COMPARISON OF TESTING AND EVALUATION METHODS FOR NEW RESOURCES IN STATISTICAL EDUCATION

Bruce Dunham, Melissa Lee, and Gaitri Yapa
Department of Statistics, University of British Columbia, Canada
b.dunham@stat.ubc.ca

New resources for use in statistical education are developed each year. Often though either the development process or evaluation of new resources does not directly involve students, the prospective end-users of the products. Described here are the development and evaluations of videos and web visualization tools created as part of a large multidisciplinary flexible learning project for creating new resources for use in introductory statistics courses. Specifically, the use of individual interviews and focus groups are compared with regards the information they can yield on learning gains, student usage and opinions. The findings are based on studies involving 85 undergraduate students at the authors' institution.

INTRODUCTION

There is a large quantity of resources available online for teaching and learning statistics, including videos, textbooks, assessment tools, and applets/simulations/visualization tools. Our focus here is on the development of such materials and in particular how involving learners, the prospective end-users, can inform the process. The work described forms part of a large-scale project for creating flexible on-line resources for teaching introductory statistics at the authors' institution.

In proposing a general framework for the development of new online tools for statistical education, Ooms and Garfield (2008) suggest a so-called Interactive Evaluation Model (IEM). The model has four main components, briefly summarised as (1) Evaluation planning: team-building, end-users identification, needs assessment (including a review of existing resources); (2) Evaluation of educational value, including evaluations of (i) the design whereby an evaluator observes users interacting with the product and (ii) content; (3) Evaluation of use: such as surveys of users and non-users, and (4) Evaluation of impact: such as student and instructor surveys and student tests. A key aspect of IEM is its iterative nature, in that progression through the stages is non-linear with feedback from each phase informing the next cycle of the development process.

Although IEM appears a sensible framework for development of new learning resources neither it, nor anything very similar, seems to be routinely adopted in practice in the creation of online statistical education materials. A common omission appears to be phase (2), which involves observation of potential end-users interacting with prototype versions of the new resource. There are two main reasons for this: engaging members from end-user groups in observations can be expensive, time-consuming, and greatly extend the development process. Moreover, the development team may include experienced instructors who believe they can anticipate how students will interact and learn from the new resource.

While recognizing the challenges involved, we advocate for student involvement in the development of a new educational resource. Including such a phase in the process has been demonstrated to be insightful in the creation of learning materials. For instance, in the development of the PhET suite of simulation tools for teaching physics (see for example Adams et al. 2008, McKagan et al. 2008, and Wieman et al. 2008), over 250 individual interviews were conducted in "think aloud" format in which students interacted with prototype simulations. These interviews were apparently very insightful and greatly informed the development process, revealing interface, pedagogical, and programming issues (McKagan et al. 2008). Such problems also arise with simulations tools for learning statistical concepts but, perhaps due to overlooking IEM phase (2) in development, may go undetected. For example, reflecting on data from student interactions with their *Sampling Distributions* applet and instructional materials, delMas et al. (1999) commented how students could become overwhelmed with information when interacting with the tools and also displayed disappointing learning gains, highlighting the types of issues that may arise in practice.

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

Here we describe two common approaches for obtaining information on how students engage with new online instructional resources: individual interviews and focus groups. These research methods have been used in the assessment and development of new applets and videos intended to improve the learning of key concepts encountered in introductory statistics courses. To begin we give context to the work, describing the background project and some related previous research. Our implementations of interviews and focus groups are described, compared and contrasted, and suggestions for best practice proposed. The results of the testing described are not the main focus here, and will only be discussed in broad terms.

BACKGROUND

The authors' institution is not unique in having many different units involved in teaching introductory statistics courses. This proliferation of statistics teaching prompts various concerns, one being inefficiency in the development of instructional materials. A large-scale project, Flexible Learning Introductory Statistics (FLIS), is ongoing at the authors' institution and involves instructors from seven units and three faculties. The project aims include the development of online teaching materials that address topics known to be challenging to learners taking introductory statistics courses, resources that are open-source, tested and assessed, and align with pedagogical research. An important deliverable from FLIS is a new online repository for related resources, StatSpace (https://statspace.elearning.ubc.ca/), where all learning materials described here can be found.

