it is compatible with most operating systems, each of which are aspects that are well aligned with *Coursera's* philosophy to offer free course content worldwide.

The structure of the 12-week course was as follows. Each week, two lecture videos were released, along with the lecture slides in PDF format. The lecture videos also contained embedded quizzes, often referred to as "in-video quizzes", a signature feature of many *Coursera* courses. The lectures were accompanied by a lab video, describing how to do the type of analyses described in the lectures in a practical R example. Finally, students were instructed to complete a "homework" assignment based on both the lecture and lab contents of the week. Lectures were recorded at the Princeton University Broadcast Center, a professional studio designed for quality film shooting. Labs were recorded using the software package ScreenFlow.

There was also a midterm exam and a final exam. There was not a homework assignment the week of the midterm exam so there were 11 total assignments. So in total we administered 13 assess-

ments of student learning. These 13 assessments will be the focus of our analysis later in the chapter. In our opinion, this is a small number of assessments for a 12-week course but given the fact that *Statistics One* does not offer course credit, it seemed like a sufficient number to gauge student progress.

Student Demographics. A total of 144,950 students enrolled in the course. However, that number is misleading because only 67,497 students (47%) watched the first lecture video, a figure in line with prior research studies on MOOCs showing that approximately 50% of the students who initially enroll do not return to the course (Koller et al, 2013). There was a male to female ratio of two to one, similar to other *Coursera* courses (see Figure 1).

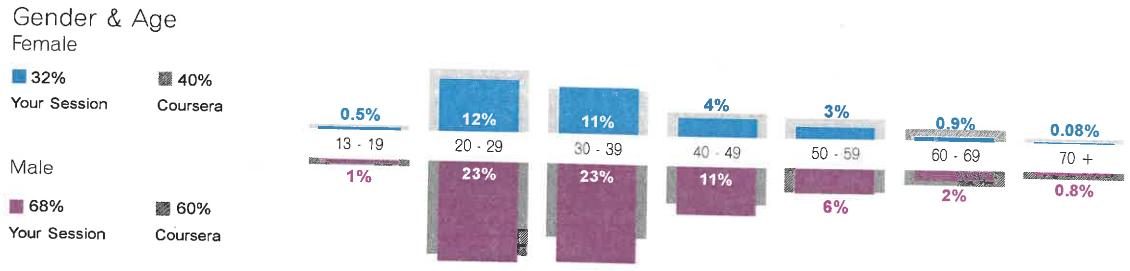

Figure 1. Student gender and age in *Statistics One*, compared to *Coursera's* average, as of Summer 2014.

It has been argued that MOOCs provide access to quality course content to students across the globe. Data from our course partly support this view, with students from various geographical locations, but in uneven proportions. A majority of students enrolled from North America (40%), but a substantial amount also participated from Europe (27%) and Asia (23%). A minority of students enrolled from other continents. A closer look showed that most students came from the United States (35%), followed by India (10%) and the United Kingdom (5%). For a detailed view, see Figures 2 and 3.

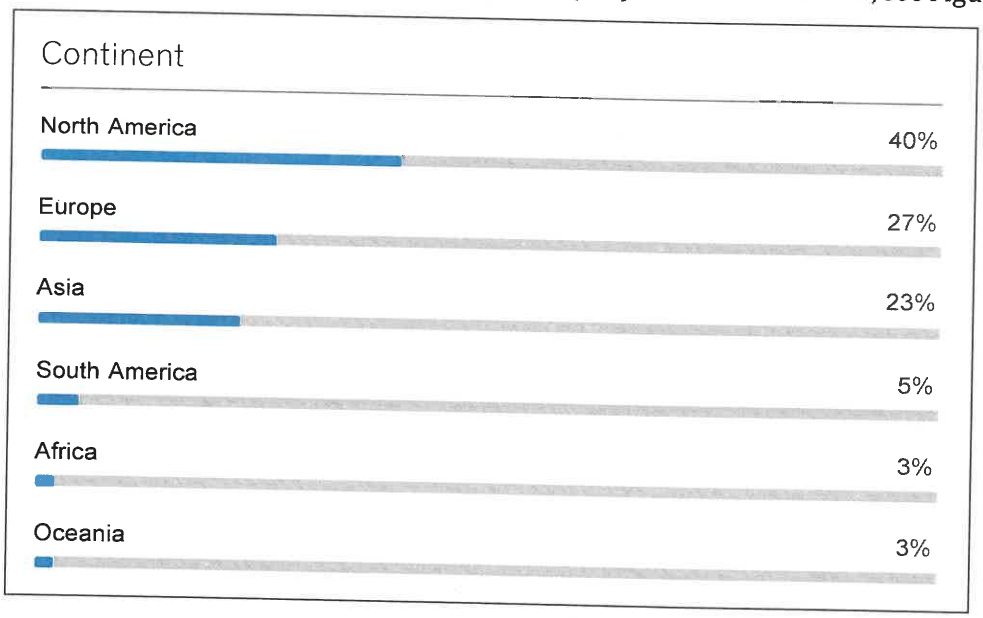

Figure 2. Student location grouped by continent.

