Test-specificity of the advantage of retaking cognitive ability tests

Nadine Matton, Stéphane Vautier, Éric Raufaste

▶ **To cite this version:**

Nadine Matton, Stéphane Vautier, Éric Raufaste. Test-specificity of the advantage of retaking cognitive ability tests. International Journal of Selection and Assessment, 2011, 19 (1), pp 11-17. 10.1111/j.1468-2389.2011.00530.x . hal-01021627

HAL Id: hal-01021627

https://enac.hal.science/hal-01021627

Submitted on 18 Jul 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Running head: TEST-SPECIFICITY OF RETAKERS' ADVANTAGE

Test-Specificity of the Advantage of Retaking Cognitive Ability Tests

Nadine Matton, Stéphane Vautier and Éric Raufaste

Université de Toulouse

Abstract

In selection settings, when people retake the same cognitive ability tests, scores are generally positively biased. Our approach aimed to test whether these previous exposure effects are test-specific or transferable to other tests. We compared the differences between scores for first-time test takers and retakers for two kinds of material: "old" tests, known only to the retakers, and "new" tests, unknown to both groups. The current study used data collected during two sessions–S and $S + 1$–of a selection process for entry into the French national air transport pilot training system, with at least 500 first-time test takers and 130 retakers in each session. For Session S, on average, retakers scored higher on the "old" tests, but not on the "new" tests. Moreover, the material that was new to retakers at Session S was old at Session $S + 1$, and the finding for old tests could be replicated at Session $S + 1$. The finding that the acquired skills that led to higher scores on old tests were only test-specific is discussed.

Test-Specificity of the Advantage of Retaking Cognitive Ability Tests

Retaking tests

Many organizations allow applicants to retake cognitive ability tests used in a selection process if applicants fail the first time (Lievens, Buyse, & Sackett, 2005; Lievens, Reeve, & Heggestad, 2007). Thus, people who retake the tests (the "retakers") have already been in contact with the test materials. The question is whether having taken the tests a first time has an impact on how they perform on the tests the second time. Should practitioners interpret cognitive ability test scores in the same way for first-time test takers and retakers? Besides, even if organizations do not allow retesting, they can be confronted with the problem of test-taking preparation or experience with similar cognitive ability tests taken within another selection process. Therefore, better knowledge about the effects of retaking tests is of primary concern for selection practitioners.

As shown by numerous studies on test-retest effects, substantive mean gain scores can be expected between the first and second sessions for a given group of applicants (e.g., Cliffordson, 2004; Coyle, 2006; Hausknecht, Trevor, & Farr, 2002; te Nijenhuis, Voskuijl, & Schijve, 2001; te Nijenhuis, van Vianen, & van der Flier, 2007; Reeve & Lam, 2005, 2007; Kulik, Bangert-Drowns, & Kulik, 1984; Kulik, Kulik, & Bangert-Drowns, 1984). These mean gain scores are labeled *practice effects*, which are defined as "changes in a person's test score from one administration to the next" (Hausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007, p. 374). In their recent meta-analysis, Hausknecht et al. (2007) revealed a mean effect size of +0.26 standard deviations from one test administration to the next. Nevertheless, this mean effect can be interpreted in different ways.

Interpreting Gain Scores between Two Sessions

Gain scores in cognitive ability testing from one session to another can be interpreted in different ways (Chamorro-Premuzic & Furnham, 2010). Anastasi (1981) attributes the improvement to the development of actual abilities, as long as the individual invested substantial effort. Many other authors believe that scores' improvements are not necessarily related to increases in the target abilities (Cole, 1982; Jensen, 1998; Lubinski, 2000; Messick & Jungeblut, 1981; Sackett, Burris, & Ryan, 1989; Snow, 1982; Sternberg, Ketron, & Powell, 1982). The observed gains in scores could be due to an increase in "test-wiseness," test-familiarity, motivation, self-confidence[1], sheer test-taking practice, and/or a reduction in stress or anxiety, and/or the memory of questions and answers.

