Testing

I think we all have had the experience that we did not do too well passing the school test but knew afterwards the subject inside out. The test did it.

Testing has almost disappeared from the classroom, with the exception of exams – and then it is too late. Evidence suggests that regular testing enhances memory and retrieval of the learnt items.

Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention

Taking a memory test not only assesses what one knows, but also enhances later retention, a phenomenon known as the testing effect. The authors studied this effect with educationally relevant materials and investigated whether testing facilitates learning only because tests offer an opportunity to restudy material. In two experiments, students studied prose passages and took one or three immediate free-recall tests, without feedback, or restudied the material the same number of times as the students who received tests. Students then took a final retention test 5 min, 2 days, or 1 week later. When the final test was given after 5 min, repeated studying improved recall relative to repeated testing. However, on the delayed tests, prior testing produced substantially greater retention than studying, even though repeated studying increased students' confidence in their ability to remember the material. Testing is a powerful means of improving learning, not just assessing it.
(I. Roediger, H. L. & J. D. Karpicke, 2006)

ASSESSMENT AND TESTING

In this brief article, the author discusses the relationship between language testing and the other sub-disciplines of applied linguistics and also the relationship, as she sees it, between testing and assessment. The article starts with a brief

exploration of the term ‘applied linguistics’ and then goes on to discuss the role of language testing within this discipline, the relationship between testing and teaching, and the relationship between testing and assessment. The second part of the article mentions some areas of current concern to testers and discusses in more detail recent advances in the areas of performance testing, alternative assessment, and computer assessment. One of her aims in this article is to argue that the skills involved in language testing are necessary not only for those constructing all kinds of language proficiency assessments, but also for those other applied linguists who use tests or other elicitation techniques to help them gather language data for research.
(Clapham, 2000)

Testing of second language pragmatics:
Past and future

Testing of second language pragmatic competence is an underexplored but growing area of second language assessment. Tests have focused on assessing learners’ sociopragmatic and pragmalinguistic abilities but the speech act framework informing most current productive testing instruments in interlanguage pragmatics has been criticized for under-representing the construct. In particular, the assessment of learners’ ability to produce extended monologic and dialogic discourse is a missing component in existing assessments. This paper reviews existing tests and argues for a discursive re-orientation of pragmatics tests. Suggestions for tasks and scoring approaches to assess discursive abilities while maintaining practicality are provided, and the problematicity of native speaker benchmarking is discussed.
(Roever, 2011)

Size and strength: do we need both to measure vocabulary knowledge?

This article describes the development and validation of a test of vocabulary size and strength. A model for administering the test in computer adaptive mode is also proposed. The study has implications both for the design and delivery of this test as well as for theories of vocabulary acquisition.
(Laufer, Elder, Hill, & Congdon, 2004)

The Yes/No test as a measure of receptive vocabulary knowledge

Performance on the Yes/No test was assessed as a predictor of scores on the Vocabulary Levels Test (VLT), a standard test of receptive second language (L2) vocabulary knowledge. The results indicate the Yes/No test is a valid measure of the type of L2 vocabulary knowledge assessed by the VLT, with implications for classroom application.
(Mochida & Harrington, 2006)

Examining the Yes/No vocabulary test: some methodological issues in theory and practice.

This article evaluates the characteristics of the Yes/No test as a measure for receptive vocabulary size in second language (L2). This evaluation was conducted both on theoretical grounds as well as on the basis of a large corpus of data collected with French learners of Dutch. The study focuses on the internal qualities of the format in comparison with other more classical test formats. The central issue of determining a meaningful test score is addressed by providing a theoretical framework distinguishing discrete from continuous models. Correction formulae based on the discrete approach are shown to differ when applied to the Yes/No test in comparison with Multiple Choice (MC) or True/False formats. Correction formulae based

on the continuous approach take the response bias into account but certain underlying assumptions need to be validated. It is shown that both correction schemes display several shortcomings and that most of the data relative to the reliability of the Yes/No test presented in the literature are overestimated. Finally, several future research options are proposed in order to attain a straightforward but reliable and valid instrument for measuring receptive vocabulary size.
(Beeckmans, Eyckmans, Janssens, Dufranne, & Van de Velde, 2001)