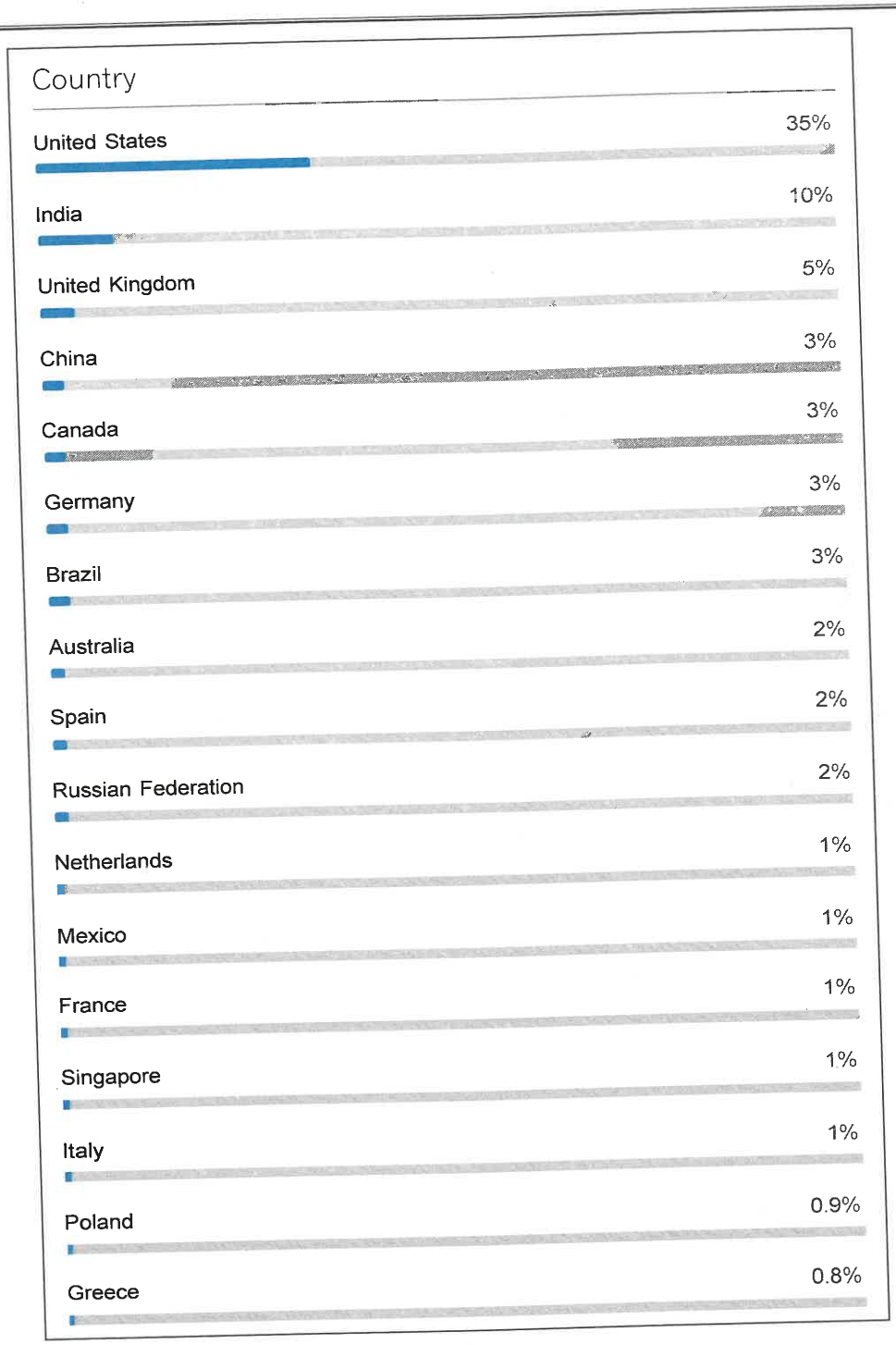

Figure 3. Student location grouped by nation.