Our FLIS project commenced by creating resources for concepts that are common to (nearly) all introductory statistics classes and where research indicates learners have difficulties (see for example Chance et al., 2004, and Castro Sotos et al., 2007). These topics include sampling distributions, Central Limit Theorem, confidence intervals, and hypothesis tests. Four new web visualization tools (referred to as applets here for brevity) were developed along with three videos. The authors comprised the assessment team responsible for testing these new resources as part of the development cycle.

Many applets have been developed over the past twenty years that aim to help the learning of statistical concepts. Popular applet suites include the Rossman and Chance applet collection (www.rossmanchance.com/ISIapplets.html), the Statistics Online Computational Resource (www.SOCR.ucla.edu), and Lane et al*.'s* online statistics book (www.onlinestatbook.com). There is strong evidence applets can be effective for learning when compared with other instructional methods, at least when applets are augmented with an auxiliary resource such as priming questions to guide the learner's interactions (for examples, see Lane & Tang, 2000, Lane & Peres, 2006, and McDaniel & Green, 2012).

The four applets created by the FLIS project in turn target (A1) the sampling distribution of the mean when sampling from a Normal distribution, (A2) confidence intervals for a Normal mean, (A3) the Central Limit Theorem, and (A4) association for categorical variables via two-way tables and the Chi-squared test. Informed by the research findings (Lane & Peres, 2006, Adams et al. 2015) that applets can be ineffective to learning unless the student's interaction is suitably structured, tutorials were built into the first three applets, and a structured worksheet developed to accompany the fourth. The fourth applet targets various concepts, including type I and II errors, and did not lend itself to a built-in tutorial.

Compared to the evidence from structured use of applets, the evidence of learning directly from videos is somewhat patchy (see for example Kay, 2012, for a meta-analysis of related research up to 2011). In short, although learners tend to like online videos, their effectiveness for learning abstract concepts is probably low, at least if time-on-task is taken into consideration. For instance, DeVaney (2009) used videos in an online statistics course, the contents apparently mainly SPSS tutorials; although students were positive about the videos, in comparison with another section where the videos were not available there was no statistically significant improvement with regards academic performance. Given the inherently passive nature of video viewing it is not surprising studies have found the level of interactivity permitted with a video may impact learning (see for instance Schwan & Riempp, 2004, and Zhang et al., 2006), though any effect may vary with the type of material to be learned (for example, Merkt et al., 2011). Our three videos are

respectively on (V1) the sampling distribution of the sample mean, (V2) confidence intervals for a mean, and (V3) the one-sample t-test.

The suggestion is that the resources be used flexibly and in conjunction to form a suite of materials adopted by learners throughout an introductory statistics course. It is proposed that repeated and connected engagement with the videos, applets, and related WeBWorK questions (see Cubranic et al., 2014, for an overview of the WeBWorKiR resources) will optimize the learning gains from these materials.

ASSESSMENT METHODS

After approval was obtained from the authors' institution's research ethics board, student volunteers were recruited from six statistics classes via emailed invitations. Interested students were invited to indicate their availabilities via an online scheduling tool, providing their preferred contact emails. Once times were determined, students were contacted with the details and then sent a reminder and the consent form the day before. In total, 85 student volunteers participated. All were given C$15 by means of recompense for their time. The goals of the assessments were to gain information about how students interact with resources, how students perceive the resources, and possible learning gains arising from engaging with each resource. The assessment team members were not the originators of the resources under consideration.

Two formats were used for the assessment sessions: individual interviews and focus groups. The motivation for using individual interviews came partly from previous research involving the development of the PhET simulations (McKagan et al. 2008), learning tools close in style in spirit to the applets developed in the FLIS project. One-one sessions during which students are closely observed in their interactions with an applet yield fine details about student behaviour when engaging with the tool. Interviews also permit a "think aloud" mode where the learner may give a verbal commentary of their thinking in real time, something our volunteers were encouraged to do. One can be reasonably sure the opinions voiced by a student in such interviews are their own, or at least are not directly influenced by other students. Interviews were only used to assess applets and not videos, as it was considered unlikely much information would be obtained from observing students watching a video alone.

