Some Issues In Retesting Policy

Anne Thissen-Roe, Ph.D.
September 15th, 2016

One policy issue germane to certification and licensure programs is whether, and under what circumstances, to allow failing candidates to reapply and/or retest. Credentialing organizations are generally sympathetic to their candidates, but also often have real and serious concerns about the quality of candidates who pass their test only after two or more tries.

This white paper aims to shed some light on questions of retest policy. It considers applicable standards, characterizes retesting candidates, explores measurement issues in retesting, and, finally, discusses some available policy options.

Standards
Various applicable standards differ on the issue of retesting policy, but in general allow testing organizations the latitude to set reasonable limits on retesting.

The right to retake a failed examination is not enshrined in the 2014 Joint Standards for Educational and Psychological Testing (AERA/APA/NCME, 2014), which state that "Test users should explain to test takers their opportunities, if any, to retake an examination" (Standard 9.18, emphasis added).

The National Commission for Certifying Agencies (NCCA) 2016 Standards for the Accreditation of Certification Programs (Institute for Credentialing Excellence, 2014) do not explicitly require that retesting be permitted. Like the Joint Standards, the NCCA Standards call for the policy and its rationale to be made public (Standard 6F and 7D) and require that an appeals process be standardized and documented (Standard 6G and 7F); an appeals process does not necessarily involve retesting. However, the NCCA Standards state that "The certification program must not unreasonably limit access to certification" (Standard 7C), which constrains the severity of an appeals process or retesting policy to that which can be justified as "reasonable."

The Principles for the Validation and Use of Personnel Selection Procedures (SIOP, 2003) favor the ability to retest: "Generally, employers should provide opportunities for reassessment and reconsidering candidates whenever technically and administratively feasible. In some situations, as in one-time examinations, reassessment may not be a viable option." In the SIOP Principles, the employer or potential employer of a test taker is the test user; certifying and licensing authorities, and other third-party testing organizations, are delegates whose compliance with the Principles must be ascertained by the employer prior to use of the test or resulting credential for selection purposes. If licensed or certified professionals are generally
self-employed, or the credential is not generally used for employment purposes, the SIOP Principles do not entirely apply.

**A Spotter's Guide to the Retesting Candidate**

Not all retesting candidates are alike. We observe different behaviors, reflecting different motivations and perspectives, among different candidates. Some of these behaviors are positive and appropriate. Others are dangerous and dishonest.

This white paper considers the candidate who retests after failing. In some cases, passing candidates do retake an exam in order to improve their scores, but they are relatively rare. Incidental retesting, in which score improvement may not be a goal at all, occurs more often in educational testing and selection testing than in certification and licensure: a student may take a test such as the SAT repeatedly at different times in order to comply with the timing requirements of different colleges and/or scholarship programs; an "off the shelf" employment test may be administered more than once to the same individual in the service of multiple job applications. These cases do not have ready equivalents in certification and licensure.

**Healthy retesting.** The prototypical, or ideal, retesting candidate fails the test once, studies or practices in order to improve her qualifications, retakes the test and passes. A process of self-evaluation and calibration occurs after the failure, through which the candidate realizes that she was unprepared, rectifies her self-understood deficiencies, and seeks the opportunity to repeat the test. It is this candidate that best justifies retesting.

As a credentialing authority making policy decisions, it is worth noting that even if a candidate studies or practices, it is sometimes difficult to evaluate how much real change has occurred (at least, prior to taking the test), or to judge whether "sufficient" study or practice has taken place. Some candidates are more effective and efficient at study and practice than others, so the same number of hours might translate to different amounts of change in qualifications. Qualitatively, study might be general and applicable to professional practice, or specific to passing the test. These complications are compounded by the presence of instructors with incentives to secure a retesting opportunity (and ultimately a credential) for the candidate; reports of study may be exaggerated or outright false, or an instructor might (knowingly or unknowingly) encourage and enable dishonest behavior. I'll return to these issues in the next section. For now, suffice to say that we don't always know whether a candidate has honestly improved.

