Australasian Journal of
Educational Technology

2010, 26(2), 226-237

Use of audience response systems for summative assessment in large classes

Terence M Hancock
University of Louisville

The audience response system – technology that allows immediate compilation and display of a group's multiple choice input – is being shown effective in the classroom both in engaging students and providing real time, formative assessment of comprehension. This paper looks at its further potential as an alternative for summative assessment, replacing conventional examinations and testing as a basis for student grades. After brief discussion as to practical benefits of doing so and a review of ARS – hardware, software, and studies of its impact on learning – we develop and report outcomes of two distinct trials utilising ARS for both formative and summative assessment. Results indicate synergies in combining the two forms of assessment, though student attitudes are found to be particularly sensitive to the instructor's approach to design.

Overview

A growing body of research indicates classroom use of audience response systems (ARS) can have a significant, positive impact on learning outcomes (Simpson & Oliver, 2007), and nowhere is this of more interest than university classes with large enrolments (Barnett, 2006; Sharma et al., 2005). These classes present special challenges to student success. They are often foundational, involving core knowledge or tools upon which later classes are built, and therefore are both critical, and concentrated in early academic careers when students are most vulnerable. They create a "lost in the crowd" feeling that stifles active engagement, intimidates some from even asking questions, permits others to hide, come unprepared, or with little excuse, not at all.

Absent are reliable cues as to class comprehension, indicators that help teachers gauge whether it is necessary to recapitulate or safe to go on. Subsequent instruction may be meaningless because some early, key concept is broadly missed or misunderstood, and none of this reveals itself – to either student or teacher – until after an examination.

Audience response systems specifically target these problems (Dufresne & Gerace, 2004). Based on audience polling devices seen on television (e.g. *Who Wants to be a Millionaire?*), they have become remarkably reliable, inexpensive and easy to use. Multiple choice questions, generally but not necessarily developed before class, are presented at strategic points throughout the lecture. Students indicate their choice by pressing the appropriate key on their "clicker," a handheld, wireless remote device. Upon close of polling, by way of *formative assessment*, the instructor can display histograms of voter response and indicate the correct answer, providing student and teacher alike immediate feedback as to comprehension, and of course, an opportunity for questions and clarification. Incentives can be added through participation credit. Each clicker has a unique ID allowing responses to be captured and scored individually according to a variety of options available to the instructor.

Given this last capability, this paper explores whether an ARS could be as effective in *summative assessment* (measuring for purposes of determining a student's final grade in a course) as in *formative assessment*, at least in large enrolment classes where multiple choice questions are already standard for student testing. Research focusing on the benefits and challenges of using ARS for summative assessment is otherwise lacking (Kay & LeSage, 2009).

First is a brief discussion of problems particular to student testing in large enrolment classes, followed by a review of ARS hardware and software and studies related to its use. Finally, we describe and report outcomes from two very different approaches utilising a clicker system in lieu of conventional, paper based testing.

Summative assessment in large classes

Student learning in large university classes is typically measured by exams ("tests") that can involve a fair mass of paper. Attached reference pages – tables, formulas, etc. – are often separated and spread out for easy access beside the open test, calculator, pencils and scan sheet, and invariably working space between adjacent test-takers overlaps. Space suitable for note taking during standard lectures becomes wholly inadequate for exams (they occur in the same classroom) and with large classes there may be few empty seats to permit student dispersion.

This crowding is at best a distraction; at worst, a rich opportunity to cooperate. Teachers create multiple test versions to discourage cooperation, but this is hardly foolproof, and every step from distribution to grading becomes decidedly more complicated. There are security issues outside the classroom as well. Exams themselves represent massive copy jobs that invariably put tests in hands other than the instructor's. Student assistants, student workers at the copy centre and others with prior access can be induced to share advance copies. In multiple section classes, tests from earlier sessions can escape the classroom and find their way to later-section students.

Tests can also involve significant expense. Beyond the obvious production costs – time, paper, equipment, etc. – old tests cannot simply be thrown away. There may be costs for storage and disposal and certainly for the machine-scannable answer sheets. Use of these "scan sheets" is ubiquitous in large enrolment classes in American universities, but tradeoffs go beyond monetary cost and even the limited scope of possible questions. For one, scan sheets can accommodate only a single correct choice per question. This works well for simple true-false dichotomies. But in most cases, besides one correct or best response, other choices run "warm to cold." The graded scan sheet cannot differentiate this range of learning. "Nearly correct" is no better than "not even close" in assessing response.