All interviews involved two of the authors. One would conduct the interview, while the second acted as an observer. The observer would either sit a few metres behind the student, close enough to see and make notes on the student's interactions, or would observe remotely via Skype. The latter had some advantages as interactions could be observed more closely and although the students were informed they were being observed remotely we suspect students tended to forget during the sessions. Prior knowledge and possible learning gains were measured by a set of pre and post questions based on ARTIST questions (delMas et al., 2007), the pre-test priming the students in a way believed to enhance learning from a simulation (Jong & van Joolingen, 1998).

The initial set of interviews we performed attempted to engage students with two related applets, A1 and A2. Given the one-hour time constraint and the variability in the time students took to complete tasks, attempting to assess two resources in a single session proved untenable, as the sessions became overly long or rushed. Also, the students were set different tasks before and after engagement with the applets, making assessment of learning difficult. Subsequent interviews involved a single applet, and where (as for applets A2 and A3) exposure to applet A1 was desirable, it was ensured students had been given opportunity to engage with A1 in their course prior to their interview. Within each session the same concept-based questions were used pre and post.

Subsequent interviews had the following format (with approximate timing, in minutes): welcome, overview of the project, preamble, consent form (5), pre-test (10), interaction with the applet via built-in tutorial or (for A4) a separate activity (25), post-test (10), discussion of student impressions, feedback and suggestions (5), review of test questions, thanks and payment (5).

Focus groups have often been used in educational research, mainly for facilitating discussions to elicit opinions. For examples, Dunn et al. (2015) used focus group to explore students' perceptions of and motivations for watching videos on statistical topics, and Hund & Getrich (2015) similarly used focus groups to gauge the opinions of video materials in a graduate biostatistics course. Most of our focus group sessions also involved student discussions of videos,

but were rather different in that for much of the sessions students were viewing the videos, providing opinions, or attempting test questions in isolation. Only near the end were there group discussions. All sessions were held in computer laboratories and students were requested to bring ear/headphones, although spare sets were available for the few that forgot. Six focus group sessions ran with between 7 and 16 students, involving a total of 59 volunteers overall.

The structure of the focus group sessions involving a video were as follows: welcome, preamble, initial set-up, consent forms (5 mins), pre-test (10), first video view, payment (10), on-line survey (5), second video view (5), post-test (5), students split into groups of three or four to discuss their opinions (5), groups feedback to the room (5), review of test questions, thanks (5).

One focus group involved applet A4 after it had been tested in six one-one interviews. The structure was similar to that for the video sessions, except that rather than viewing a video the students interacted with the applet and the accompanying activity. This session was arranged to explore whether a similar quality of information could be gleaned from a group session involving an applet compared to multiple individual interviews.

Following each set of assessment sessions on a prototype resource, summaries of students' responses, behaviour patterns, and opinions were written up, and a short report indicating any issues and suggestions for modifications was provided to the development team for the resource. An overview of the main issues identified was also discussed at a FLIS project team meeting.

STUDENT RESPONSES

A more comprehensive description of how the students interacted with, responded to, and learned from the assessment sessions is planned for a subsequent article. A few of the most striking points are mentioned here.

Overall our student volunteers nearly all engaged in the sessions enthusiastically and appreciated the opportunity to be involved in the development of new resources. Feedback on the resources was very positive, particularly regarding the applets ("Fun. ... Graphics are clean. ... [helps] visualise what we learned in class.", "Ask us before we do it ... helps me think.").

In some instances students hit upon issues that had previously escaped the attention of the development team, including minor bugs and a typing error in the applets. More insights came from observing closely how the students interacted with the resources and their "think aloud" comments. For example, applet A1 initially had a button labelled "Calculate Many Means", which would simulate many samples of a fixed sample size and create the distribution of the resulting sample means. Some students found this confusing, apparently thinking that multiple means could be computed from the same sample, and the button was subsequently relabeled as "Means for Many Samples". In testing for applet A4, the accompanying activity asks students to simulate fifty contingency tables, which would take around a minute to complete; however, the applet includes a "Faster" button to speed up simulations, and all students observed in interviews used this feature rather than waiting. This caused a problem as stopping at exactly fifty was difficult, and the aim was to have the students observe the outcomes and the corresponding Chi-squared statistics to help them appreciate the associated variations. For the focus group, the instructions were changed to specifically dissuade the learner from using the "Faster" button on that part.

