論文

Low-Stakes Testing

Stephen M. Ryan[1]

Keywords: Testing, Low-stakes, Brain Science

High-stakes testing, where the test-taker has a lot to gain from a good score and as much to lose from a poor one, has always received a lot of attention from researchers seeking to inform both the people who make the tests and the teachers who prepare their learners to take them. At least until recently, though, low-stakes tests, typically designed and implemented by classroom teachers as wayposts along their students' journey to learning, have received less attention. Thanks to a renewed focus on evidence-based teaching, however, it is now becoming clear that low-stakes testing has much to offer in terms of support for and enhancement of learning.

High-Stakes Testing

The goal of a high-stakes test is to assess the amount of learning the test-taker has achieved and the test-taker's ability to make use of that learning (Association of Language Testers in Europe, 1999). The "high stakes" are usually important decisions about the test-taker's future: What job will they get? Which school will they be admitted to? What level of class will they be able to join? In the world of English Language Teaching, high-stakes tests include not only the globally available TOEFL, TOEIC, IALETS and their local equivalents such as the EIKEN suite of tests in Japan, but also admissions and placement tests used by schools, universities, and language centres.

Since the stakes for the test-taker, and often for the teacher or institution that prepares them to take the test, are high, the issues for researchers are fairness, validity, and accuracy (Young et al., 2013). To achieve the credibility a test needs in order to be used for high-stakes purposes, great efforts are made to see that the test will produce the same results for all test-takers with comparable abilities (fairness); that it actually measures what it appears to be measuring (validity); and that it measures that ability without other intervening variables that may distort the result (accuracy). Consequently, test-makers, such as the Educational Testing Service, Pearson Language Tests, and the EIKEN Foundation of Japan, devote considerable resources to making fairer, more valid, more accurate tests; schools and teachers preparing learners to take these tests scrutinise and analyse them to understand exactly what kind of items the experts are using; and learners devote a large proportion of their time and energy for preparing for

[1] 山陽学園大学総合人間学部言語文化学科

those testable items.

Clearly, there is a certain amount of distortion involved here (Jones et al., 2003). A particular skill or area of knowledge is likely to be selected for testing, not because it is a core to the ability the test claims to measure, but because it is testable under the conditions imposed by the need for high levels of fairness, validity, and accuracy. An obvious example, from language testing, is the issue of testing speaking fluency (De Jong, 2018). "They are fluent speakers of X" is a commonly heard layperson's assessment of a learners' ability to use language X, but fluency has proved very difficult to assess with the high degree of accuracy required for this kind of testing. Accuracy in grammatical structure and in pronunciation are more amenable to high-stakes testing, so it is often these skills that are tested, studied, and practiced, rather than the more intuitive but also more elusive fluency. Since the stakes for learners are high, the conscientious language teacher has little choice but to teach to the somewhat distorted focus of "speaking" tests such as the IELTS Speaking Test, the TOEIC Speaking and Writing Test, or the EIKEN interview test.

Researchers have also shown considerable interest in variables that are beyond the control of test-setters, notably the level of stress experienced by individual learners taking high-stakes tests. We have known since Yerkes and Dodson conducted their famous experiment with mice in 1908 that, while a certain amount of stress enhance performance, too much stress is detrimental to it. Unfortunately for those who seek to standardise the experience of being tested, it is equally clear that the level of stress experienced by test-takers depends not only on how high the stakes are but also the individual's susceptibility to test anxiety (see, for example, Gorjian, 2012) and thus varies from person to person.

Low-Stakes Testing

Low-stakes testing "involves the frequent use of evaluation instruments that have little impact on a student's course grade" (Oswego State University of New York, 2022). Despite superficial similarities to the high-stakes variety, low-stakes testing is much more concerned with helping a learner to learn than it is with determining their future prospects. As part of a teacher's everyday practice, it offers a chance for learners to practice certain items repeatedly, to space out the practice of those items over time, to allow the learner to understand where their strengths and weaknesses lie, and to show the teacher where remedial work might be required. Examples from the language classroom include frequent vocabulary testing, challenges to use recently acquired knowledge in a new context, requests for information previously studied, and review quizzes which do not form a substantial part of a student's grade. While fairness, validity, and accuracy are desirable in such testing opportunities, the desire for them does not dominate decisions about what is to be tested and how. Stress levels are intentionally low so that the learner has little to lose from trying out new knowledge or from recognising areas in which their knowledge is, as yet, insufficient.