A framework for second language vocabulary assessment

Vocabulary tests are used for a wide range of instructional and research purposes but we lack a comprehensive basis for evaluating the current instruments or developing new lexical measures for the future. This article presents a framework that takes as its starting point an analysis of test purpose and then shows how purpose can be systematically related to test design. The link between the two is based on three considerations which derive from Messick's (1989) validation theory: construct definition, performance summary and reporting, and test presentation. The components of the framework are illustrated throughout by reference to eight well-known vocabulary measures; for each one there is a description of its design and an analysis of its purpose. It is argued that the way forward for vocabulary assessment is to take account of test purposes in the design and validation of tests, as well as considering an interactionalist approach to construct definition. This means that a vocabulary test should require learners to perform tasks under contextual constraints that are relevant to the inferences to be made about their lexical ability.
(Read & Chapelle, 2001)

Modern language testing at the turn of the century: assuring that what we count counts.

In the past twenty years, language testing research and practice have witnessed the refinement of a rich variety of approaches and tools for research and development, along with a broadening of philosophical perspectives and the kinds of research questions that are being investigated. While this research has deepened our understanding of the factors and processes that affect performance on language tests, as well as of the consequences and ethics of test use, it has also revealed lacunae in our knowledge, and pointed to new areas for research. This article reviews developments in language testing research and practice over the past twenty years, and suggests some future directions in the areas of professionalizing the field and validation research. It is argued that concerns for ethical conduct must be grounded in valid test use, so that professionalization and validation research are inseparable. Thus, the way forward lies in a strong programme of validation that includes considerations of ethical test use, both as a paradigm for research and as a practical procedure for quality control in the design, development and use of language tests.
(Bachman, 200)

Testing the testing effect in the classroom

“Laboratory studies show that taking a test on studied material promotes subsequent learning and retention of that material on a final test (termed the testing effect). Educational research has virtually ignored testing as a technique to improve classroom learning. We investigated the testing effect in a college course. Students took weekly quizzes followed by multiple choice criterial tests (unit tests and a cumulative final). Weekly quizzes included multiple choice or short answer questions, after which feedback was provided. As an exposure control, in some weeks students were presented target material for additional reading. Quizzing, but not additional reading,

improved performance on the criterial tests relative to material not targeted by quizzes. Further, short answer quizzes produced more robust benefits than multiple choice quizzes. This pattern converges with laboratory findings showing that recall tests are more beneficial than recognition tests for subsequent memory performance. We conclude that in the classroom testing can be used to promote learning, not just to evaluate learning”.
(McDaniel, Anderson, Derbish, & Morrisette, 2007)

Why tests appear to prevent forgetting: A distribution-based bifurcation model

Retrieving information from memory produces more learning than does being presented with the same information, and the benefits of such retrieval appear to grow as the delay before a final recall test grows longer. Recall tests, however, measure the number of items that are above a recall threshold, not memory strength per se. According to the model proposed in this paper, tests without feedback produce bifurcated item distributions: Retrieved items become stronger, but non-retrieved items remain weak, resulting in a gap between the two classes of items. Restudying items, on the other hand, strengthens all items, though to a lesser degree than does retrieval. These differing outcomes can make tested items appear to be forgotten more slowly than are restudied items—even if all items are forgotten at the same rate—because the test-induced bifurcation leaves items either well above or well below threshold.
(Kornell, Bjork, & Garcia, 2011)

Retrieval mode distinguishes the testing effect from the generation effect

A series of four experiments examined the effects of generation vs. retrieval practice on subsequent retention. Subjects were

first exposed to a list of target words. Then the subjects were shown the targets again intact for Read trials or they were shown fragments of the targets. Subjects in Generate conditions were told to complete the fragments with the first word that came to mind while subjects in Recall conditions were told to use the fragments as retrieval cues to recall words that occurred in the first part of the experiment. The instruction manipulated retrieval mode—the Recall condition involved intentional retrieval while the Generate condition involved incidental retrieval. On a subsequent test of free recall or recognition, initial recall produced better retention than initial generation. Both generation and retrieval practice disrupted retention of order information, but retrieval enhanced retention of item-specific information to a greater extent than generation. There is a distinction between the testing effect and the generation effect and the distinction originates from retrieval mode. Intentional retrieval produces greater subsequent retention than generating targets under incidental retrieval instructions.
(Karpicke & Zaromb, 2010)

Interacting in pairs in a test of oral proficiency: Co-constructing a better performance.

