LANGUAGE
TESTING

Special Issue

Repeated test-taking and longitudinal test score analysis

Language Testing
2020, Vol. 37(4) 475–481
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0265532220934202
journals.sagepub.com/home/ltj
SAGE

Tony Green
University of Bedfordshire, UK

Alistair Van Moere
MetaMetrics Inc, USA

Keywords
Assessment, equity, longitudinal, practice test, repeated testing, retest, test familiarity

This Special Issue of *Language Testing* is devoted to research into repeated test taking and associated issues. Although repeated test taking is beginning to attract increased research attention, it is far from a new phenomenon. According to Elman (2000), most of those entering for the highly competitive civil service examinations in Qing era China repeatedly entered and repeatedly failed: "For most, persistence. . . was a way of life" (p. 147). It was also regarded as a sign of strength of character and rectitude. Zhang Jian, winner of first place in the palace examinations (the highest level of the system) in 1894, was celebrated as the embodiment of resoluteness. His eventual triumph only came after almost 20 years of repeated failure.

Although examination success may represent the culmination of many years of study, many repeating candidates have been frustrated to find that the value of their cumulative investment in learning, as expressed by their test scores, may fall as well as rise over time. In his classic paper on the statistics of examinations, Edgeworth (1888) observed that "It is said to be a frequent occurrence at some of our civil service examinations that a candidate, when examined for the second time, after a year's hopeful study of a subject, obtains fewer marks in it than he had obtained at the first examination" (p. 606).

Edgeworth understood this element of luck in examination results as a problem of measurement error, attributable to a wide range of factors. He also saw that taking multiple tests would reduce the relative impact of error:

> let not the aspirant to academic honours be cast down by a first disappointment. Let him press on to higher and perhaps less aleatory examinations. He is playing a game in which superior skill must make itself felt in the long run. "Only go on long enough," as Dr. Venn says in his Logic of Chance, and the bias in favour of merit will be brought out. (1888, p. 616)

Corresponding author:
Alistair Van Moere, MetaMetrics Inc, 1000 Park Forty Plaza, Durham, NC 27519, USA.
Email: avanmoere@lexile.com

Repeated testing, like adding to the number of items on a single test, is a means of increasing the number of observations and so reducing the relative contribution of error to the results.

Candidates for civil service examinations or university entrance followed the principle of "try and try again" to gain their reward, but in the era of mass education, efficiency, and accountability, governments have made increasing use of tests as tools for technocratic management. They set out to monitor change and compare outcomes, hoping to use results to propagate the most successful educational practices. A leading voice in the drive for educational efficiency, Thorndike (1912, p. 290) set out his progressive vision for a new science of education to support these objectives: "education concerns the production and prevention of changes in human beings; and a science of education must identify these changes, compare them, and relate them to their causes. To do this it must measure them." Measurement of outcomes using common units would allow managers to identify and promote best practice, giving rise to a more effective education system.

Although more pluralistic, in allowing for a variation of objectives and trajectories in language learning, tools such as the Common European Framework for Languages (CEFR) embody a similar vision to Thorndike's. Stakeholders are attracted by the common basis that shared frameworks seem to offer for monitoring a learner's progress in functional terms. As the papers in this issue make clear, however, it is challenging even to monitor individual language learning progress through repeated measurement with parallel forms of the same instrument. The added difficulties associated with relating results obtained from different assessments suggest that any such attempt will be imprecise at best and potentially misleading for score users (Green, 2017).

Repeated testing has also long been recognized as an aid to retention and learning (Abbott, 1909; Roediger & Butler, 2011). The benefit of repeatedly recalling information to carry out test tasks can be leveraged when accompanied by information on the extent to which test performance has satisfied the criteria for success. If this information is provided in a form that learners are able to use to build their knowledge or skills, it contributes to a process of feedback that enhances learning (Boyd et al., 2019). Conversely, there are many equally well-rehearsed arguments warning that a reliance on testing to monitor and audit learning breeds anxiety, leads to a narrowing of the curriculum, and encourages teachers and students to restrict what is taught and learnt to what is tested (Green, 2019; Herbert, 1889).