Although low-stakes testing has been part of the teacher's tool-kit for a long time (Howatt, 1984), recent findings from the field of Mind, Brain, and Education (Sousa, 2010; Tokuhama-Espinosa, 2014) have led to renewed attention being paid to it. These findings are themselves part of a movement towards "evidenced-based teaching."

Evidence-Based Teaching

While the tendency of educators to "teach as they were taught" or to use "tried and trusted" methods in the classroom has long been noticed, it has been difficult to produce objective, scientific evidence that these ways of teaching are in fact the most efficacious and likely to lead to learning success. The main explanation for this lack of scientific evidence lies in the complexity of the teaching-learning process: there are so many variables in and around the classroom that it is impossible to control all extraneous variable in order to isolate and measure the effects of a particular teaching approach or methodology. This is why, for example, efforts to prove the superiority of Communicative Language Teaching over the Grammar Translation Method have so often ended in ambiguity and defeat (Natsir & Senjaya, 2014). Lack of convincing evidence may also explain why learners have, for generations, used study techniques (such as underlining, or reading their textbook repeatedly) which their teachers know to be inefficient (Dunlosky et al. 2013).

Recently, however, the ability to scan brains in the act of learning has provided new opportunities to examine effective ways of learning and teaching. The results, from cognitive neuroscience, are seldom surprising to experienced teachers, but do provide a solid, evidential guide to best practice. The increasing confidence of book titles such as *How we Learn* (Dehaene, 2020) and *Make it Stick: The Science of Successful Learning* (Brown, 2014) reflects growing assurance that our understanding of learning processes is becoming clearer and more complete.

In particular, neuroscientists have identified a number of strategies that lead to more effective learning. This list of strategies is taken from a guide (Pashler et al., 2007) produced by the U.S. Department of Education:

Spaced Repetition. Studying the same information at different times.

Interleaving. Studying a variety of topics one after another before returning to the initial topic.

Retrieval Practice. Recalling previously learned information.

Elaboration. Consciously connecting new information to things already known.

Concrete Examples. Connecting new knowledge to previous experiences.

Dual coding. Accessing the same information in various forms (visual, auditory, etc.)

While elements of all six of these strategies can be in play in low-stakes testing, an examination of the first three (Spaced Repetition, Interleaving, and Retrieval Practice) provides a convincing explanation of the efficacity of its use.

Spaced Repetition

Since Ebbinghaus (1885) pointed out that most information is lost to memory within a very short time of being learned, the search has been on for better mnemonics, better ways of ensuring that learners do not forget their lessons.

It is now understood that our brain tends to remember things that are salient to it (Sousa, 2010). An encounter that represents a threat to our survival is likely to be remembered very well, thanks to a release of dopamine, a neurochemical that enhances neuronal connections (LaLumiere, 2014). Similarly, a pleasant surprise, especially one that enhances our prospects of survival, is also likely to increase dopamine production and so be remembered. These bio-neural mechanisms are easy to understand in terms of species and organism survival but, unfortunately, difficult to reproduce in the classroom.

Fortunately, there is another factor that increases the salience of a particular experience or piece of information: encountering the same thing repeatedly. Repetition of new language ("Repeat after me") has long been an important feature of the language classroom. What neuroscience is adding to our understanding is that, to be effective, the repetition should be spaced out over time: not "repeat . . .again . . again . .again. . .OK, you've got it" but "repeat now, an hour later, after lunch, again tomorrow, a few days after that, next week, and so on." This specific pattern—repeating a few minutes later, an hour or so later, two days later and a week later—is recommended by Smolen et al. (2016).

Reviews (Carpenter at al., 2012; Cepeda et al., 2006) of studies comparing spaced repetition with its opposite ("massed practice") find overwhelming evidence that spaced repetition results in better long- and short-term retention in both laboratory and classroom settings.