Student Activity. At the onset of the course, most students intended to complete all the assignments and to watch all the lectures (see Figure 4). However, a closer look at students actual behavior shows that many visited the course and watched one or a few lectures, but only a small percentage submitted assignments or used the forums (for a more detailed view of student activity, see Figures 5 and 6).

This seems to be typical of the consumer behavior of many students in MOOCs. The possibility to access course content for free, without any negative consequences for passive students, probably exacerbate this behavior. In addition, *Coursera* does not allow viewers to access content without first enrolling in a course, whereas some other MOOC providers do so.

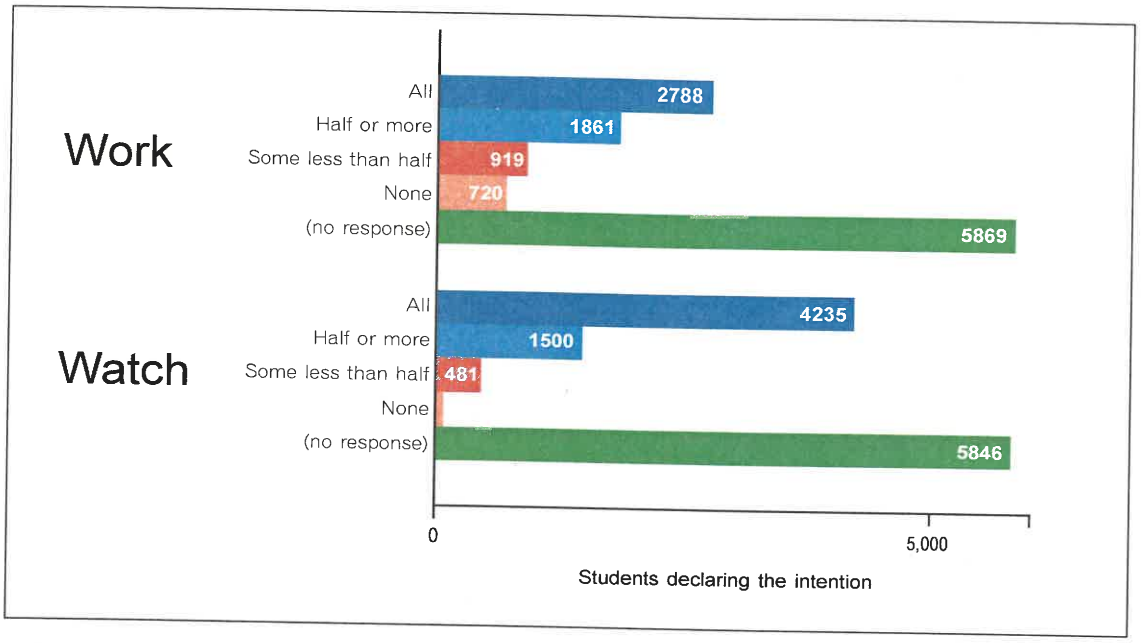

Figure 4. Student intentions at the time they enrolled in the course, in terms of submitting assignments ("Work") and watching lectures ("Watch"). Note: This survey was optional, which is why the total number of respondents does not match the total number of students enrolled in the course.

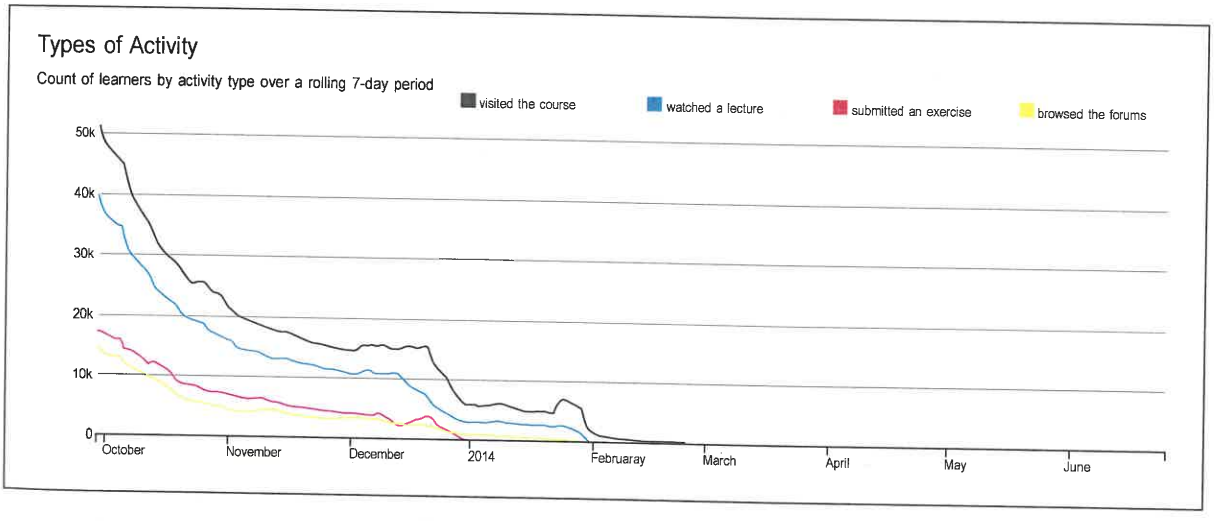

Figure 5. Evolution of the number of students per types of activity throughout the course.