It is also worth a brief note that some tests, according to the theory behind them, measure qualifications which the candidate cannot change. In a working adult population, improvement is not usually expected on tests of general mental ability, for example. However, knowledge, practical skill, speed or fluency, and physical abilities such as muscle strength are all attributes that candidates can work to improve.
There are other situations in which retesting is honest and healthy. A "near miss" candidate, whose score on the first try was marginally below the standard, may have been a false negative: a candidate whose qualifications actually meet the standard, but who failed "by chance" due to particulars of the test and their interactions with the candidate. All test scores have a margin of error; good test development practices reduce, rather than eliminate, chance false positives and false negatives. Decision consistency statistics are a good way to evaluate how well a test avoids false positives and false negatives.

A form of false negative also occurs when testing conditions are sufficiently non-standard to affect a candidate's score -- for example, a power failure or weather event ends the administration early, or a candidate becomes acutely ill. Often, such nonstandard conditions are obvious at the time of testing, and the score can be canceled, but that is not always the case and not always desirable; many incidents occur during testing that appear likely to have a severe effect on candidate scores, but upon later analysis, do not.

Test administrators and score users tend to acquire a healthy skepticism of after-the-fact claims of testing irregularities, as many such claims are nothing more than excuses. Some evidence of an actual false negative might include: 1) a report of the irregularity prior to score reporting; 2) independent third-party reports or hard evidence confirming the irregularity; and/or 3) ancillary evidence about the candidate that suggests that the failing score is atypical, such as several high passing scores on other tests. In these and even in weaker cases, it is often in everyone's best interest to repeat a measurement whose validity is doubted.

**Pathological retesting.** On the other hand, there are pathological forms of retesting, in which a liberal retest policy is abused in order to secure a credential for an unqualified candidate.

The usual indicators of pathological retesting are many retests, or rapid retests, or both. We see candidates who test more than once in the same day, if they can, or on consecutive days if that's the minimum delay. With a next-day minimum, even where there is an instructional requirement, we see candidates testing eight times in two weeks. There's a locally legendary candidate who spent a day driving from place to place around Los Angeles, testing as a walk-in wherever a testing center had a seat, until closing time; that candidate understandably inspired a policy change.

These candidates may be capitalizing on chance. Marginally underqualified candidates can expect that, if they retest enough times, the test's margin of error will favor them and they will become false positives (and, more importantly for them, credential holders). These candidates are characterized by a steady pattern of just-below-standard scores, eventually followed by a low passing score. Their actual qualifications are unchanged.

Other candidates may realize that they will eventually see items again, and recognize them. This can happen when anchor items are present on multiple test forms for equating purposes, or when a retesting policy allows more retakes than there are test forms. Such a candidate can
study specific items that she remembers having difficulty with. Her scores will increase gradually as she memorizes the item bank, until (again) she barely passes. However, her knowledge of the profession has not meaningfully increased.

Item harvesters are pathological from the first take, but can appear as retesters. These are candidates who are offered compensation to memorize and share items seen during testing, in order to enable later candidates to cheat. (Incidentally, it’s a good idea to have a signed, contractual candidate agreement that states that the candidate will do no such thing. Otherwise, some candidates won’t realize what they’re doing is wrong.)

Because item harvesters have an incentive to take the test that typically covers testing costs and then some, and that incentive is not limited to the first take, item harvesters may intentionally fail in order to repeat the test. Such candidates are characterized by atypical right-wrong patterns in their responses, such as clusters of wrong answers to sequential questions, or wrong answers to easy questions combined with right answers to hard questions. (It is worth noting that such patterns can also result from perceived time pressure or other contextual factors.)

Other item harvesters take advantage of liberal pre-qualification policies to fake their way in the door, knowing that they are not qualified professionals and will never want or need the credential; these harvesters may willingly take the test and fail even when a retest is not available, and the marks on their records are permanent. These professional harvesters may have a hidden recording device (or conspire with a proctor to ignore such a device), and work quickly through the test, giving random answers, solely in order to witness and record every item. Harvesters working by memorization may work slowly, particularly on new sections of the test; their item response times can be expected to be more correlated than the average candidate with readability statistics (e.g., word count) for those sections being memorized.

A professional harvester might also abuse the retesting policy by falsifying a testing irregularity, with or without a proctor’s assistance, and immediately request score cancellation. Such a candidate might fake becoming ill, for example, or "accidentally" unplug her computer. This is another reason for score users and test administrators to be skeptical of irregularity reports.