There was some evidence of learning gains from the studies, though patterns were not similar across all resources tested. In particular, video V1 appeared helpful to learning whereas V2 and V3 seemed more likely to increase confusion. For the 42 students who engaged with the videos, only 10 improved their scores on the learning test from pre to post. Of the 22 students engaging with the applets, 8 increased their scores on the learning test. About half the students engaging with either resource type had scores that did not change from pre to post.

COMPARISON OF METHODS

In comparison of the modes of assessments, we summarise here what we found to be the main "pros and cons" for both individual interviews and focus groups.

Interviews with a single student permit observing detailed individual interactions with resources, the sessions being student-paced. The information gleaned from "think aloud" commentaries is especially insightful, and one can be sure that any student's opinions are not influenced by those of other students. However, such interviews come with appreciable time cost,

particularly if as suggested two researchers are involved with each session. It can be difficult to arrange mutually convenient times to schedule an interview, and should the student fail to make the appointment much time is wasted. Some students struggled to "think aloud" during the interviews, and on rare instances a volunteer appeared to show signs of stress at being observed. Moreover, it is possible a single student may feel inhibited to express negative opinions during such sessions.

Focus groups enjoy an economy of scale compared to individual interviews – many more students can be accommodated in a single session. Students appear to enjoy the group interactions and seem more likely to freely express their opinions, particularly if validated by their group. The sessions are less stressful for the volunteers, and a further benefit is that no time is wasted if a small proportion of the volunteers expected fail to turn up. On the other hand, there are difficulties in closely observing how individual students interact with the resource during a focus group, and no "think aloud" is possible. If there are problems (say with the technology or the resource) they may be difficult to resolve in real time. For instance, a glitch in the online survey tool in our first focus group resulted in some aspects of the data being lost. Students may vary greatly in the time taken to complete activities, as for example, in the focus group for applet A4 students took between around 15 and 25 minutes to complete the activity. Finally, it may be difficult to prevent students from "cheating" on the pre and post tests since it may not be practical to operate such tests under anything like examination conditions and there would not be repercussions for students who saw the responses of another student.

CONCLUSIONS

As part of a team involved with creating new, flexible online resources for learning concepts encountered in introductory statistics courses, the authors have engaged with potential end-users via both individual interviews and focus group sessions. These sessions have been discussed in some detail, and the perceived advantages and disadvantages described for each. The proposed role of such research in the interactive evaluation model of Garfield and Ooms has been highlighted, along with guidelines on how to conduct each type of session.

It is natural to question which, if either, type of research method is preferable. Given their relative merits and demerits, the authors propose that it is ideal to engage potential end users in both types of sessions when assessing a resource in development. However, and mindful of the time cost of individual interviews, we suggest that relatively few individual sessions may be required. From our experience next to nothing was gained after conducting three or four individual sessions with each applet, the fine details obtained from each session being mostly replicated in those subsequent. Hence our suggested approach is to schedule a few individual sessions for a prototype resource, provide feedback to the resource development team who can then modify the resource before it is subsequently assessed again via at least one focus group session.

Creating online learning materials can be costly and once the development is completed it may be impractical to make substantial revisions to a resource. Hence it becomes important to assess a prototype resource on potential end-users, gauging their opinions, difficulties, and learning gains. Interviews and focus groups can provide complementary data to inform the development team on how to improve the final product and give guidance to both teachers and learners on usage of the resource. We propose that activities of this nature greatly inform the creation of new online resources and should routinely form part of the development cycle for such artefacts.

ACKNOWLEDGEMENTS

The authors wish to thank Doug Bonn, Nancy Heckman, Mike Marin, Zachary Rothman, Mike Whitlock, and Eugenia Yu for their contributions. The project would not have been possible without the generous support of UBC's Teaching and Learning Enhancement Fund.