This study, framed within sociocultural theory, examines the interaction of adult ESL test-takers in two tests of oral proficiency: one in which they interacted with an examiner (the individual format) and one in which they interacted with another student (the paired format). The data for the eight pairs in this study were drawn from a larger study comparing the two test formats in the context of high-stakes exit testing from an Academic Preparation Program at a large Canadian university. All of the test-takers participated in both test formats involving a discussion with comparable speaking prompts. The findings from the quantitative analyses show that overall the test-takers performed better in the paired format in that their scores were on average higher than when they interacted with an examiner.

Qualitative analysis of the test-takers' speaking indicates that the differences in performance in the two test formats were more marked than the scores suggest. When test- takers interacted with other students in the paired test, the interaction was much more complex and revealed the co-construction of a more linguistically demanding performance than did the interaction between examiners and students. The paired testing format resulted in more interaction, negotiation of meaning, consideration of the interlocutor and more complex output. Among the implications for test theory and practice is the need to account for the joint construction of performance in a speaking test in both construct definitions and rating scales.
(Brooks, 2009)

Repeated retrieval during learning is the key to long-term retention

Tests not only measure the contents of memory, they can also enhance learning and long-term retention. The authors report two experiments inspired by Tulving's (1967) pioneering work on the effects of testing on multitrial free recall. Subjects learned lists of words across multiple study and test trials and took a final recall test 1 week after learning. In Experiment 1, repeated testing during learning enhanced retention relative to repeated studying, although alternating study and test trials produced the best retention. In Experiment 2, recalled items were dropped from further studying or further testing to investigate how different types of practice affect retention. Repeated study of previously recalled items did not benefit retention relative to dropping those items from further study. However, repeated recall of previously recalled items enhanced retention by more than 100% relative to dropping those items from further testing. Repeated retrieval of information is the key to long-term retention.
(Karpicke & Roediger III, 2007)

Learners’ choices and beliefs about selftesting.

Students have to make scores of practical decisions when they study. The authors investigated the effectiveness of, and beliefs underlying, one such practical decision: the decision to test oneself while studying. Using a flashcards-like procedure, participants studied lists of word pairs. On the second of two study trials, participants either saw the entire pair again (pair mode) or saw the cue and attempted to generate the target (test mode). Participants were asked either to rate the effectiveness of each study mode (Experiment 1) or to choose between the two modes (Experiment 2). The results demonstrated a mismatch between metacognitive beliefs and study choices: Participants (incorrectly) judged that the pair mode resulted in the most learning, but chose the test mode most frequently. A post-experimental questionnaire suggested that self-testing was motivated by a desire to diagnose learning rather than a desire to improve learning.
(Kornell & Son, 2009)

The last sentence describes a common recurrence. I think it has to do with the fact that learners generally do not know that testing is also a learning tool. They encounter testing almost exclusively as an assessment tool. They have to be told. Maybe they use it then more often.
JH

The Power of Testing Memory: Basic Research and Implications for Educational Practice

A powerful way of improving one’s memory for material is to be tested on that material. Tests enhance later retention more than additional study of the material, even when tests are given without feedback. This surprising phenomenon is called the testing effect, and although it has been studied by cognitive psychologists sporadically over the years, today there is a renewed effort to learn why testing is effective and to apply

testing in educational settings. In this article, the authors selectively review laboratory studies that reveal the power of testing in improving retention and then turn to studies that demonstrate the basic effects in educational settings. They also consider the related concepts of dynamic testing and formative assessment as other means of using tests to improve learning. Finally, they consider some negative consequences of testing that may occur in certain circumstances, though these negative effects are often small and do not cancel out the large positive effects of testing. Frequent testing in the classroom may boost educational achievement at all levels of education.
(H. L. Roediger & J. D. Karpicke, 2006)

The critical role of retrieval practice in long-term retention

Learning is usually thought to occur during episodes of studying, whereas retrieval of information on testing simply serves to assess what was learned. We review research that contradicts this traditional view by demonstrating that retrieval practice is actually a powerful nemonic enhancer, often producing large gains in long-term retention relative to repeated studying. Retrieval practice is often effective even without feedback (i.e. giving the correct answer), but feedback enhances the benefits of testing. In addition, retrieval practice promotes the acquisition of knowledge that can be flexibly retrieved and transferred to different contexts. The power of retrieval practice in consolidating memories has important implications for both the study of memory and its application to educational practice.
(Roediger & Butler, 2011)

Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention

Five experiments were conducted to examine how unsuccessful retrieval influences learning and subsequent memory. We used

a cued-recall paradigm that produces many unsuccessful retrieval attempts (followed by feedback) and allows comparisons to be made between later memory for these trials and trials that only required reading or studying the pairs. On read trials participants studied cue–target pairs that were either weakly associated (DOOR–EXIT) or unrelated but identical in length (DOOR–SHOE). On test trials participants were given only the cue (either without [Exps. 1–3] or with [Exps. 4–5] prior experience with the pair items) and asked to guess the target which they were told was either semantically related or identical in length to the cue; then they received the correct cue–target pair to study. Unsuccessful retrieval attempts (i.e., guessing) relative to studying benefited retention for weakly associated pairs but impaired retention for unrelated pairs. This pattern of results occurred regardless of study duration (Experiments 1A and 1B), level of processing of the cue (Experiment 2), whether relatedness was manipulated between or within subjects (Experiment 5), and when guessing involved episodic as opposed to semantic retrieval (Experiments 4 and 5). However, this pattern was partly mediated by the ability to retrieve incorrect guesses during a final cued-recall test which may provide a link between the cue and target (Experiment 3). The current study demonstrates that unsuccessful retrieval attempts with immediate feedback not only enhance, but also can impair learning. This effect is robust and depends on elaborative semantic activation related to the answer and the effectiveness of incorrect guesses as mediating cues.
(Knight, Hunter Ball, Brewer, DeWitt, & Marsh, 2012)

Retrieval effort improves memory and metamemory in the face of misinformation

Retrieval demand, as implemented through test format and retrieval instructions, was varied across two misinformation experiments. Our goal was to examine whether increasing retrieval demand would improve the relationship between confidence and memory performance, and thereby reduce

misinformation susceptibility. We hypothesized that improving the relationship between confidence and memory performance would improve controlled processes at retrieval. That is, when confidence and memory performance were well calibrated, participants would be able to withhold incorrect responses if given the opportunity. To examine the relationship between memory retention, confidence, and controlled withholding, we compared older and younger adults' performance on a forced memory test, where participants could not withhold responses, and on a free test, where participants were encouraged to withhold responses. Confidence judgments were collected after forced responding. Retrieval demand was manipulated indirectly through type of test (cued recall vs. recognition) and directly through retrieval instructions. The results demonstrated that increasing retrieval demands improved memory retention, metamemorial monitoring and effective withholding. This was particularly pronounced when participants received misleading information. Finally, older adults required explicit direction to effectively monitor memory and institute successful controlled withholding.
(Bulevich & Thomas, 2012)

Separate mnemonic effects of retrieval practice and elaborative encoding

Does retrieval practice produce learning because it is an especially effective way to induce elaborative encoding? Four experiments examined this question. Subjects learned word pairs across alternating study and recall periods, and once an item was recalled it was dropped from further practice, repeatedly studied, or repeatedly retrieved on repeated recall trials. In elaborative study conditions, subjects used an imagery-based keyword method (Experiments 1–2) or a verbal elaboration method (Experiment 3) to encode items during repeated study trials. On a criterial test 1 week after the initial learning phase, repeated retrieval produced better long-term retention than repeated study even under elaborative study

conditions. Elaborative studying improved initial encoding when it occurred prior to the first correct recall of an item, but while repeated retrieval enhanced long-term retention, elaboration produced no measurable learning when it occurred after successful retrieval. Experiment 4 used identical item word pairs (e.g., castle–castle) to reduce or eliminate verbal elaboration, and robust effects of repeated retrieval were still observed with these materials. Retrieval practice likely produces learning by virtue of mechanisms other than elaboration.
(Karpicke & Smith, 2012)

Semantic Information Activated During Retrieval Contributes to Later Retention: Support for the Mediator Effectiveness Hypothesis of the Testing Effect