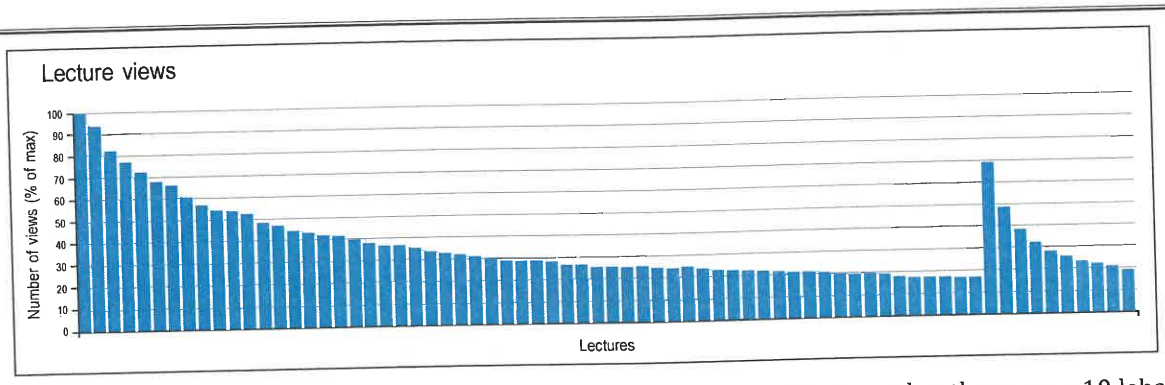

Figure 6. Number of views per lecture video (1 – 24) and lab video (1 – 10). Note that there were 10 labs, not 12, because there was not a lab the week before the midterm exam or the week before the final exam.

Assignments and Exams. The administrator interface on *Coursera* has greatly improved since the first iteration of this course in 2012. However, some aspects could still be upgraded, to facilitate deployment and grading of assignments and exams (see "Limitations" section below). Creating assignment and exam questions on the *Coursera* platform was a bit daunting at first. Each question had to be edited separately, rather than all at once, for example by uploading source code. This seems to be something *Coursera* is currently working on and should improve in the near future. Similarly, in order to provide feedback, every possible answer to a question that we could imagine had to be computed and uploaded; this means that all possible answers needed to be identified and entered in the system. This was especially time-consuming given that most answers in our course required writing R code, which can be expressed in various ways. To be blunt, creating each assignment and exam was tedious and very time-consuming. Moreover, if we, as instructors, failed to upload a form of an answer that could be considered correct then the process was extremely frustrating to students. For example, if the correct answer to a question was "50%" and a student entered "0.50" the system would score that answer as incorrect unless we uploaded "0.50" as a candidate for correct answers (in this example, others would include ".50", ",50", "50/100", and so on). It was difficult for us to anticipate the variety of forms a correct answer might take and this in turn made the experience frustrating to students. We learned the hard way with the first assignment and unfortunately we received plenty of negative feedback about it. Indeed, an inspection of assignment activity suggests that we lost many students after assignment 1 (see Figure 7).

Title	# of Users Submitted	Recent Activity	Raw Score Histogram	Min	Max	Avg	Std
Week 1	26458			0.00	10.00	6.64	2.46
Week 2	17414			0.00	10.00	7.46	2.43
Week 3	13110			0.00	10.00	7.36	2.33
Week 4	9657			0.00	10.00	7.36	2.59
Week 5	8005			0.00	10.00	6.69	2.65
Week 6	7123			0.00	10.00	6.69	2.92
Week 7	6042			0.00	10.00	5.80	2.59
Week 8	5527			0.00	10.00	7.26	2.62
Week 9	5062			0.00	10.00	7.25	2.50
Week 10	4692			0.00	10.00	7.15	2.50
Week 11	4351			0.00	10.00	7.50	2.28
Empty Title	0			0.00	10.00	0.00	0.00
Midterm	6800			0.00	25.00	18.41	6.01
Final	4533			0.00	25.00	20.01	5.35

Figure 7. Students participation in assignments across the course.