Interleaving

If repetition is to be spaced, what should come between the repeated iterations? The answer seems to be "other topics." The human mind is primed for variety (Sousa, 2010): it seeks out difference. Again, this is most likely a survival technique, as, in the environment in which humans evolved, what was different could easily be fatal or, alternatively, could present an opportunity for enhanced survival. Consequently, large cognitive resources are devoted to recognising patterns in the environment and then paying attention to things that do not fit established patterns. As the new element becomes more familiar (i.e. becomes integrated into existing patterns, modifying them as necessary), the brain pays less and less attention to them as it seeks more novel experiences.

In the classroom, this leads to a limit to the length of time can pay attention to a single topic. After a while (10 – 20 minutes) they begin to give more attention to things other than the topic being studied. This is why neuroscientists recommending a variety of topics, a variety of activities and a variety of focuses for each lesson or study session.

Changing the topic, however, does not mean that learners will no longer process information about the previous topic. There is ample evidence that the brain unconsciously continues to look for patterns and differences during any spare time it has. If this happens during sleep, it is called dreaming; if the learner is awake, it is known as “mind-wandering.” Both are known to enhance both understanding and learning (Sousa, 2020).

The evidence that interleaving enhances learning and memory is summarised by Pan (2015) who highlights a study by Rohrer et al. (2015), which reported a 25% advantage on a final test for learners who were given practice mixing up two different types of math problems over classmates who practiced the two types of problems separately,

Retrieval Practice

Just like a photo album which, when opened and looked at, improves recall of the people and events depicted there, retrieving memories improves recall of those memories. In essence a memory is a group of neurons that fire together. The more they do this (fire together), the stronger the connection between them becomes, and so the memory becomes more firmly fixed. Each time, we retrieve a memory, we strengthen it.

Exercises that require learners to recall previously learnt material function as retrieval practice, strengthening the recall of that material. Further, by recalling the material in a different context to the one in which it was initially encountered, the learner is expanding the network of associations that material has for them. In other words, they are expanding the number and variety and neurons activated by the remembered material. A learner who encounters a new word may make efforts to memorise the word and basic elements of its meaning, but it is repeated retrieval of the word that both solidifies the memory of the world and provides nuance and extended range to the occasions when the learner might use the word (Karpicke, 2012; Pyc & Rawson, 2010).

Retrieval practice has other obvious advantages. It alerts learners to gaps in their knowledge and, if done as a low-stakes test, has the potential also to bring such gaps to the attention of the teacher.

Dunlosky et al. (2013) had students learn Swahili words and their English equivalents. The experimental group was told to quiz themselves on the words (a form of retrieval practice) in preparation for the test, whereas the control group were told to read and re-read the words. The experimental group out-performed its peers by over 50%.

Low-Stakes Testing, Spaced Repetition, Interleaving, and Retrieval Practice

Low-stakes testing draws on the strengths of each of these three study strategies. It allows the teacher to bring an item back to the attention of the learner at various intervals after it is initially encountered, by including it in subsequent tests. It allows for the mixing up of items taught in different lessons and lesson stages in a single test. It affords practice in recalling items—or not recalling them if the item has been forgotten, in which case further study is in order.

Furthermore, it allows teachers to add salience to certain items by repeating them more often than others as test items. This indicates to learners the relative importance the teacher attaches to these items, consolidates the more important ones in the learner's memory, and focuses the learner's (limited) attention where the teacher believes it belongs.

Because the stakes are not only low but also demonstrated to the learners to be low, test anxiety should not be a problem for learners taking these tests. They can focus on checking their understanding and recall of the test items rather than worrying about the consequences their test performance might have for their future. Attention remains where it should be in the classroom: on learning.

Additionally, low-stakes testing makes possible learning-friendly practices that would be unthinkable if the stakes were higher. Educational psychologists, for example, report that students learn best when they learn in a social context (Cozolino, 2014). This finding is reinforced by the discovery that the brain uses two different neural networks for "social learning" (learning from each other) and for "analytical learning" (the kind of analysis-and-memorisation activities familiar in most school settings), and that the social learning network is more effective than the analytical is (Lieberman, 2013). Yet, working with a classmate to understand and answer questions would simply not be acceptable in a high-stakes test. When the stakes are low, though, and the focus is on learning, consulting a classmate can be not only desirable but encouraged.