Item harvesters work in groups, and their traces are easier to detect in the aggregate than individually. A pattern of "fail-fail-pass" across several candidates may be a hint, especially where the retesting policy limits attempts; a consistent pattern of faster responses to items when they have been seen by another candidate is a giveaway. Unfortunately for those of us doing the data forensics, the groups may not be geographically related or otherwise obvious. Test security teams use multiple convergent streams of information to identify groups of candidates who may be working together in one way or another.
The Retest Effect

It is well established in the field of testing that candidates who test more than once are generally able to improve their scores on the second attempt. Test users often have questions like: Which score is more valid? Is the second score deceptive?

The existence of a "retest effect," or average score gain across retesting candidates, is reported in the psychometric literature across testing contexts (e.g., education and employment as well as certification) and content types (ability, knowledge, skill, and personality). The scale of the retest advantage has generally been reported in the vicinity of a quarter of a standard deviation, but may be higher or lower depending on the content type and specifics of the test (Hausknecht, Halpert, Di Paolo & Moriarty Gerrard, 2007; Schleicher, Van Iddekinge, Morgeson and Campion, 2010; Thissen-Roe, Baysinger & Morrison, 2012).

The data are consistent with the retest effect being a combination of two sources of improvement: 1) real improvement in the construct the test is intended to measure, and 2) reduction or elimination of errors or omissions made due to the candidate's unfamiliarity with the test format; that is, improvement in "testwiseness."

The familiarity effect occurs immediately, appearing essentially identically for a second administration on the same day, and a second administration after months. It is greatest between the first and second administration of a test, with diminishing returns thereafter (Mangos, Thissen-Roe & Robinson, 2012).

Real improvement, however, takes place over time, and typically increases with greater time elapsed between tests, although there are diminishing returns for each elapsed day as weeks and months pass. (A linear increase in knowledge cannot indefinitely translate to a linear increase in a test score; the candidate approaches a maximum possible score.)

If we plot the score increase seen by a retesting candidate against days elapsed (on a logarithmic scale, for a general sense of "hours to days to weeks to months"), we see something like Figure 1. The average retest effect has a positive intercept, which is the familiarity or practice effect, and a positive slope, which reflects real improvement on underlying knowledge. Individual candidates vary widely, because the "chance" (or person by test interaction) error component in a test score is independent for the two administrations, and thus the error of the difference is larger than the error of a single score. Some candidates improve by twice or more the average retest effect; some candidates actually score worse on the second try, even though real improvement has taken place.

Neither real improvement over time, nor practice effects, render the second score on a knowledge test invalid. Practice and familiarity, or being "testwise," don't give a candidate more knowledge, or, on a well-developed test, even the illusion of more knowledge. They do remove a source of error contributing to the test score, which is generally negative but not uniform across candidates. To the extent that being "testwise" is needed on a particular test,
the second score is actually a better representation of a candidate's true knowledge. This, too, is borne out by the data, at least where relevant data are available. In education and employment testing, criterion-related validity for retest scores exceeds that of initial scores (Van Iddekinge, Morgeson, Schleicher, & Campion, 2011). That is, second scores predict actual job performance better than first scores.

Figure 1. Score increases between first and second attempts by simulated candidates on a fictional 100-item knowledge test. The average score rise, in pink, shows both an immediate practice/familiarity effect (intercept) and an effect of learning and study that increases with elapsed time (slope). The thickness of the line indicates the number of candidates contributing to the average; the blue point cloud shows the considerable individual variation around the averages. Not every candidate benefits from retesting, even if she studies.
Test scores are not unique in this behavior; they resemble other repeated measures of performance in which a stable individual difference combines with a task-specific practice effect. An obvious example is skilled job performance, where there is a job-specific learning curve, followed by a period of stable performance for the remainder of an employment spell. Initial performance is generally lower, but also more widely variable; trainees are expected to make some mistakes, but not a particular number of mistakes. Analogous to the finding of higher validity for later test scores, an employee is usually considered to be better judged according to her long-term performance than according to what she did on her first day (or week) on the job.