Quantitative questions suffer this issue and worse. An inadvertent misstep – a bad punch on the calculator – carries the same weight as a true conceptual error. This is particular issue with the common practice of creating a series of test questions from a single problem. If each question represents a successive stage in a solution process, an early mistake can "cascade" and be the sole cause for every subsequent answer being wrong. A student who may genuinely know the material, who has otherwise done the process perfectly, is accorded no credit for failures unrelated to anything being taught or tested for.

So for reasons of cost, convenience, security and validity, technologies that might permit alternative means of summative assessment certainly merit study.

Audience response technology

There are a number of competing audience response systems, each with their own balance of cost, features, and ease of use (Barber & Njus, 2007). The institutional trend is toward campus-wide adoption so students need no more than one clicker for all classes. At this university, the standard is *i>clicker*, acquired by Macmillan in 2005 from physicists who developed it for their own classes at the University of Illinois. The student remote is sold at both on and off campus bookstores. Receiving units, software and instructor remotes are furnished at no charge by the manufacturer. Remotes have five keys labeled A through E in addition to an on-off button, and with periodic change of batteries, are expected to last the student's college career. They can be registered to a particular student either over the web or with a couple clicks in the classroom at the start of the semester.

Instructor remotes are identical to student clickers except for distinguishing color, if desired. Keys can start and stop polling (A), display a histogram of student response (B), index *PowerPoint* slides (forward C, reverse D), and designate correct response (E). The receiving unit is about sandwich-size and powered through a USB port on a standard classroom computer. It can handle as many as 1500 clickers from as far away as 250 feet, with a switch of sub-frequency preventing interference between units in nearby classrooms. Votes successfully received are confirmed with an indicator light on the student's remote. While polling is open, elapsed time and total respondents are shown continually, and students can repeat or change their choice as often as they wish.

The software can stand alone or interact with course management systems like *ANGEL, Blackboard* and *WebCT*, downloading student information or uploading grades. Where questions are meant to elicit opinion or discussion, polling can be anonymous, without grade effect, or with any choice weighted equally. Otherwise, each possible response can be given unique weight, full or partial credit at the instructor's discretion.

Clicker questions can be created on the spot simply by defining alternatives to the class and pressing the start key. Typically, however, they are prepared in advance, requiring no more than inserting a *PowerPoint* slide into the class presentation. A screenshot is captured when polling is activated and displays with response statistics for reference in the database. Any correct choice, grade weight, etc. can be assigned in advance, by default, or sometime later.

The technology has spread from science classrooms to virtually every academic discipline, applications as different as the material being covered and the style and intent of the instructor (Draper, Cargill & Cutts, 2002). While clickers do not automatically boost exam scores, they have never been shown to harm them (Knight &

Wood, 2005). Their impact depends largely on how questions are crafted and fit to outcomes (Beatty, Gerace, Leonard & Dufresne, 2006), how effectively they are used to pace lectures (Poulis, Massen, Robens & Gilbert, 1998); and what weight participation bears on grades (Hake, 1998).

A growing body of research is defining best practice and the genuine potential of ARS as a learning tool. However, "reports of using clickers for summative high-stakes testing are relatively rare" (Caldwell, 2007). Given the shortcomings of traditional, paper-based testing, and where clickers are already familiar and trusted, using ARS for summative assessment seems an alternative worth studying.