Discussion Forums. Students' involvement in the forums was mostly helpful and constructive. Figure 5 offers a glimpse at forum activity throughout the course. Actively monitoring new and "up-voted" threads allowed us to rapidly identify and fix errors or bugs. We should mention here that at the moment *Coursera* still allows anonymous comments on the forums. In our view, this can sometimes have a pernicious effect, with some individuals unleashing harsh comments toward their fellow students, or about course content, or about the instructor and staff. Disabling anonymous comments on the forums could be an easy fix to eradicate this kind of behavior, but unfortunately *Coursera's* interface does not permit this yet. We think this could be improved in the near future, to make everyone's experience more pleasant and constructive.

Retention. As mentioned above, 144,950 enrolled in *Statistics One*. However, only 67,497 students watched the first video (47% of total enrolled). Figure 6 illustrates the number of students who watched each lecture but we also have more detailed data on the number of students who watched each segment of each lecture (all lectures were divided into 2-3 segments). There was a good bit of variance across lectures, with a high of 63,503 viewers for lecture 1, segment 1 and a low of 11,637 viewers for lecture 24, segment 4 (the final segment). The median number of viewers for lecture segments was 20,416. Figure 6 also illustrates the number of students who watched each of the lab videos. These too varied with a high of 46,767 viewers for lab 1 and a low of 13,161 viewers for lab 10 (the final lab). The median number of viewers for labs was 18,462. Figure 7 shows that the number of students who completed the first assignment was 26,458. The number who completed the final assignment was 4,351.

Each of the 11 assignments consisted of 10 questions and each exam consisted of 25 questions. Therefore, the total number of points a student could earn in the course was 160. If we consider 60% "passing", which is a common standard to earn a D and pass a college course, then 4,581 students "passed" *Statistics One*. When considered from the view of initial enrollment (144,950) that number may seem low, indeed, it is approximately 3%. However, considering that this is a course about statistics, and it does not offer a certificate or a statement of accomplishment, and it requires learning challenging software, 4,581 is pretty impressive. Also, if we consider the number of students who "passed", i.e., 4,581 relative to the number who completed the first assignment, then "retention" was 17%. And we think we can improve upon that because as mentioned, assignment 1 was problematic for various reasons. If we consider the number of students who passed relative to the number who completed assignment 2, then "retention" jumps to 26%.

Overall, our experience in developing and implementing *Statistics One* was positive. But we were pioneers on *Coursera*. *Statistics One* was one of the first courses offered on the platform. Most people in higher education have now come to recognize that teaching a MOOC takes much more time than teaching a traditional course. Indeed, that was our experience, but for us, it was worth it.

PSYCHOMETRICS

Psychometrics can be loosely defined as the measurement of the mind. The over-arching goal of psychometrics is to develop and evaluate objective measurements of products of the human mind, such as skills, knowledge, abilities, attitudes, traits, and academic achievement. In the field of Psychology there are two primary disciplines that have been largely shaped by psychometrics: intelligence and personality. The fields of intelligence and personality both involve the creation of consistent (*reliable*) and appropriate (*valid*) measurement tools that produce observable and quantifiable data. In order to

produce quantifiable observations from psychological constructs such as intelligence or personality, one must first develop a mode for assessing the construct, however, it is more essential that the mode is both reproducible and appropriate. To achieve this goal, psychometricians must put their own tools to the test of human unpredictability. If a measurement tool does afford a consistent pattern of results across different random samples of students then it is a likely conclusion that the tool is reliable. In other words, the tool is consistently tapping into some ability or trait. However, in order for any measurement tool to be useful it must also be measuring the real ability or trait that the assessment tool in question claims to assess. In other words, it must be valid.

The degree to which a psychometric tool, for example an IQ test or a personality survey, *consistently* measures an ability or trait over time aligns with its reliability. The degree to which it measures what it is supposed to measure is in accordance with its validity. Hence, reliability and validity are two central concepts in psychometrics and assessment. In the remainder of this section we discuss in more detail reliability and validity, respectively.

Reliability. According to *classical test theory*, also known as *true score theory*, any score on a test or survey, etc., is a reflection of a "true score" and measurement error (Novick, 1966). If we let X = "raw score" then X = T + e, where, T = true score and e = measurement error. As such, the more consistently a person obtains a similar score on a test or a survey, the more reliable it becomes and the less it is likely to be contaminated by error. Moreover, the reliability of measurement maps onto the amount of error variance in a test, which can be expressed as follows:

$$\text{var}(X) = \text{var}(T) + \text{var}(e)$$

Thus, the less error variance, the more a score reflects the nature of a true ability or trait or skill, etc. Again, as measurement error reflects the amount of random error, and random error variance in observations, the extent to which a test reflects systematic error variance, is reflective of measurement error or bias. The value of e, however, is unknown, which is why reliability cannot be calculated exactly and instead must be estimated.