However, to the rule that later scores are as good as or better than first scores, certain exceptions exist outside the domain of knowledge tests. Tests of fluid intelligence measure an individual’s ability to reason about completely novel situations, to which experience, knowledge, and acquired skills do not apply; some of these tests rely on the unfamiliar format of the test to provide novelty. The substitution of practiced skill (crystallized intelligence) is undesirable, because it introduces variation due to a second construct. In this case, the first administration of a test produces the most valid score. Tests of fluid intelligence, however, are generally inappropriate for certification and licensure, which are predicated on the notion that, in a particular profession, experience, knowledge, and acquired skills do apply. We are not born lawyers, doctors, chefs, pilots or public accountants. We learn skills, acquire expertise, and then, finally, are certified or licensed.

Caps, Delays, and Issues in Score Reporting
With these issues in mind, credentialing organizations use a variety of rules and policies to manage retesting candidates and their scores.

Delays. Some programs allow failed candidates to retest after a certain amount of time has passed. For example, individuals who intend to fly small unmanned aircraft systems ("drones") for commercial purposes in the United States must first pass an Airman Knowledge Test administered by the FAA; if they fail, they may retake the test after 14 days.

The imposition of a delay may be pragmatic: some testing programs require advance notice to set up a test for a candidate, or administer tests only on certain dates throughout the year. However, generally, as in the case of the Unmanned Aircraft Systems test, the idea is to encourage candidates to study in between testing opportunities. Allowing retests after a delay supports healthy retesting patterns without undue inconvenience, except among certain special populations (e.g., residential college students testing at the end of a semester; candidates who must travel to distant locations in order to test). The delay is a small imposition if a candidate would take some time to study anyway, but a significant one for a pathological retester who has no use for the intervening time.
**Caps.** Other programs limit the number of opportunities to retake an examination. A candidate might, for example, have only three chances to pass an examination. A variation on the theme is a fixed number of opportunities to retest before some larger effort is required to requalify, such as completing an educational program or submitting evidence of supervised work in the field. Some programs use both a delay and an absolute cap; others limit the number of retake opportunities within a period of time (e.g., twice in a year).

Like delays, caps are intended to be permissive of healthy retesting patterns, but discourage pathological retesting. Whereas delays are a "soft" deterrent, caps actually cut off candidates who appear to be pathological retesters, or hopelessly unqualified. In selecting a cap, a credentialing organization should consider the frequency and impact of false positives (in this case, marginal candidates whose healthy retesting patterns ran afoul of the cap), as well as the ways in which pathological retesters might game the system. Caps do not eliminate the practice of item harvesting.

**Progressive requirements.** Programs may require candidates to submit some form of justification for a request to retake a test, such as the aforementioned completion of a program of study, or an excuse for the original poor showing. The required evidence may vary depending on the number of times a candidate has already taken the test.

Like delays, study requirements act as a general deterrent on retesting, but have less practical effect on candidates who would be likely to study anyway.

In workplace testing, different cut scores are sometimes used for the first, second, third and nth take (Clauser & Wainer, 2016); however, I know of no certification or licensure program that applies this strategy.

**Differential reporting.** Some programs choose to pass the question of what to do about retesters on to the users of the test scores, for example by printing a candidate's take count on her certificate. If a user, such as an employer, wants to favor certified professionals who passed on the first take, and who are more likely to have passed by a large margin, that user is welcome to do so. If another user values persistence, self-improvement and second chances, then a take count of two or more may not be a disadvantage.

**Policies involving multiple tests or sections.** Some credentials require more than one test, or a test with multiple sections that produce multiple scores. These scores can be combined as a weighted sum or average which is compared to a single cut score (compensatory scoring), or using a rule that requires a passing score on each individual test (noncompensatory scoring or "multiple hurdles").

In both cases, a candidate can fail overall while passing, or doing adequately on, some sections of the test. In noncompensatory scoring, a single failed section is always sufficient to deny a candidate the credential. In compensatory scoring, a candidate can do quite well on one or
more sections, but badly enough on another section or sections to bring the combined score down below the cut score. Either way, the candidate may really only need to improve on one section's score to pass.