Two Examples of Low-Stakes Test

Vocabulary Test

As an English teacher, I am aware of the need for my learners to become familiar with, remember, and use more and more words. They themselves express the difficulties they have in understanding and using English in terms of their need for new words.

I am also aware that the approach to vocabulary adopted by most textbook-based lessons is insufficient to ensure retention and future use of the word. Typically, a new words is presented in the context of a particular textbook unit. It may be explained or exemplified. Its morphological, phonological, and orthographic features may be highlighted. It may even appear in a glossary at the end of the chapter or book or be quizzed in a "Unit Review." Then, however, attention moves on to the next chapter and the next, where the word will most probably not re-occur. Some enlightened textbook authors make a feature of "recycling vocabulary," ensuring that certain words occur in not one but two or three chapters. This is still, though, a long way from the "at least 12 encounters" that Nation (2014) says are needed in order to learn a new word.

In an attempt to remedy the situation, I begin each lesson with a low-stakes vocabulary test. The stakes are low because, though the tests are graded, scores on all the vocabulary tests combined (a total of some 700 individual items over a semester) account for only 10% of the learner's grade for the semester. Learners know the stakes are low because they are told this at the beginning of the semester, and also because the teacher's attitude is not judgmental ("You didn't do well on this test") but supportive ("We need to look again at this word").

Obviously, the tests provide an opportunity for retrieval practice. Spaced repetition is provided by including not only words from the previous lesson but a selection of words from all previous lessons. The fact that items from various lessons are mixed together in the test provides interleaving of semantic fields encountered when studying different topics.

The design of the test is intended to provide motivation for review of previously encountered vocabulary (spaced repetition and retrieval practice) and learners who do this are rewarded with good scores and a sense of progress. Not all students, though, are willing to review vocabulary between lessons. For these students, the test itself functions as an opportunity for spaced repetition and retrieval, as they encounter prompts for words they have been taught in the past.

Feedback on test answers is provided as quickly as possible. Agarawal and Bain. (2019) list immediate feedback as one of the elements of effective retrieval practice. Feedback on the vocabulary tests occurs in two ways: after the test papers have collected in, the teacher asks learners individually to answer orally the questions from the test, continuing until a correct answer is heard for each question; and, in anticipation of this, many learners consult their vocabulary notebooks as soon as they surrender their test paper to check on answers they were not sure of. The pleasure that learners experience on discovering (whether from another learner or from their notes) the correct answer to a question that has been puzzling them is the result of a dopamine release, the neurochemical that helps to fix things in memory (LaLumiere, 2014).

The fact that the testing procedure takes up a substantial part of the lesson (20 out of 90 minutes most days) communicates to the learners the seriousness they should accord to vocabulary learning and review. Additionally, the teacher's selection of words for each test and the frequency with which certain words are recycled provides salience to the words that are chosen, indicating that these words are especially important and useful, playing to the brain's tendency to commit more resources to remembering salient items.

A social element can be added to the test when it is clear that some learners are having difficulty with answers. A second learner can be assigned to help their struggling classmate. For the helper, providing an answer to a classmate affords another opportunity to retrieve a word from memory. For the helped, learning from a classmate activates the social learning network, which Lieberman (2013) has told us is a more powerful way of learning than facing alone a list of new words. For both helper and helped, solving a puzzle may result in a memory-consolidating shot of dopamine (LaLumiere, 2014).

Finally, if there are some questions which no one can answer, learners can be allowed to consult their notebooks, but not copy from them. They need to hold the answers in their mind long enough to close the notebook, put it away, pick up a pencil and write. The same item is then included in the test given in the next lesson, giving learners a chance to retrieve the memory trace.