So how can we estimate the reliability of an educational assessment tool? First, let's consider an analogy, for example, the measurement of body temperature. If a nurse uses a thermometer to measure the temperature of a patient twice, one measurement immediately after the other, then the measurements from time 1 to time 2 should be stable. In other words, measure 1 should be correlated with measure 2. The degree to which they are correlated reflects the thermometer's reliability.

It is a bit more difficult to evaluate reliability when dealing with things like intelligence or personality or student learning in a MOOC. As mentioned, these assessments are influenced by measurement error and/or bias. Actually, even something as simple as body temperature is susceptible to measurement error. For example, consider again a nurse trying to measure the body temperature of a patient. Suppose an oral thermometer is used and between measurements the patient takes a sip of ice water. The second measure would be inaccurate. The point here is that measurement is always susceptible to chance error and/or bias.

To recap, any raw score X is actually the true score, plus some bias if it exists, plus some measurement error. As a measure X approaches the true score, it is considered to be more reliable. The problem is that we don't know the true score. Reliability, therefore, must be estimated.

Three primary methods used to estimate reliability are (a) test/re-test, (b) internal consistency, and (c) parallel tests. The test/re-test method involves administration of an assessment twice and examining the correlation across time, as in the body temperature example. This is a common procedure in research settings but it is clearly not appropriate in an educational context (you can't administer the same midterm exam twice). We therefore used methods of internal consistency and parallel tests to evaluate the reliability of our assessments in *Statistics One*. The notion of internal consistency is that items within a test, or assignments within a course, should be stable. In other words, students who perform well on a test item that is designed to measure knowledge of a particular concept should also perform well on other items that are also designed to measure knowledge of the same, or similar, concepts. Likewise, the notion of parallel tests is that two assessments that are designed to measure the same thing should yield consistency.

To preview, we evaluated the reliability of our assignments by calculating Cronbach's alpha, α, across all 11 assignments, a method commonly used in the behavioral sciences. We were not able to create "parallel tests" for each assignment but some assignments are more alike than others. For example, Assignment 3 is more like Assignment 4 than it is to Assignment 8. We should therefore expect a stronger correlation between Assignment 4 and Assignment 5 than between Assignment 4 and Assignment 8.

Validity. To understand the concept of validity in psychometrics it is first essential to clearly define a *construct*. A construct can be thought of as an object or entity that is not directly observable (as opposed to real objects). For example, apples are real objects that can be measured in various ways, such as size, weight, type and so on. They are pretty easy to measure. In the behavioral sciences we are forced to deal with more abstract constructs, for example, intelligence. For over a century researchers have been arguing about what intelligence really means and how it should be measured. This is partly because it is a psychological construct that cannot be observed directly.

To gain traction on this problem, the first step is to operationalize a construct. That is, one must first define the construct but then come up with some way to quantify the construct. For example, in the case of intelligence, researchers made the construct observable and quantifiable using intelligence tests. There is of course a good deal of controversy around how that is done in the area of intelligence research but the point here is that there is a process of defining a construct and then operationalizing it that allows for empirical research and the evaluation of validity.

There are four dimensions of validity: (a) content validity; (b) convergent validity; (c) divergent validity; and (d) nomological validity (Cronbach & Meehl, 1955). To illustrate these four types of validity, consider a simple example that's not very controversial – verbal ability in young children. Verbal ability is a construct and it is a very important construct for academic achievement. How might we operationalize it? One way would be to develop a vocabulary test. That is, a researcher might design a vocabulary test and then administer it to a sample of children. The researcher could then assess if the test is a valid measure of verbal ability. Content validity is very simple; it is just *face value*. That is, teachers could simply examine the test and determine if it is appropriate for the population of children being evaluated. For example, a vocabulary test consisting of German words is not a valid test of verbal ability in children whose native language is English.

The next two dimensions, convergent and divergent validity, are typically evaluated together. In the verbal ability example, the vocabulary test will demonstrate convergent validity to the extent that it correlates with other established measures of the same construct. So suppose the researcher

who designed the new vocabulary test has access to a well-established, reliable and valid, measure of reading comprehension. The researcher could then administer the new vocabulary test as well as the established measure of reading comprehension to a sample of children and examine the correlation between the two measures. If the vocabulary test-scores are correlated with scores on the reading comprehension test then that indicates a degree of convergent validity. In other words, the vocabulary scores are converging on other measures of the construct, verbal ability. However, it is not enough to just have convergent validity, and unfortunately, a lot of researchers stop there. It is critical to also demonstrate divergent validity. To continue with the verbal ability example, the researcher should be able to demonstrate that the vocabulary test scores do not correlate, or correlate to a lesser degree, with some other measure that it theoretically should not correlate with, for example, body temperature.