Previous research has proposed that tests enhance retention more than do restudy opportunities because they promote the effectiveness of mediating information—that is, a word or concept that links a cue to a target (Pyc & Rawson, 2010). Although testing has been shown to promote retention of mediating information that participants were asked to generate, it is unknown what type of mediators are spontaneously activated during testing and how these contribute to later retention. In the current study, participants learned cue–target pairs through testing (e.g., *Mother:* _____) or restudying (e.g., *Mother: Child*) and were later tested on these items in addition to a never-before-presented item that was strongly associated with the cue (e.g., *Father*)—that is, the *semantic mediator*. Compared with participants who learned the items through restudying, those who learned the items through testing exhibited higher false alarm rates to semantic mediators on a final recognition test (Experiment 1) and were also more likely to recall the correct target from the semantic mediator on a final cued recall test (Experiment 2). These results support the mediator effectiveness hypothesis and demonstrate that semantically related information may be 1 type of natural mediator that is activated during testing.

(Carpenter, 2011)

How good is your test?

This article reports on a study of the validity and reliability of tests administered in an EFL university setting. The study addresses the question of how well face validity reflects more objective measures of the quality of a test, such as predictive validity and reliability. According to some researchers, face validity, defined as the surface credibility or public acceptability of a test, has no theoretical basis since it is based on the subjective perceptions of stakeholders such as teachers and students. However, due to lack of time or resources, or due to a perceived lack of competence, practitioners tend to rely on the 'appeal' of language tests, rather than seek empirical evidence. This article describes several ways of evaluating achievement tests, comparing their results in order to shed light on what measures can and should be taken to ensure that achievement tests accomplish their purposes.
(Küçük & Walters, 2009)

Bachman, L. F. (200). Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing, 17*(1), 1-42.

Beeckmans, R. , Eyckmans, J., Janssens, V. , Dufranne, M. , & Van de Velde, H. (2001). Examining the Yes/No vocabulary test: some methodological issues in theory and practice. . *Language Testing, 18*(3), 235-274.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. . *Language Testing, 26*(3), 341-366.

Bulevich, J. B., & Thomas, A. K. (2012). Retrieval effort improves memory and metamemory in the face of misinformation. . *Journal of Memory & Language, 67*, 45-58.

Carpenter, S. K. (2011). Semantic Information Activated During Retrieval Contributes to Later Retention: Support for the Mediator Effectiveness Hypothesis of the Testing Effect. . *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(6), 1547-1552.

Clapham, C. (2000). ASSESSMENT AND TESTING. *Annual Review of Applied Linguistics, 20*, 147-161.

Karpicke, J. D. , & Roediger III, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. . *Journal of Memory and Language, 57*, 151-162.

Karpicke, J. D. , & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory & Language*, 1-13.

Karpicke, J. D. , & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *62*, 227-239.

Knight, J. B. , Hunter Ball, B. , Brewer, G. A. , DeWitt, M. R. , & Marsh, R. L. . (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. . *Journal of Memory & Language, 66*, 731-746.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. . *Journal of Memory and Language, 65*, 85-97.

Kornell, N., & Son, L. K. . (2009). Learners' choices and beliefs about selftesting. . *Memory, 17*, 493-501.

Küçük, F., & Walters, J. (2009). How good is your test? . *ELT Journal, 93*(4), 332-341.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: do we need both to measure vocabulary knowledge? . *Language Testing, 21*, 202.

McDaniel, M. A., Anderson, G, L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*(4/5), 494-513.

Messick, S. (Ed.). (1989). *Educational Measurement* (3rd ed.). New York: MacMillan.

Mochida, K., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. . *Language Testing 23*, 73.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. . *Science, 330*, 335.

Read, J., & Chapelle, Carol A. (2001). A framework for second language vocabulary assessment. . *Language Testing, 18*(1), 1-32.

Roediger, H. L., & Butler, A. C. . (2011). The critical role of retrieval practice in long-term retention. . *Trends in Cognitive Sciences, 15*, 20-27.

Roediger, H. L., & Karpicke, J. D. . (2006). The power of testing memory: Basic research and implications for educational practice. . *Perspectives on Psychological Science, 1*, 181-210.

Roediger, III, H. L., & Karpicke, J. D. (2006). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science, 17*(3), 249.

Roever, C. (2011). Testing of second language pragmatics: Past and future. *Language Testing, 28*(4), 463-481.