It should be noted that all of these helping techniques—the immediate feedback, the help from a classmate, and the notebook consultation—serve to communicate that it is not their performance on the test but the chance to learn that is the most important part of this procedure. This further amplifies the message that the stakes for the test are low.

Facts-and-Figures Test

I recently had the chance to apply the principles of low-stakes testing in another context. This was not a language course, where skills and practice are at least as important as memory in determining outcomes, but an information-based course about life in various English-speaking countries, where remembering facts and figures about the country was the heart of the course.

The test consisted of a pile of note cards. After each lesson, I wrote on the note cards information from the lesson that I wanted learners to remember, with one piece of information on each card. A typical card would say "It's independence day is July 1" or "Many people celebrate Christmas on the beach."

In the next lesson, all cards from all previous lessons were used. One card was handed to each learner. The learner was to identify the country the information on the card referred to and go and stand in a corner of the classroom where that country's flag was displayed. Once all learners had chosen a country, each read the words from their card aloud. If the answer was correct, the card was discarded (until the next lesson). If not, the card was returned to the bottom of the pack. The second round began with a card from the top of the pack being distributed to each learner, and so on until all cards had been correctly identified with a flag and discarded.

No limits were placed on the means used to answer the questions. Some learners remembered answers from week to week, providing retrieval practice and spaced repetition for them. Others consulted classmates, adding a social learning aspects to the retrieval of information. Still others consulted their notes or even their smartphones, providing spaced repetition of the information, if not practice in retrieving it from their own brains.

What was evident in the classroom was a desire to find the answer. Learners wanted to solve the puzzle of the card in their hands. They were engaged with the search for knowledge. Engagement, focussed attention, is one of the pre-requisites for learning (Lodge & Harrison, 2019). Consolidation of the learning was provided by the dopamine release that was evident in their satisfaction at finding the answer (LaLumiere, 2014).

Interleaving of topics was achieved simply by shuffling the cards between lessons, so that topics covered recently were dealt alongside those further back in the course. Retrieval practice was evident in learners' answering specific questions they (or their less knowledgeable classmate) were dealt. Spaced repetition was available to learners for all questions as their classmates read their cards aloud during the checking phase of the test. The power of spaced repetition was demonstrated by the way in which learners, quite unconsciously, became more and more familiar with the flags of the various countries week by week.

A further aspect of this test should be noted: it was multi-modal and multi-sensory. Dual coding (expressing information in more than one way) was mentioned by Pasler et al. (2007) as one of the six main evidence-based learning strategies. Learners *read* the question on their card and *say it aloud* during checking; they *hear* classmates read their own cards; and they get to express their answer by *moving* to a certain part of the classroom. It used to be thought that each learner had a favoured learning style: visual, auditory, or kinesthetic (Walsh, 2011), but more recent research suggests that all able-bodied learners make use of all of these modes of learning and that the use of multiple modes allows each mode to reinforce the others (Riener & Willingham, 2010). So, the use of various of various learning modes is another strength of this kind of test.

As with the first example, it is the low-stakes nature of this test that allows the inclusion of elements, such as social learning and multi-modality, that support and encourage learning. In this case, learners are not even awarded a score for their performance. The focus is on finding the answer, on learning.

Conclusion

Although high-stakes and low-stakes tests may appear similar and, at times, the experience of taking them may feel similar, in fact they are very different. The purpose of a high-stakes test is to provide a snapshot of what the test-taker is capable of, so that decisions about the test-taker's future can be made. The purpose of a low-stakes test is to encourage and support learning, and only secondarily to provide feedback to the learner and teacher about areas that need extra support.

High-stakes tests are designed and implemented under stringent conditions designed to remove as many intervening variables as possible from the assessment of the test-taker's ability. The absence of these constraints in low-stakes testing allows for the introduction of elements that are known to enhance learning. The evidence supporting these elements is being provided by neurological and cognitive psychological studies into the very nature of learning.

As more such evidence becomes available, teachers should feel free to abandon a strict understanding of what a test should be and how it should be conducted in favour of evidence-based practices that use tests as just one of the strategies available to help students learn.

References

Agarawal, P. K., & Bain, P. M. (2019). *Powerful teaching: Unleash the science of learning*. Jossey-Bass.