Finally, there is nomological validity, which is more like "meta validity". The scores on the vocabulary test should be consistent with more general theories that exist in related fields of research. So, for example, if the researcher is investigating verbal ability in children then scores on the vocabulary test should be consistent with what we know about developmental psychology, cognitive psychology, and neuroscience. For instance, a child with neural damage or disease to a particular brain region that is known to be important for the development of verbal ability should score less well on the vocabulary test than a healthy child. That empirical pattern would be consistent with broader theories, or nomological theories, hence the term nomological validity.

EVALUATION OF ASSESSMENTS

In this section we evaluate the reliability and validity of the 13 assessments administered in *Statistics One* (11 weekly assignments and 2 exams). Descriptive statistics for all assessments are provided in Table 1. Correlations among all measures are reported in Table 2.

Table 1. Descriptive statistics for all assessments.

Assessment	Mean	SD	Skew	Kurtosis
Assessment 1	6.65	3.51	-0.58	-1.00
Assessment 2	7.07	3.37	-0.78	-0.66
Assessment 3	7.69	3.10	-1.19	0.20
Assessment 4	7.64	3.21	-1.10	-0.05
Assessment 5	7.66	3.10	-1.14	0.12
Assessment 6	8.20	3.02	-1.54	1.14
Assessment 7	7.24	3.23	-0.82	-0.55
Assessment 8	7.99	2.97	-1.42	0.86
Assessment 9	8.27	2.83	-1.62	1.60
Assessment 10	8.39	2.68	-1.75	2.08
Assessment 11	8.62	2.39	-1.91	2.92
Midterm Exam	20.43	7.65	-1.96	2.47
Final Exam	21.76	5.60	-2.12	4.08

Table 2. Correlations among all assessments

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	Midterm	Final
A1	1.00												
A2	.50	1.00											
A3	.43	.50	1.00										
A4	.34	.40	.52	1.00									
A5	.35	.39	.49	.56	1.00								
A6	.32	.38	.46	.52	.63	1.00							
A7	.31	.33	.36	.43	.52	.51	1.00						
A8	.27	.33	.40	.43	.51	.49	.50	1.00					
A9	.28	.35	.40	.43	.51	.53	.50	.64	1.00				
A10	.26	.31	.37	.42	.47	.51	.48	.53	.63	1.00			
A11	.22	.28	.36	.37	.41	.41	.41	.54	.55	.60	1.00		
Midterm	.29	.35	.45	.46	.54	.56	.45	.47	.47	.44	.39	1.00	
Final	.17	.24	.30	.33	.37	.34	.31	.41	.45	.47	.52	.38	1.00

There are three features of the summary statistics and correlations that we would like to highlight. First, the grade distributions for all the assessments are relatively normal but all demonstrate a slight negative skew, which is to be expected for criteria-referenced assessments. That is, most of the students do quite well yet there is still some variability in student learning and performance. Second, the correlations among the assignments from week to week are stronger than the correlations among assignments with a longer time lag. To observe this pattern, examine the correlations just below the diagonal in the matrix relative to the correlations further from the diagonal. The correlations just below the diagonal are, for the most part, stronger than the others. Third, scores on assignments were predictive of exam performance but assignment 1 illustrates the lowest correlation with exam scores among all the assignments, confirming our anecdotal evidence, and feedback from students in the discussion forums, that assignment 1 was a poorly designed assessment.

With respect to reliability and validity, the data suggest that most of our assessments are satisfactory but could use improvement. And again, the data suggest that assignment 1 is unsatisfactory. We formally tested the reliability of the 11 Assignments using Cronbach's alpha and we found alpha = .91, which suggests a strong degree of reliability.

To assess validity, we conducted regression analyses predicting exam scores from assignment scores. If the assignments are valid then they should demonstrate *predictive validity*, that is, they should predict scores on the exams. There were 6 assignments before the midterm exam so we conducted a multiple regression analysis with midterm exam score as the outcome variable and assignments #1-6 as predictors. The model $R^2 = 0.43$, which suggests that 43% of the variance in midterm scores can be predicted from the assignments. Another way to think about this result is that the correlation between midterm scores predicted by the regression model and the actual scores is $r = 0.66$ (the square root of the model R^2). As we suspected, all the assignments were significant predictors of midterm score (for all $p < .05$), with the exception of assignment 1 ($p = .19$). In fact, if we remove assignment 1 from the regression model, the R^2 value remains = 0.43, again suggesting that assignment 1 was a poor assessment and lacks convergent and predictive validity.