ARS questions – focus and discipline

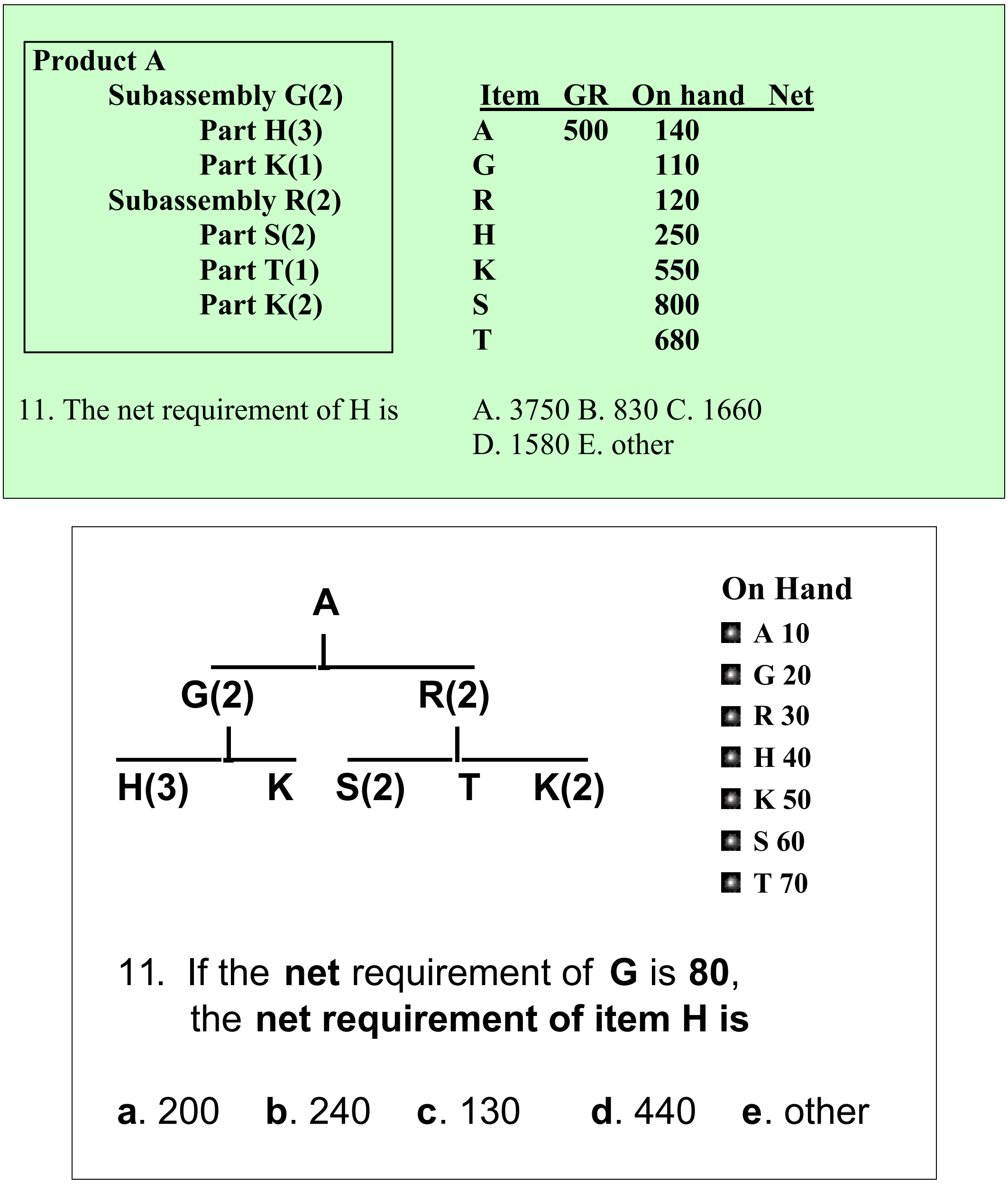
The author presently uses clickers in two courses, both senior-level undergraduate classes taught in multiple sections, each with typical enrolment of 60 to 70 students. *Operations Management* is a survey course required of all College of Business students, delivered primarily in lecture format. *Project Management* is a focused, skill-based elective combining lecture, lab, and hands on project work.

Clickers were first introduced in the *Operations* class as student engagement seemed more critical there; commitment is often lower in required than in elective classes. At the time, test scores were averaging only 56 percent, requiring an unearned grade adjustment ("semester curve") of 21 points to achieve a desired median grade of B-. Clicker questions are sprinkled throughout the lecture, usually 3 to 6 per 75-minute session, a quantity consistent with recommended practice (Preszler, Dawe, Shuster & Shuster, 2007). They may involve material just being covered or perhaps readings and problems assigned from a previous class. At present, older material is revisited as well. Correct answers and response statistics are shown immediately, with discussion encouraged.

Half credit is awarded any response (simply participation), another half if correct. Throughout the semester, a student's total can be compared to cumulative points possible, and that ratio becomes basis of bonus credit up to 20 points at semester's end. Clicker participation is optional, but students are warned of little or no curve otherwise. Expectation of regular use and at least modest grade impact seem all that is necessary for much of the benefit associated with ARS; students are more apt to attend, prepare, stay abreast and involve themselves if there is daily consequences for doing otherwise.

