Listening. Learning. Leading.®

Research Memorandum
ETS RM-13-09

Supporting an Expiration Policy for English Language Proficiency Test Scores

Donald E. Powers

Venessa Lall

October 2013

ETS Research Memorandum Series

Supporting an Expiration Policy for English Language Proficiency Test Scores[1]

Donald E. Powers and Venessa Lall

Educational Testing Service, Princeton, New Jersey

October 2013

Action Editor: James Carlson

Abstract

An operational requirement for all testing programs, including English language testing programs such as those for the *TOEFL*® and *TOEIC*® tests, is the need to establish a period of time after which test scores will no longer be reported to test score users. This paper provides support, both empirical and logical, for the TOEFL and TOEIC programs' current policy of not reporting scores that are more than 2 years old.

Key words: TOEFL, TOEIC, score expiration, test score validity, score reporting

Because most tests are designed to provide only a snapshot of test takers' capabilities at a given point in time when a test is taken, test scores may become less trustworthy indicators of those capabilities as time passes. Therefore, an operational requirement for all testing programs, including English language testing programs such as the *TOEFL*® and *TOEIC*® test programs, is the need to establish a period of time after which a test score will no longer be reported—either to test takers or to test score users.

An Analogy From Food and Drugs

In the food and drug industries, products typically have a *shelf life*, defined as the length of time after which a product is no longer considered suitable for sale or consumption. *Unsuitable* may mean that a product's quality (that is, its freshness in the case of food, and its potency in the case of drugs) is no longer guaranteed. Alternatively, for some products, an expiration date may signify that the product can no longer be used *safely*, because of the risk of food poisoning, for instance.

In reality, however, there is usually no definite period of time after which a drug becomes completely ineffective or a food product totally unfit for consumption. Rather, both freshness and potency are matters of degree: Both tend to exhibit relatively gradual declines, rather than precipitous ones. The same is true regarding test score validity. Therefore, establishing an exact expiration date for test scores is a somewhat arbitrary exercise.

Of course, the analogy from food and drugs to test scores doesn't apply completely. For example, pharmaceutical companies may have reasonably definitive research and therefore relatively clear criteria with regard to when a drug stops working. Criteria for the meaningfulness of test scores, however, are much less prescriptive. The analogy seems close enough, however, to be at least somewhat instructive when thinking about expiration dates for test scores. For example, it is probably rare that either drugs or test scores suddenly lose their potency (validity), and it is probably equally rare that the use of an expired test score will result in actual harm to a test score user or to a test taker. However, the meaning of test scores may not be as "crisp" for older test scores as for more current ones. In any event, determining an appropriate expiration date is not an easy process.

For some tests—for example, those used to facilitate postsecondary admissions decisions (like the *Graduate Record Examinations*® General Test)—the abilities tested (e.g., general academic reasoning abilities) tend to develop relatively slowly over long periods of time. They also tend to decline rather slowly. Accordingly, it has been customary to establish a reasonably long expiration date for these kinds of tests—typically about 5 years.

Proficiency in a foreign language, on the other hand, is believed to be more amenable to improvement from specific training than are more general reasoning abilities. It is also thought to decline more rapidly as a result of disuse. Therefore, it seems prudent to establish somewhat shorter expiration periods (i.e., less than 5 years) for foreign language proficiency test scores. Just *how much* less is debatable, but the longer the interval between testing and test score use, the less reflective the test score will be of a test taker's proficiency. Specifically, if a test taker has continued to learn since he or she was tested, then an old test score may *under*estimate his/her proficiency. If, on the other hand, the time since previous testing has involved little if any use of previously acquired language skills, then language proficiency may decline, thus rendering older test scores as *over*estimates of proficiency.

The first case (i.e., an older test score *under*estimating proficiency) is arguably the less serious of the two problems—at least from the perspective of the *test score user*, if not from that of the *test taker*. This situation will probably be self-correcting in the following sense. Test takers whose language skills have sharpened significantly since they last tested will probably wish to retest in order to show themselves in the best light possible. On the other hand, test takers whose skills have declined because of nonuse (or any of several other possible reasons) will probably have little if any incentive to retest. It is this latter case, primarily, that the test score expiration policies for programs such as TOEFL and TOEIC are intended to address. That is, mainly, the concern arises not from further language *acquisition*, but rather from possible language *attrition*.

Research on Language Attrition

When we look to the research on language attrition for advice, little direct guidance is forthcoming. We base this conclusion primarily on a very recent and extremely comprehensive review (Bardovi-Harlig & Springer, 2010), which entailed a critical evaluation of 126 individual studies of language attrition.