We conducted a similar analysis to predict final exam scores. Between the midterm and the final exam there were 5 assignments (7-11) so we conducted a multiple regression analysis with final exam score as the outcome variable and assignments 7-11 as predictors. The model $R^2 = 0.34$, which suggests that 34% of the variance in final scores can be predicted from the assignments. Again, another way to think about this result is that the correlation between final exam scores predicted by the regression model and the actual scores is $r = 0.58$ (the square root of the model R^2). All of the assignments were significant predictors of final exam score (for all $p < .05$), with the exception of assignment 7 ($p = .68$). This was somewhat surprising to us and suggests that we should re-evaluate assignment 7 before the next iteration of the course.

SUMMARY AND LESSONS LEARNED

Overall, our experience designing and teaching *Statistics One* was a pleasant one. We enjoyed preparing and delivering content to students around the world. However, we can identify a few limitations in the process, especially related to the implementation of assignments. We detail these here, so that our experience can benefit those who intend to implement an online course.

Creating an assignment on *Coursera* can sometimes be cumbersome – various possibilities are offered, but their implementation is often tedious. For example, we sometimes encountered difficulties with regard to response formats. *Statistics One* required submitting R code, which slightly complicated the pre-defined response modalities. Progress has been made on that end – *Coursera* now allows submitting programming assignments directly with servers that check outputs for correctness, and efforts are being underway to ease the process from the instructor's perspective (e.g. more user-friendly interface, multiple access routes merged into single paths, etc.). It is therefore likely that creating assignments will become more straightforward in the near future, especially for courses in which advanced data processing is required.

The vocabulary surrounding assignments can also be puzzling to new users. For example, *Coursera* distinguishes between a 'soft' and a 'hard' deadline. The former refers to a preferred deadline for an assignment, in order to stay on track or to get full credit. The latter is the actual deadline – submissions past this date are not accepted by the system. This terminology was sometimes confusing to students. We had no intention to forbid access to particular assignments after a deadline, especially given that our course did not offer a certificate of completion or a statement of accomplishment. This aspect, however, created some confusion among students who were not clear about which deadline to refer to as the course progressed. Again here, *Coursera* offers a help page to cover these issues, yet not all students take the time to refer to them. Therefore, it might be a good idea to prevent confusion by defining these terms at the onset of the course.

Another limitation we have identified pertains to the discussion forums. Although our assignments sometimes required students to reflect back on the course content, all questions were designed to call upon the content covered in the lectures and the labs. However, instead of referring back to the videos, some students tended to use forums to seek answers and explanations. This is perfectly acceptable, but from an instructor's perspective it requires dedicating a substantial amount of time to monitor and answer threads. Because we had limited staff working on the course, we could not adopt a proactive strategy on potential issues arising or on topics needing clarifications. Rather, we were forced to fix problems after they arose, in a process that was certainly not optimal. In most cases, a

collaborative interaction emerged from students willing to help each other. However, we should also note that forum threads were not always constructive – a few students used anonymous comments to post unethical remarks directed toward other students or the staff. In our opinion, enabling anonymous comments is superfluous, as only deviant behavior seems to emerge from it. As *Coursera's* interface improves, the option to disable anonymous comments should be provided; otherwise, additional effort should be made on the instructor's end to prevent these issues from happening.

The students in *Statistics One* were quite vocal about their frustration with Princeton University's policy of not offering either a certificate of completion or a statement of accomplishment. This is an internal issue at Princeton so we will not discuss it further here but our advice to anyone considering teaching a MOOC is to discuss University policy before you start.

We also advise prospective MOOC instructors to encourage live "meet-ups" among students. Some of the students in *Statistics One* spontaneously created local groups and those students seemed to thrive. Another idea that we entertained as we were teaching *Statistics One* was to offer several "meet the Instructor" sessions. For example, a MOOC instructor could "go on tour" the week before the midterm exam and conduct in-person review sessions. Of course this would come at a cost, in terms of both time and money, but it would be a fun way to connect teachers and students and enhance the MOOC learning experience.

References

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Koller, D., Ng, A., Do, C., & Chen, Z. (2013). Retention and intention in Massive Open Online Courses. *EDUCAUSE Review*, May/June 2, 62–63.

Novick, M. R. (1966) The axioms and principal results of classical test theory *Journal of Mathematical Psychology, 3*, 1-18.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing. Vienna, Austria: The R Foundation for Statistical Computing.