Quick, simple questions suffice. Greater benefit arises, though, the more questions support the most challenging material or wherever comprehension is least certain (Allen & Tanner, 2005). Typical textbook problems do not translate well. ARS questions need to be sufficiently honed that they fit neatly on a single *PowerPoint* slide and assess comprehension with some accuracy, some specificity, without consuming much class time. Text problems may involve multiple dependent steps, peripheral aspects intended to place new material in some broader context, and significant investment of time. Figure 1 illustrates some of these issues and how they might be addressed.

The upper half of Figure 1 shows a conventional dependent demand problem, the bill of material at left representing the raw materials and component parts that build into the finished goods. For example, 3 H and 1 K assemble to make a single G; then 2 of these G along with 2 R assemble to become one Product A.



Product A
Subassembly G(2)
Part H(3)
Part K(1)
Subassembly R(2)
Part S(2)
Part T(1)
Part K(2)

Item	GR	On hand	Net
A	500	140	
G		110	
R		120	
H		250	
K		550	
S		800	
T		680	

11. The net requirement of H is A. 3750 B. 830 C. 1660 D. 1580 E. other

On Hand

- ■ **A 10**
- ■ **G 20**
- ■ **R 30**
- ■ **H 40**
- ■ **K 50**
- ■ **S 60**
- ■ **T 70**

11. If the **net** requirement of **G** is **80**, the **net requirement of item H is**

a. 200 **b**. 240 **c**. 130 **d**. 440 **e**. other

Figure 1: Conventional dependent demand problem (top); reformulated as ARS question (bottom)

The net required of H depends on quantities calculated for G, which in turn depends on A. A careless error any point in the process makes every subsequent calculation wrong, and the probability of error accumulates with every punch of the calculator. Where time permits, the motivated student responds by checking and rechecking work (though sometimes just repeating errors). Where time doesn't permit, genuine effort may be little superior to random guessing.

The lower half of Figure 1 illustrates how adjustments might be made. First, information is presented about how best to assimilate (the indented bill of materials is recast as a "product structure tree"). Numbers are simplified to ease calculator input; and if possible, to where a question might be accessible to someone without a calculator (on-hand inventories shrink from 3- to 2-digit multiples of 10). The scope of the problem is narrowed, perhaps focusing on a single, critical step rather than the full solution process (an intermediate value is provided, allowing an independent starting point for this item). And finally, anything peripheral to the explicit pedagogical purpose is stripped away (Fies & Marshall, 2006) (for example, a question would include no mixed units of measure unless converting between units is the specific ability being assessed).

Polling time

The period allowed for response – polling time – can be fixed in advance, giving equal time to each question, or controlled flexibly from the instructor's remote. In the *Operations* class, expectations depend on problem type: 30 seconds for simple true-false, 1 minute for standard multiple choice, and 2 minutes if more than rudimentary calculation is involved. Within reason, this is adjusted according to vote count display, with a verbal call well before polling is closed. Voting speed can be nearly as meaningful as the vote itself in assessing student comprehension. When responses are unexpectedly slow, a choice is made to either extend polling or to close and return immediately to discussion. Little effort is needed to reformulate a question for re-using in a later class.

Figure 2 compares grade results in the *Operations* class before and after clickers were introduced. Each column represents the compilation of 4 course sections, between 216 and 244 students. Column 2 (ARS, T1) shows average scores after integrating clickers within lectures, all else being the same. As before, three traditional paper tests were scaled to 30 points each and a cumulative final was scaled to 10 points, totaling 100 semester points. Comparing Column 1 (Pre-ARS, T1), test scores jumped 13% from 56 to 63 with clicker use. Where previously a curve averaging 21 points had been necessary to bring the class to 77%, the desired mean grade, points awarded for clicker use, 15 of 20 possible, achieved virtually the same overall grade.

First ARS testing approach

After 2 years of faultless operation and student feedback consistently high with regard to clickers in the classroom, using them in place of traditional exams seemed a natural step. A review of articles found no more than anecdotal mention of ARS use in summative assessment, certainly little to discourage it. So, at the start of year 3, "paperless" testing was added to the *Operations* class and introduced as a "green" or "environmental" initiative.

Testing remained "closed book." Students typically received a single reference sheet, anything needed in the way of formulas or tables, the back of which could serve as scratch paper. This involved little to produce, hand out and keep track of, and students understood they had to sign and return the sheet to receive their grade.