As a research tool, repeated testing has inevitably featured in second language acquisition and pedagogic research that investigates the effects of interventions or background variables on learning. The success of such studies is generally limited by small sample sizes, which restrict both the number of factors that can be investigated and the generalizability of the results. Their size also leaves them more vulnerable to the effects of measurement error. Such studies tend to involve pre-post designs, over relatively brief periods, making them insensitive to longer term development and the dynamic interplay of factors involved in language learning. This led Ortega and Byrnes (2009) to complain a decade ago that, "after some 40 years of disciplinary history, we know little about the longitudinal pace and pattern of development in second language and literacy" (p. 5).

National or international testing programs, which are the focus of this Special Issue, could provide a complementary longitudinal perspective (Isbell et al., 2019). On the one hand, repeated test taking on such tests does involve very substantial numbers of language learners and often occurs over extended periods, spanning a school or college career, or even longer. On the other hand, it cannot offer researchers the same degree of control over sampling or the focus of the tasks found in experimental studies. Relatively little data is generally collected about the test takers' backgrounds, or what they do between testing occasions.

What is radically different in the digital age is our capacity to collect and store large quantities of language assessment data. When assessments are conducted digitally, it becomes very much easier to keep track of individual and group test performance over time. Automated scoring means that information can quickly be shared with the learner or test taker and is not limited to whether a response was correct, or to a numeric score. Automated writing evaluation software can, increasingly, offer qualitative feedback, suggest revisions and allow for multiple submissions to support improved performance. This enhances the potential for repeated test taking to inform guidance and direction in the language learning process.

Research initiatives such as the English Profile (https://www.englishprofile.org/) are able to take advantage of language corpora derived from test performances to seek new insights into language learning, with repeated test taking offering a longitudinal dimension. In addition, new technology affords greater access to the processes involved in responding to tasks (whether designed as learning tasks or test tasks). It is now straightforward not only to collect responses, but to record the length of time taken to choose or compose them; the changes made as a text is written; the direction of the respondent's gaze as they carry out a reading task. Where learners participate in virtual learning environments, it is becoming easier to capture information about their backgrounds and learning behaviours as well as their performances over time. Changes in many aspects of language performance and learning behaviours are thus opened to investigation and this offers promising new avenues for longitudinal investigation.

The papers

With access to more complex data on repeating test takers, there is a need for sophisticated analysis techniques. A raft of issues further complicates the task of the analyst including sampling, attrition, missing data, irregularly spaced testing occasions, and correlations between measures. Several papers in this special issue offer a methodological contribution. Van Moere and Hanlon sketched out how Bayesian statistics may be used to limit the impact of the error inherent in single test administrations when monitoring progress over time. Illustrating the point with data from a large-scale testing program and from a classroom setting, they showed how weighting data over a series of test administrations can provide a more credible picture both of a learner's developmental trajectory and of their current status than that provided by the most recent test event alone.

Demonstrating an alternative method to Bayesian modeling, Cho and Blood investigated a multilevel modelling approach to measure young EFL learners' score change on TOEFL® Primary™. They analysed the effects of elapsed time between tests taken, showing how rate of change among test takers varies depending on the initial test score, test taker age, and test level difficulty. Cho and Blood pointed to unexpectedly large score changes for individuals between test occasions, which they partially attributed to “fleeting attention span and other developmental limitations” (p. xx) in this young learner population; this is the kind of “score fluctuation” (p. xx) that Van Moere and Hanlon attempted to mitigate through Bayesian modeling.

Keeping with the theme of measuring the effects of dynamic changes over time, Lin and Chen investigated whether duration between test attempts was associated with changes in dimensions of language quality. The authors analysed the performance of students who sat three administrations of a writing exam over periods ranging from three to six months, including the time elapsed between test administrations, and the proficiency of the test takers. The findings indicated that test scores were relatively stable across multiple attempts, but that the most significant changes observed over test events were lexical features; moreover, these improvements were sustained over both shorter (one to two months) and longer (three to six months) periods of time.