Association of Language Testers in Europe. (1999). *Multilingual glossary of language testing terms.* Cambridge University Press.

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S.H.K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instructi*on. Educational Psychology Review, 24*, 369–378.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*, 354-380.

Cozolino, L. (2014). Attachment-based teaching: Creating a tribal classroom. W. W. Norton.

Dehaene, S. (2020). *How we learn: The science of education and brain.* Viking.

De Jong, N. H. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly, 15*(3), 237-254. https://doi.org/10.1080/15434303.2018.1477780

Dunlosky, J., Rawson, K. A., & Willingham, D. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. Psychological Science in the Public Interest, 14(1), 4–58. https://doi.org/10.1177/1529100612453266

Ebbinghaus, H. (1885). *Memory: A contribution to experimental psychology.* Dover.

Gorjian, B. (2012). Effects of learners' individual differences on test anxiety among the students majoring in translation in Islamic Azad University of Abadan. *Procedia Technology, 1*, 395–399. https://doi.org/10.1016/j.protcy.2012.02.090

Howatt, A. P. R. (1984). *A history of English language teaching.* Oxford University Press.

Lieberman, M. D. (2013). *Social: Why our brains are wired to connect.* Crown.

Jones, G. M., Jones, B. D., & Hargrove, T. (2003). *The unintended consequences of high-stakes testing.* Rowman and Littlefield.

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning.
Current Directions in Psychological Science, 21(3), 157-163.

LaLumiere, R. T. (2014). Dopamine and memory. In A. Meneses (Ed.), *Identification of neural markers accompanying memory*, Elsevir.

Lodge J. M., & Harrison W. J. (2019). The role of attention in learning in the digital age. *Yale Journal of Biological Medicine, 92*(1), 21-28.

Nation, P. (2014). How much input do you need to learn the most frequent 9,000 words? *Reading in a Foreign Language, 26*(2), 1–16. https://files.eric.ed.gov/fulltext/EJ1044345.pdf

Natsir, M., & Senjaya, D. (2014). Grammar translation method (GTM) versus communicative language teaching (CLT): A review of literature. *International Journal of Education and Literacy Studies, 2*(1), 58-62.

Oswego State University of New York. (2022). *Low-stakes testing.* Center for Excellence in Learning and Teaching. https://www.oswego.edu/celt/low-stakes-testing

Pashler, L., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., &

Metcalfe, J. (2007). Organizing instruction and study to improve student learning. Institute of Education Sciences. https://files.eric.ed.gov/fulltext/ED498555.pdf

Pan, S. C. (2015, 4 January). The interleaving effect: Mixing it up boosts learning. *Scientific American.* https://www.scientificamerican.com/article/the-interleaving-effect-mixing-it-up-boosts-learning/

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*(6002), 335. https://doi.org/10.1126/science.1191465

Riener, C., & Willingham, D. (2010). The myth of learning styles. *Change: The Magazine of Higher Learning, 42*(5), 32-35. https://doi.org/10.1080/00091383.2010.503139

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology, 107*(3), 900–908. https://doi.org/10.1037/edu0000001

Smolen, P., Zhang, Y., & Byrne, J. H. (2016). *The right time to learn: Mechanisms and optimization of spaced learning. Nature Reviews Neuroscience, 17*(2), 77–88. https://doi.org/*10.1038/nrn.2015.18.*

Sousa, D. A. (2010). *Mind, brain, and education: Neuroscience implications for the classroom.* Solution Tree.

Tokuhama-Espinosa, T. (2014). *Making classrooms better: 50 practical applications of mind, brain, and education science.* W. W. Norton.

Walsh, B. E. (2012). *Vak self-audit: Visual, auditory, and kinesthetic communication and learning styles: Exploring patterns of how you interact and learn.* Walsh Seminars.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology, 18(5), 459–482.* https://doi.org/*10.1002/cne.920180503*

Young, J. W., Young, Y. S., & Ockey, G. J. (2013). *Guidelines for best test development practices to ensure validity and fairness for international English language proficiency assessments.* Educational Testing Services.