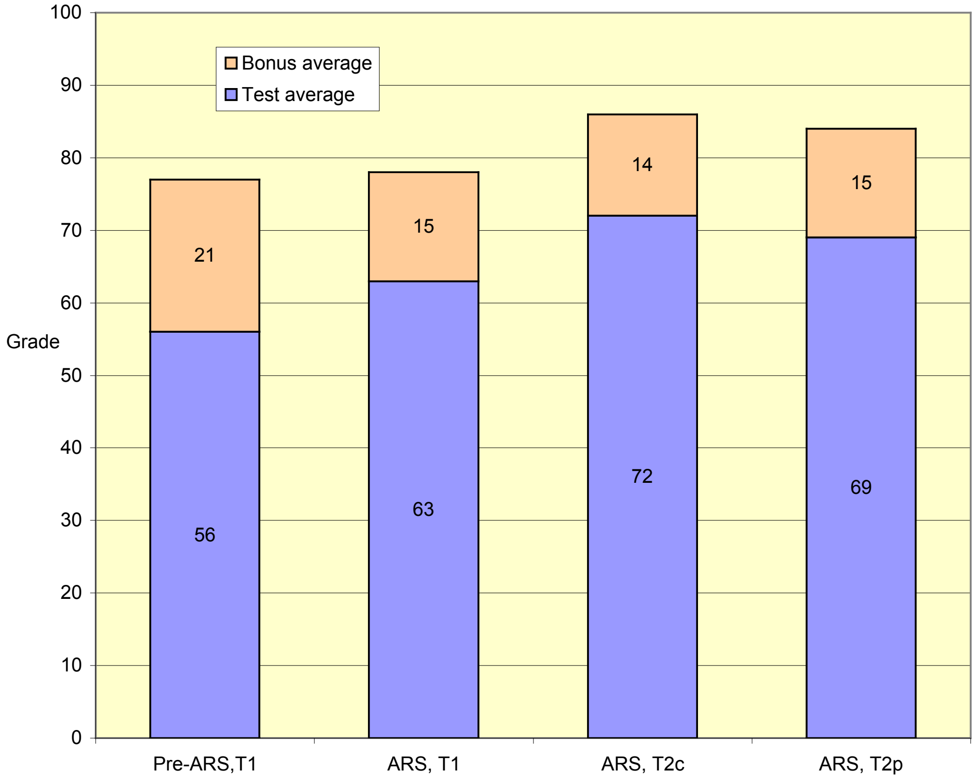


Figure 2: Average total grades in *Operations Management*
Column 1: No clicker use, Traditional tests with scan sheet
Column 2: Clickers during class only, Traditional tests with scan sheet
Column 3: Clickers during class, Tests by clicker with ARS style questions
Column 4: Clickers during class, Tests by paper but with ARS style questions

Test questions were presented as a series of *PowerPoint* slides indexed from the back of the classroom where student activity could be monitored. Students were reminded genially that they had no need to direct attention anywhere but straight ahead. Clickers were to be kept close to the person and out of range of any other student's peripheral vision. As with previous testing, students were required to bring and present on demand student ID (picture identification).

Questions paralleled the style and nature of those integrated within lectures, so students came to the exam with a good understanding what to expect. Each was deliberately focused and independent of others, targeting a specific critical idea or process step. Response choices "nearly but not entirely correct" were accorded partial grade credit (as much as half), though this was limited to 1 in 3 questions owing to uncertainty about its impact on grade distribution. Polling time followed roughly the same protocol used in class depending on question type and complexity, including fair warning before closing one question for the next.

Since no paper test was printed, leaked copies were not an issue. Mobile phones, programmable calculators – all things that could store or transmit information – were not permitted. As further precaution, the sequence of questions was shuffled between sections.

Findings of first trial

Results are shown in Column 3 (ARS, T2c) of Figure 2. The new testing process more than doubled the previous impact of ARS on test scores. While clicker bonus points remained approximately steady, 14 of 20 possible, average test score rose to 72. This was not entirely unexpected for reasons already suggested. First, ARS forced a discipline on test questions that reduced the possibility of "false negatives" – wrong answers for the wrong reasons, such as an error in one question being allowed to "cascade" into another. Second, students were more specifically prepared for test questions as they mirrored the style and nature of ARS questions practised in class. And third, capability of partial credit was exploited in a way not possible with traditional scan sheets. Zeroing partial credit, average test score falls from 72 to 67, reducing by half the impact of ARS testing.