Echoing the experience with the Chinese Civil Service examinations, Knoch, Huisman, Kong, Elder, and McKenna investigated repeaters steadfastly attempting to gain the scores they needed to satisfy Australian immigration requirements by taking the Pearson Test of English-Academic up to 22 times. They found evidence of learners changing their approaches over time, seeking help from experienced friends, test preparation programs and tutors, and moving between test familiarization activities (such as practicing test tasks), test-wise score boosting tactics (attempting to speak without pausing, or changing the quality of their voices to increase their speaking scores), and efforts at general language improvement (including practising reading and listening in real-world contexts).

Finally, Kim, Barron, Sinclair, and Jang used latent growth curve modeling (LGCM) to investigate how monolingualism or multilingualism in students’ homes is associated with their literacy development as they progress through school. Longitudinal data from tests taken at approximately 9, 12, and 16 years of age revealed students’ developmental trajectories, while a covariate in the LGCM model grouped these students according to whether their home environments were monolingual, multilingual, or shifting from one to the other over time. Their findings showed that students whose home language environment shifted across time exhibited lower literacy performance compared to their peers in earlier grades, but higher performance over time.

Emergent themes and directions

We see three main themes emerging from these papers which impact the validity of test score interpretations. The first theme asks the following question: Does repeated testing impact the interpretation of test scores? Concern arises because learners and test providers may view repeated testing from very different perspectives, which in turn may affect their behaviours and their whole approach to the assessment paradigm. Particularly in high-stakes contexts, for many learners the primary objective is to pass the test, as it is required for immigration or university entrance; proficiency in the target language may become a secondary consideration. For those learners whose chief goal is to pass the test rather than acquire a necessary standard in the target language, their behaviours may include intensive test preparation, acquiring or anticipating item bank content, or otherwise treating the test as a code to be cracked through test-wise techniques (as discussed by Knoch et al., in this issue). Learners (and instructors of intensive preparation courses) may use repeated test taking as an opportunity to identify and exploit features of the test design with the intention of maximizing scores rather than improving target abilities.

To meet this threat, test providers are obliged to put expensive controls in place to counter the effects of repeated test taking. These controls include, for example: constant development of new items; protection of item bank security; limiting the number of test windows per year; and scoring protocols that work against formulaic or template responses, especially in writing and speaking tasks. In other words, test providers constrain test design and take actions to counter the very behaviours that would seem to be encouraged by repeated testing. Defence against repeated test taking requires a level of investment that only the better-resourced test providers can manage.

Closely related to this, the second theme concerns equity and access to repeated test-taking opportunities. What does it say about test score interpretation if some learners, owing to time and financial resources, can repeat the test as many times as they wish, but others cannot? Again, learners and test providers have different perspectives on this. For learners, repeated testing may represent a form of intense test familiarization and a chance to reduce test-taking anxiety and refine their preparation strategies. If they manage to exploit weaknesses in the test design or benefit from measurement error by earning higher scores, they may secure access to opportunities at the expense of their more proficient, but under-resourced, peers (see Van Moere & Hanlon, this issue).

In contrast, test providers clearly have a commercial incentive to offer repeated testing, as it increases revenue and also makes their test more attractive to learners over competing tests that do not allow frequent repeated testing. With computerized testing, large time banks, and sufficient capacity in professional test centers, there is the potential for learners to retake high-stakes tests two or three times a month. A recent trend towards "super-scoring" (i.e., when learners can take several tests and choose the best sub-score from each attempt to use in their total reported score) exacerbates the problem, as it provides even more incentive for learners to retest; whereas test providers may portray this approach as a fair way of giving learners the benefit of the doubt, it is advantageous only for learners who are able to retake expensive, time-consuming tests many times.

The following question arises: Are there more equitable ways to deal with repeated testing than are currently being employed? Van Moere and Hanlon proposed one approach, which is to aggregate repeated test scores, with more recent test scores carrying greater weight. But there could be other approaches (such as looking at the *trajectory* of an individual's scores over time and test occasions) that provide a balance between equity for learners while also preserving commercial interests. Many test providers have supported substantial research into understanding test preparation, test practice, and washback (e.g., Chappell et al., 2019; Gu & Xi, 2015; Yu et al., 2017), but there remains a need for a concerted research focus on equity in the context of repeated testing.