REFERENCES

Adams, W. K., Armstrong, Z., & Galovich, C. (2015). Can students learn from PhET Sims at home, alone? In A. Churukian, D. Jones, & L. Ding (Eds.), *Proceedings of Physics Education Research (PER) Conference on Critical Examination of Laboratory-Centered Instruction and Experimental Research in Physics Education* (pp. 23-26).

Adams, W. K., Reid, S., LeMaster, R., McKagan, S.B., Perkins, K.K., Dubson, M. & Wieman, C.E. (2008). A Study of Educational Simulations Part II - Interface Design. *Jl. of Interactive Learning Research, 19*(4), 551-577.

Castro Sotos, A.E., Vanhoof, Stijn, Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*, 98-113.

Chance, B., delMas, R. & Garfield, J. (2004). Reasoning about sampling distributions. In *The Challenge of Developing Statistical Reasoning and Thinking* (pp. 295 – 323). Kluwer Academic Press.

Cubranic, D., Dunham, B., & Kim, D. (2014). *On-line homework in probability and statistics: WeBWorK incorporating R*. Contributed paper at 9th International Conference on Teaching Statistics, Flagstaff, Arizona, USA.

delMas, R., Garfield, J., & Chance, B. (1999). A Model of Classroom Research in Action: Developing Simulation Activities to Improve Students' Statistical Reasoning. *Journal of Statistical Education, 7*(3).

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 6*(2), 28-58.

DeVaney, T. A. (2009). Impact of Video Tutorials in an Online Educational Statistics Course. *Journal of Online Learning and Teaching, 5*, 600–608.

Dunn, P.K., McDonald, C. & Loch, B. (2015). Statscasts: Screencasts for Complementing Lectures in Statistics Classes. *International Journal of Mathematical Education in Science and Technology*, *46*(4), 521-532.

Hund, L. & Getrich, C. (2015). A Pilot Study of Short Computing Video Tutorials in a Graduate Public Health Biostatistics Course. *Journal of Statistics Education, 23*(2).

Jong, T. de & van Joolingen, R. (1998): Scientific Discovery learning with Computer Simulations of Conceptual Domains, *Review of Educational Research*, *68*, 179-201.

Kay, R.H. (2012). Exploring the use of video podcasts in education: A comprehensive review of the literature. *Computers in Human Behavior*, *28*(3), 820-831.

Lane, D.M. & Peres, S.C. (2006). *Interactive simulations in the teaching of statistics: Promise and pitfalls*. Presented at the 7th International Conference of Teaching Statistics, Salvador, Brazil, http://icots.info/7/pages/session.php?s=7D.

Lane, D.M. & Tang, Z. (2000). Effectiveness of simulation training on transfer of statistical concepts. *J. Educational Computing Research, 22*(4) 383-396.

Merkt, M., Weigand, S., Heier, A., & Schwan, S. (2011). Learning with videos vs. learning with print: The role of interactive features. *Learning and Instruction, 21*(6), 687-704.

Ooms, A. & Garfield, J. (2008). A Model to Evaluate Online Educational Resources in Statistics. *Technology Innovations in Statistics Education*, *2*(1).

McDaniel, S. N., & Green, L. B. (2012). Using Applets and Video Instruction to Foster Students' Understanding of Sampling Variability. *Technology Innovations in Statistics Education, 6*(1).

McKagan, S.B., Perkins, K.K., Dubson, M., Malley, C., Reid, S., LeMaster, R. & Wieman, C.E. (2008). Developing and researching PhET simulations for teaching quantum mechanics. *American Journal of Physics*, *76*, 406-417.

Schwan, S. & Riempp, R. (2004). The cognitive benefits of interactive videos: learning to tie nautical knots. *Learning and Instruction*, *14*, 293-305.

Wieman, C.E., Adams, W.K., & Perkins, K. K. (2008). PhET: Simulations That Enhance Learning. *Science, 322*(5902), 682-683.

Zhang, D., Zhou, L., Briggs, R.O., & Nunamaker, J.F. (2006). Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness. *Information & Management*, *43*(1) 15-27.