In soliciting feedback, these issues folded into the discussion. Granted, primary benefits for ARS testing may be invisible to students (e.g. cost savings, improved security, reduced handling), but in their direct interest, test grades were arguably more valid and certainly higher. Nonetheless, reaction among *Operations* students was mixed – bright students feeling harnessed by the pace of the test, slow students feeling deprived of the flexibility to ruminate particular questions indefinitely, and a few throughout the spectrum lamenting the loss of scope to answer questions in random order or the opportunity to revisit earlier questions, as conventional tests allow.

Beyond its use in testing, however, overall impressions of ARS remained highly positive, particularly the novelty of it, the instantaneous feedback and statistics it made possible, as well as the bonus credit it garnered. As to non-test negatives, nothing was expressed beyond lighthearted grumbling about having to buy and "haul around" the clicker itself. There seemed no concern for its reliability. Figure 3 shows students' general attitude toward ARS, polled anonymously at semester's end. While 89% felt either positive or very positive about clicker use in everyday classes, this percentage shrank to 46% with respect to how it was used to replace conventional testing.

In one further experiment, then, ARS style questions were simply put to paper, setting aside issues of cost, security and handling to assess impact of controlling pace and sequence. As before, answers were recorded on scan sheets. Results are shown in Figure 2 Column 4 (ARS, T2p). Discounting effects of partial credit, average test score rose about 3%, from 67 (72 stripped of partial credit) to 69. Opportunity to take the test as a whole rather than one question at a time does indeed benefit the test-taker, but at least in this case, it is more than offset by the loss of partial credit, which scan sheets cannot accommodate.

Whatever the reality, too many students in the prior trial perceived attributes of their testing process – the paperless series of questions answered by clicker – as significantly negative, even unfair. ARS questions were seen favorably until concentrated beyond a certain few; which begged the question: why need they be?

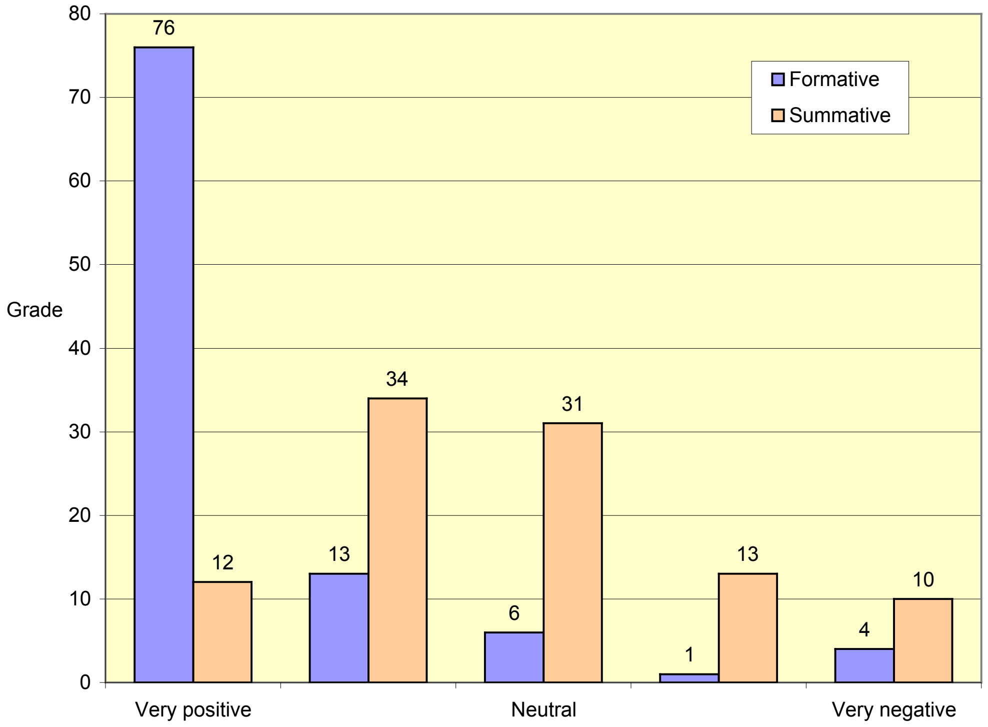


Figure 3: *Operations* students' attitudes toward clicker use
First in series: as a tool to support learning in regular class sessions
Second in series: as alternative to conventional testing

Traditionally, testing has been conducted in large, discrete blocks – entire class sessions – in part because it was difficult or inefficient to do otherwise. With ARS technology this is no longer the case. Given the ease with which comprehension can be assessed, perhaps functions of feedback and grading can be strategically combined.