A significant number of the studies reviewed by Bardovi-Harlig and Springer (2010) were concerned with attrition of subjects' *first* language and so do not apply directly to second language attrition. Generally, attrition in a second language is considered to be far more complex than is attrition in the first language, in large part because acquisition of a first language is, to some degree, invariably successful, whereas acquisition of a second language is not.

Bardovi-Harlig and Springer (2010) noted several hypotheses that have been advanced to explain the loss of *first* language proficiency. These hypotheses have been tested to varying degrees, but none has received overwhelming support. Furthermore, fewer of these hypotheses have been tested with respect to *second* language attrition. Bardovi-Harlig and Springer identified a wide variety of variables that require control in attrition studies. Apparently, however, few studies have been adequate in this respect. Moreover, many of the studies have been conducted over very short periods of time (e.g., summer vacations as the period of disuse). In addition, other studies have had serious design flaws. Chief among them has been a failure to measure peak attainment, either at an appropriate time or, in fact, at all. Thus, a substantial number of studies have failed to establish a baseline against which to measure attrition. This has been seen as "the chief problem" (Bardovi-Harlig & Springer, p. 24) in study design: Whereas in studies of second language *acquisition* the baseline can be assumed to be zero (i.e., no knowledge), there is no such common starting point for studies of language *attrition*. That is, study participants will almost always exhibit considerable variation at the outset with respect to their level of language acquisition, thus precluding a common baseline. Moreover, reduced use of a second language is not an all-or-none phenomenon. Rather, there is typically a wide range of continued input, exposure, and use. And even when the use of a second language is relatively continuous, not all aspects of language knowledge are exercised uniformly. So, there may be gains in some areas and simultaneous losses in others.

Besides identifiable flaws in their design, most studies have also been limited in other ways as well. For example, there have been relatively few longitudinal studies of attrition, and even these have tended to be only 1-2 years in duration. Moreover, the methods (and adequacy) of the assessments that have been used have varied widely, ranging from retrospective reports to standardized tests. All of this has lead Bardovi-Harlig and Springer to characterize research design in studies of second language attrition as being "weak" (p. 24).

Furthermore, although a variety of different populations have been subject to second language attrition research, none closely matches the populations served by either the TOEFL or the TOEIC programs. Studied populations have included mainly three: children returning from other countries, missionaries (after spending time abroad), and college and high school students returning to school after summer vacation. In addition, the predominant target language of second language attrition studies has been Japanese. All in all, therefore, it is very difficult to generalize from these populations to English language proficiency test takers in general, even when studies are well designed.

Thus, in light of the previously discussed limitations, there seems to be no consensus on the salient characteristics of second language attrition. However, we do know that second language attrition does not affect all components of language uniformly. For instance, production skills appear to be more vulnerable to attrition than do receptive skills, and oral skills seem to be highly susceptible.

Finally, a fairly wide variety of other variables have been implicated in language attrition. For example, retention may depend on the pedagogical methods used and on the degree of overlearning. In addition, the age at which a second language is acquired, and the duration/nature of immersion in the host country are important variables that have usually not been studied in attrition research.

Data on Test Repeaters

A second source of possible guidance is information on test takers who retake English language proficiency tests. For example, there have been at least two major studies to examine test score changes on the TOEFL. Wilson (1987) analyzed the patterns of repeat test taking and score change for some 200,000 test takers who had taken an older version of the TOEFL between 1977 and 1982. About 28% had repeated the test at least once, and a very small proportion (0.1%) had taken the test at least 11 times during the period. Wilson found a positive relationship between TOEFL score gains and the number of test repetitions: those who tested three times gained more than did those who tested twice, etc. (Of course, the number of retests was related to the interval between first and final testing.) The overall picture, Wilson noted, was one of “added gains with added testing” (p. 13), with additional gains appearing to diminish slightly with each subsequent retesting. Unfortunately, Wilson’s study was conducted on an earlier version of the TOEFL. Furthermore, he did not report information on the distribution of gains or on the extent to which some examinees’ scores may have *decreased* upon retesting.

Zhang (2008), however, focused on score changes on the *current* version of the *TOEFL iBT*® test, examining both score increases and score decreases. However, all of the examinees in Zhang’s sample—some 12,000 test repeaters—had retaken the test within a short time interval (30 days). Thus, her results pertain primarily to the stability (test-retest reliability) of test scores for the current version of the test, not to any longer-term changes resulting from real growth (or real decline) in the abilities being tested.