The final common theme across this collection of papers is a research methodology perspective which concerns the dynamic nature of so many variables that can now be investigated and modeled. Owing to sensitivity in data collection and modeling techniques, papers in this issue evaluated a wide range of variables which are both static (e.g., learner proficiency at the time of test taking) and dynamic (e.g., a change in learner proficiency between the first test and the second test). Researchers in this issue investigated, among other variables, duration between test events, different practice effects, differential growth in skills, contextual or environmental changes, and patterns of change on test performance. Moreover, when such data has been collected, multilevel modelling can be employed, which allows for within-person dependency (see in this issue: Kim et al.; Cho & Blood). Thus, rather than treating each learner's data as an independent snapshot, a learner's repeated measures are nested within that individual.

These avenues of investigation point to the complex dynamics that occur between tests and during them. How long should learners wait between tests? How should they best prepare? In which skills can they expect to make score gains? And, from the test provider's perspective: How are learning gains best tracked? To what extent does the test format encourage repeaters to focus on test cracking versus improving language proficiency? Is validity affected by repeated test-taking behaviour? How does the test construct itself shift over time or over test conditions, or change due to dynamic interactions between learner, language and task? As we come to better understand these dynamics, it is to be hoped that the educational value of repeated testing becomes clearer.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Tony Green https://orcid.org/0000-0003-4893-1798

Alistair Van Moere https://orcid.org/0000-0002-6631-5318

References

Abbott, E. E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Monographs: General and Applied*, *11*(1), 159–177. https://doi.org/10.1037/h0093018

Boyd, E., Green, A. B., Hopfenbeck, T., & Stobart, G (2019). *Effective feedback: The key to successful assessment for learning*. Oxford University Press.

Chappell, P., Yates, L., & Benson, P. (2019). Investigating test preparation practices: Reducing risks. *IELTS Research Reports Online Series, No. 3*. British Council, Cambridge Assessment English and IDP: IELTS Australia. https://www.ielts.org/-/media/research-reports/2019-3-chappell_et_al_layout.ashx

Edgeworth, F. (1888). The statistics of examinations. *Journal of the Royal Statistical Society*, *51*(3), 599–635. https://www.jstor.org/stable/2339898

Elman, B. A. (2000). *A cultural history of civil examinations in late imperial China*. University of California Press.

Green, A. B. (2017). Linking tests of English for academic purposes to the CEFR: The score user's perspective. *Language Assessment Quarterly*, *15*(1), 59–74. https://doi.org/10.1080/15434303.2017.1350685

Green, A. B. (2019). Washback in language assessment. In C. A. Chapelle (Ed.), *The concise encyclopedia of applied linguistics*. Wiley-Blackwell. https://doi.org/10.1002/9781405198431

Gu, L., & Xi, X. (2015). Examining performance differences on tests of academic English proficiency used for high-stakes versus practice purposes. *Research Memorandum No. RM-15-09*. Educational Testing Service. https://www.ets.org/Media/Research/pdf/RM-15-09.pdf

Herbert, A. (1889). *The sacrifice of education to examination: Letters from "All sorts and conditions of men"*. Williams & Norgate.

Isbell, D. R., Winke, P., & Gass, S. M. (2019). Using the ACTFL OPIc to assess proficiency and monitor progress in a tertiary foreign languages program. *Language Testing*, *36*(3), 439–465. https://doi.org/10.1177/0265532218798139.

Ortega, L., & Byrnes, H. (2009). *The longitudinal study of advanced L2 capacities*. Routledge.

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Thorndike, E. L. (1912). *Education: A first book*. Macmillan.

Yu, G., He, L., Rea-Dickins, P., Kiely, R., Lu, Y., Zhang, J., Zhang, Y., Xu, S., & Fang, L. (2017). Preparing for the speaking tasks of the *TOEFL iBT®* test: An investigation of the journeys of Chinese test takers. *ETS Research Report Series*, *2017*, 1–59. https://doi.org/10.1002/ets2.12145