Second ARS testing approach

When a decision was made to introduce clickers to a second class, *Project Management* (enrolment caps had been precipitously doubled from 30 to 60), full-period tests were eliminated in favor of routine "micro-tests." Six clicker questions were included in each 75-minute class period, three each at the start and end of class, addressing assigned readings, lab experience and any concepts or techniques previously covered. Accumulated points, which included partial credit, accounted for 80% of the semester grade (all but the course project). Splitting the questions between the beginning and end of class had the obvious effect of getting students to arrive on time and stay full period. More subtlety, the short sequence of questions reduced need to press any particular pace (more latitude to accommodate slower students); and since scope of material was generally fairly compact, students were less mindful of not being able to address questions in random order.

As the first set often dealt with homework and assigned readings, students came more consistently prepared and were more easily engaged in discussions. The second set was treated more flexibly, providing an opportunity to integrate, revisit and reinforce important concepts, especially those poorly demonstrated on a primary quiz.

Clicker introduction was coincident to other changes responding to the larger class size (individual projects became team projects, short essay test questions became multiple choice, etc.), so comparing before and after measures is not as meaningful as with the Operations class. What may be comparable are the attitudes reflected in Figure 4. While only 46% in *Operations* viewed clicker testing favorably, in *Project Management* that rose to 97%. Feedback has been entirely positive, most notably from those experiencing ARS in both classes.

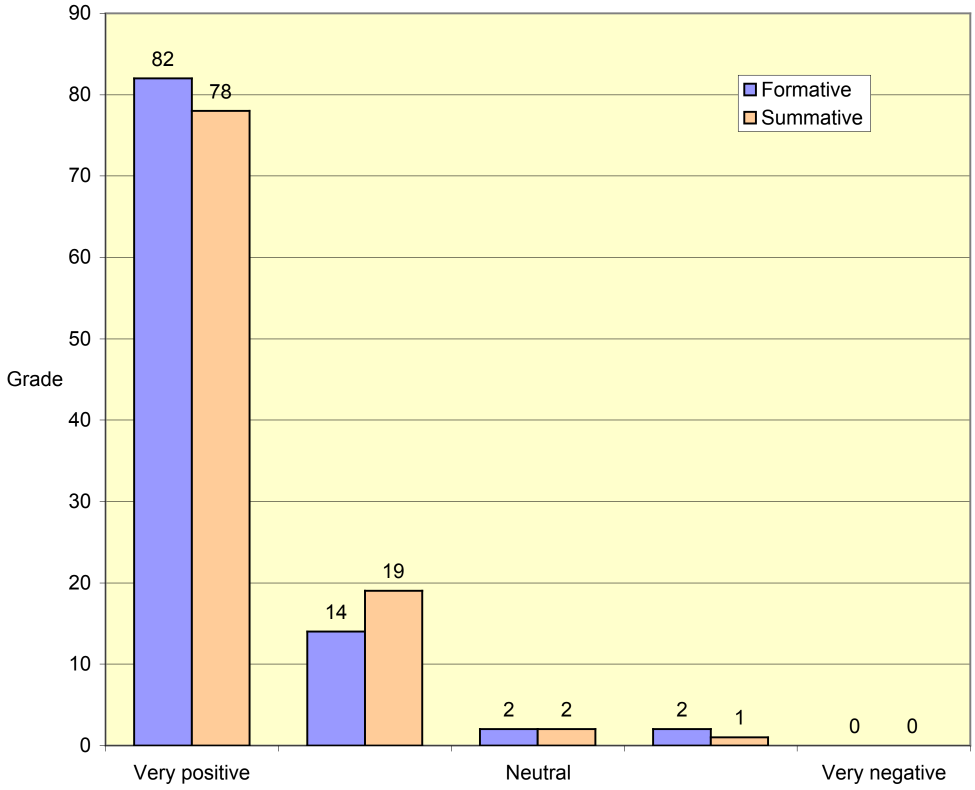


Figure 4: *Project Management*: Attitudes toward clicker use
First in series: as a tool to support learning in regular class sessions
Second in series: as alternative to conventional testing