Clauser and Wainer (2016) classify retake policies for credentials requiring multiple tests into two general categories: 1) those that require a failing candidate to retake all of the tests, and 2) those that permit a failing candidate to retake some of the tests, such as the failed sections or the sections with the lowest original scores. The first category behaves like a single test, albeit one with complex scoring, for the purpose of retest effects and retest policies. The second category, however, increases the vulnerability of a testing program to accidental false positive results, as well as pathological retesting.

How does this work? Well, classical test theory (and modern test theory, less directly) holds that no matter how related or unrelated two or more sections are in material covered, the error introduced by person-form interactions, or "chance," is independent between sections. That means that under compensatory scoring, systematic retest effects aside, a positive change in one score will on average be balanced out by a negative change in another score. If a marginal candidate can hold onto her highest score on each section -- the one most likely to include beneficial error from a person-form interaction -- she can inflate her overall score by cherry-picking her attempts. This is another form of pathological retesting, gaming the system, and the candidate may not see anything wrong with it.

Under non-compensatory scoring, the situation is slightly more complex, but the outcome is essentially the same. For a candidate who is marginally underqualified in only one area, and competent in all others, there's no real difference between a policy of retaking the whole test and a policy of retaking only failed sections. However, a candidate who is underqualified in more than one area can capitalize on chance separately for each of her weak areas. Her probability of passing two or more sections includes, rather than excluding, her probability of passing both (or all) weak area sections on two or more different attempts. Again, this group of underqualified candidates is more likely to receive credentials under a policy that allows individual sections to be retaken.

It is not always practical to readminister all tests that contribute to a credentialing decision. For example, a credential may require that the candidate hold a prerequisite credential in good standing. That prerequisite credential might certify a lower level of professional skill, in which case it is improbable that a candidate would pass the current credential's test while truly not holding the prerequisite level of skill. The prerequisite might not be under the control of the credentialing organization, e.g., a degree requirement. Or it might simply be embarrassing to the organization to jeopardize or walk back earlier credentials in the face of later optionally provided information. (On the other hand, prerequisite credentials and authorizations, other than degrees and diplomas, can and regularly do expire after a specified period of time.)
My final word on this subject is milder than Clauser and Wainer's. Those authors recommend always readministering all sections. My recommendation is that simultaneous and/or compensatory sections should be re-administered together. In the case of sequential multiple hurdles, the probability of multiple false positives should be evaluated alongside practical considerations when setting a policy.

**Scoring rules that incorporate or adjust for previous scores.** Clauser and Wainer (2016) also recommend Millman's (1989) "Asian Option" approach to stabilizing test scores for retesters. This approach involves adjusting the latest score for the scores achieved on earlier attempts; in essence, keeping a running average rather than reporting the latest score. Such an average has reduced error, equivalent to a longer test: twice as long for the second test, if there are no repeated items, three times as long for the third, and so on. Over several attempts, the average converges toward a good representation of the candidate's true qualification level. It makes it difficult, if not impossible, for an underqualified candidate to "take it 'til you pass" based on error alone.

The Millman approach has limitations. It does not account for practice effects, real learning or recognition of individually repeated items. It is based on a model "...without the confounding effects of practice, learning, forgetting, fatigue, and the like" (Millman, 1989, p. 5). If you have data that tell you there is systematic change in your candidates, don't assume there isn't any. If you have enough data, you can fully model retest trajectories over take count and time, and account for lower error variance in later scores. If (more likely) you have some data, but not that much, you might simply choose to have scores become "stale" after some time passes, such that a previous attempt from a month ago has a reduced weight in the running average, and one from two years ago has no weight at all. (A time-weighted averaging scheme like this should, if possible, be informed by data about the correlation between consecutive scores at different take counts and intervals, as compared to Cronbach's alpha or an equivalent measure of internal consistency reliability.)

Another limitation of the Millman approach is that, not being familiar to candidates, it requires a large amount of candidate education and publicity in order to actually suppress pathological retesting behavior, and in order to avoid generating angry candidates and bad publicity.

**Conclusion**
All in all, it is generally wise to allow failing candidates to retake a credentialing examination under reasonable circumstances. A candidate is likely to improve her score on a second attempt, and this improved score is likely to be a better reflection of her true qualifications. However, credentialing organizations have tools at their disposal, and should use them, to curb or deter pathological retesting behaviors.
References