The most recent data on test repetition comes from some 87,000 TOEFL examinees who took (and retook) the TOEFL between August 2010 and February 2011 at intervals of (a) 30 days or less, (b) 31 to 90 days, (c) 91 to 120 days, or (d) more than 120 days between tests (Venessa Lall, personal communication, August 5, 2011). (The median number of days between tests for those retaking it after more than 120 days was 147.) These data reveal the following. For the reading and listening measures, about two thirds of examinees experienced score gains (or no change in their scores). The percentages are very slightly higher as the length of the interval between tests increases. The corresponding percentages (of gains or no change) for the speaking and writing measures are slightly higher—about 71% overall. Changes in speaking and listening scores are somewhat more strongly related to the interval between testing than are reading and writing scores. For instance, for the speaking test whereas 69% exhibited non-negative score changes after retaking the test up to 30 days later, 75% showed comparable changes after repeating the test after 120 days.

More critical to our concerns here, however, is the fact that significant percentages of test takers exhibited score *decreases* upon retesting. See Table 1.

Table 1

Score Decreases for TOEFL Test Takers Upon Retesting

	% of test takers with score decreases upon retesting			
Section	Across all time intervals	After 120 days		SEM
		Decrease of 1 or more SEMs	Decrease of 2 or more SEMs	
Reading	29–33%	13.4%	6.4%	2.36
Listening	30-36%	15.4%	6.9%	2.32
Speaking	24-31%	14.7%	7.3%	1.57
Writing	28-30%	13.8%	4.3%	2.41
Total score		10.5%	4.0%	4.36

A significant proportion of these decreases can, of course, be attributed simply to the fact that tests are not perfectly reliable and that therefore test scores will fluctuate somewhat from one occasion to the next. However, a significant percentage of test takers exhibited relatively large score decreases when taking the test again 120 days or more later, as shown in Table 1.

On the basis of these data, it is not unreasonable therefore to assume that some of these large decreases represent more than simple random fluctuations due to test unreliability. And, as intimated earlier, test takers whose English language proficiency has deteriorated are probably less likely to retest than are those who may have improved. Thus, being largely self-selected, TOEFL test repeaters may provide an inaccurately low estimate of the proportion of all TOEFL test takers who would exhibit score decreases if retested. Therefore, in light of these data, a 2-year TOEFL score expiration policy might seem relatively liberal.

An Ideal Study

Although the studies of test repetition data such as those discussed above are at least somewhat helpful, they are still not entirely useful for establishing an empirically based score expiration policy for English language proficiency tests. They do, however, suggest that test score users should not assume that all older scores (even those that are only 5 months old) can be trusted unconditionally. Ideally, the kind of study that would be required to formulate a score expiration policy would entail an extensive longitudinal approach in which various groups of test takers would take the test with varying intervals of time between testing. It would also require careful documentation of the language experiences of test takers between the times of their testing. And finally, it would require that test scores, such as TOEFL and TOEIC, obtained at different points in time be related to some meaningful validity criterion (e.g., grades, self-assessments, or faculty ratings) in order to establish any difference in the validity of scores of varying ages (i.e., scores obtained at different intervals of time from the criterion performance). Needless to say, such a study would be difficult to implement. Moreover, because attrition may not be uniform across all four language domains, separate studies would be needed for each section of the test.

Conclusion

Thus, in summary, there is little research on either (a) foreign language attrition or (b) changes in test score validity over time that would fully justify any score expiration policy for English language proficiency tests. The policy, therefore, generally has been to set a period of 2 years. This policy is conservative in that if not perfect, it errs on the side of protecting test score users from *over*estimating test takers' language skills, on the assumption that the proficiency of some test takers will have declined since they last tested. At the same time, however, this policy is also likely to protect at least some test takers by ensuring that their English language proficiency is not *under*estimated substantially when they have made significant improvements in English language skills since they last tested. In any event, decisions made on the basis of scores from any English language proficiency test, such as TOEFL and TOEIC, should be fairer and more valid if test scores are obtained reasonably close to the point at which decisions are made.

References

Bardovi-Harlig, K., & Springer, D. (2010). Variables in second language attrition: Advancing the state of the art. *Studies in Second Language Acquisition, 32,* 1-45.

Wilson, K. M. (1987). *Patterns of test taking and score change for examinees who repeat the Test of English as a Foreign Language*™ (TOEFL Research Report No. 22). Princeton, NJ: ETS.

Zhang, Y. (2008). *Repeater analysis for TOEFL iBT*® (Research Memorandum No. RM-08-05). Princeton, NJ: Educational Testing Service.

Notes

[1] This paper is adapted from Powers, D. E., & Lall, V. (2012). *Supporting an expiration policy for TOEFL scores* (Research Memorandum No. RM 12-03). Princeton, NJ: Educational Testing Service.