Conclusion

Multiple choice questions may not be the richest measure of student learning, but in some large-enrolment university classes, they may be the only practical alternative. At scheduled intervals, some quantity of these questions are put to paper, a copy provided each student, and a class period set aside to answer them. The process is burdensome, costly, and not without issues of security and validity. ARS as a substitute is largely free from these concerns, and in fact, in the first trial here – when paper test was replaced with a series of *PowerPoint* questions, and scannable answer sheet by a clicker – a synergy was seen between use of ARS in everyday classes and the new testing process. Student familiarity with question style produced some of the benefit, presumably, and certainly more arose from the discipline ARS imposes on the focus and independence of questions. Still, many students expressed preference for paper tests when so many questions were strung together.

On reflection, the tradition of student grades being based on isolated, full-period tests may have little basis beyond convenience, and past difficulty in assessing more continuously. In truth, the practice is only sure in measuring how much students can "cram", last-minute into short-term memory.

ARS-enabled "micro-testing," as described in the second trial here, all but eliminated negatives perceived by students in trial 1. Formative and summative assessment merged into a single coherent process integrated seamlessly into every class meeting.

References

Allen, D. & Tanner, K. (2005). Infusing active learning into the large-enrolment biology class: Seven strategies, from the simple to complex. *Cell Biology Education*, 4, 262-268. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1305885

Barber, M. & Njus, D. (2007). Clicker evolution: Seeking intelligent design. *CBE-Life Sciences Education*, 6(1), 1-8.

Barnett, J. (2006). Implementation of personal response units in very large lecture classes: Student perceptions. *Australasian Journal of Educational Technology*, 22(4), 474-494. http://www.ascilite.org.au/ajet/ajet22/barnett.html

Beatty, I. D., Gerace, W. J., Leonard, W. J. & Dufresne, R. J. (2006). Designing effective questions for classroom response system teaching. *American Journal of Physics*, 74(1), 31-39.

Caldwell, J. E. (2007). Clickers in the large classroom: Current research and best-practice tips. *CBE-Life Sciences Education*, 6(1), 9-20. http://www.lifescied.org/cgi/reprint/6/1/9.pdf

Draper, S., Cargill, J. & Cutts, Q. (2002). Electronically enhanced classroom interaction. *Australian Journal of Educational Technology*, 18(1), 13-23. http://www.ascilite.org.au/ajet/ajet18/draper.html

Dufresne, R. J. & Gerace, W. J. (2004). Assessing-to-learn: Formative assessment in physics instruction. *The Physics Teacher*, 42, 428-433.

Fies, C. & Marshall, J. (2006). Classroom response systems: A review of the literature. *Journal of Science Education and Technology*, 15(1), 101-109.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1), 64-74.

i>clicker (undated). *i>clicker* mainsite. http://www.iclicker.com/

Kay, R. & LeSage, A. (2009). A strategic assessment of audience response systems used in higher education. *Australasian Journal of Educational Technology,* 25(2), 235-249. http://www.ascilite.org.au/ajet/ajet25/kay.html

Knight, J. K. & Wood, W. B. (2005). Teaching more by lecturing less. *CBE-Life Sciences Education,* 4, 298-310.

Poulis, J., Massen, C., Robens, E. & Gilbert, M. (1998). Physics lecturing with audience paced feedback. *American Journal of Physics,* 66(5), 439-441.

Preszler, R. W., Dawe, A., Shuster, C. B. & Shuster, M. (2007). Assessment of the effects of student response systems on student learning and attitudes over a broad range of biology courses. *CBE-Life Sciences Education,* 6(1), 29-41. http://www.lifescied.org/cgi/content/full/6/1/29

Simpson, V. & Oliver, M. (2007). Electronic voting systems for lectures then and now: A comparison of research and practice. *Australasian Journal of Educational Technology,* 23(2), 187-208. http://www.ascilite.org.au/ajet/ajet23/simpson.html

Sharma, M. D., Khachan, J., Chan, B. & O'Byrne, J. (2005). An investigation of the effectiveness of electronic classroom communication systems in large lecture classes. *Australasian Journal of Educational Technology,* 21(2), 137-154. http://www.ascilite.org.au/ajet/ajet21/sharma.html

Terence M. Hancock *PE PhD*
Department of Management & Entrepreneurship
College of Business, Belknap Campus
University of Louisville
Louisville, KY 40292, USA
Email: hancock@louisville.edu