Research on the question of the interpretation of the gain scores has focused on different aspects. First, some studies have modeled the psychometric structure of the scores at sessions 1 and 2 to test measurement invariance hypotheses, with mixed conclusions supporting both structural invariance (Reeve & Lam, 2005) and non-invariance (Lievens et al., 2007). Second, Jensen (1998) and te Nijenhuis et al. (2007) linked the gain scores to the hypothesized g-loadings of the tests, finding a negative correlation. Thus, the more a test is g-loaded, the less likely it is to allow for increases in scores. Third, a few studies have compared predictive validities based on either first-session scores or on second-session scores. Hausknecht et al. (2002) found that the predictive validity coefficient was significantly larger for those hired after a single session compared to those hired after multiple administrations ($r = .36$ vs. $r = .24$). Lievens et al. (2007) also found higher validity coefficients for initial test scores compared to retest scores ($r = .19$ vs. $r = .00$). Inversely, Embretson (1987) showed that the predictive validity for training in text editing was increased with the use of post-test scores obtained after an intervention. Fourth, Matton, Vautier, and Raufaste (2009) modeled the psychometric structure of the change in observed scores of a battery of cognitive ability tests through a longitudinal

SEM approach and reported a one-factorial structure of change. They concluded that there was no need to suppose change on the g-factor or in the target abilities to account for their observed gain scores. Thus they did not exclude construct-irrelevant explanations encompassing level of anxiety, general test-taking improvement, etc.

As mentioned by Jensen (1998), the best thing to do for investigating whether the observed gain can be attributed to a real increase in ability is to assess the *far transfer* or *broad generalizability* (p. 333) of test-taking skills. Transfer or generalizability of test-taking skills refers to skills acquired during the training for a given test, which can be applied to better performance on similar material. Gain scores observed for the same test administered twice do not allow for disentangling the gain due solely to test-specific familiarity from the gain due to the acquisition of skills that are sufficiently general to be transferred to similar test material. The purpose of the present study was to test whether mean score gains on a given test should be interpreted as an increase in the targeted ability or only as test-specific familiarity.

Hypotheses

As suggested by Lubinski (2000), the gain scores observed after retesting might be due to the development of test-specific strategies, or to increased test-specific familiarity. Under such a "specificity hypothesis," retakers would tap specific skills acquired during previous encounters with a test, with little or no transfer to tests that were not previously encountered. If the specificity hypothesis holds, then, compared to a group of first-time test takers, the retakers would exhibit an advantage on "old" testing material, that is on material that they already encountered on previous sessions, but no advantage on "new" testing material, i.e., on material that they had not previously encountered. Now, let us consider (i) a series of testing sessions of which two are of interest, say Session S and Session $S + 1$, and (ii) a population of applicants including some retakers. Some tests in

Session S are entirely new to all participants, whereas others were previously encountered by retakers only. The specificity hypothesis entails two predictions:

> Hypothesis 1: On average, retakers score higher than first-time test takers on "old" tests at Session S.

> Hypothesis 2: On average, retakers do not score differently than first-time test takers on "new" tests at Session S.

Now, some tests are used in both sessions, whereas other tests are presented only in Session S. Thus, it is possible to evaluate the replication of Hypothesis 1. Indeed, for the participants who were present at Session S the "new" tests from Session S become "old" at Session $S + 1$. If a single exposure to the test is sufficient to induce the test-specific familiarity effect, then we should observe at Session $S + 1$ an advantage for the retakers on the tests that were new at Session S.

> Hypothesis 3: Hypothesis 1 is replicated at session $S + 1$ on the tests that were new at Session S.

To synthesize, this approach is a within-occasion and between-persons study of the effect of the historical status of the testing material.

Method

The present study used data collected during an actual selection process that presented the required characteristics, that is, (i) a sample composed of both first-time test takers and retakers, and (ii) testing material composed of both old and new tests. The selection process was the yearly selection for entry into the École Nationale de l'Aviation Civile, the French national air transport pilot training system. In the case of failure at one session, applicants are allowed to retake the battery of tests a maximum of three times.

Retakers represented approximatively 20-30% of the total number of applicants. The second step of this multistage selection process was comprised of cognitive ability tests, including (i) "old" tests, that remained unchanged from 2001 to 2007 and (ii) "new" tests, introduced in 2007 and reused in 2008.

Participants

Two samples of applicants were of interest:

- Sample 1 comprised 656 applicants who took the battery of ability tests in 2007. They were divided into a control group ($N = 517$, 86.1% male, 5.0% experienced pilots), who took the battery of tests for the first time, and a group of retakers ($N = 139$, 94.2% male, 13.7% experienced pilots), who had already taken the old tests at one previous session.

- Sample 2 comprised 835 applicants who took the battery of tests in 2008, with 674 first-time test takers (84.4% male, 2.8% experienced pilots) and 161 retakers (91.0% male, 1.2% experienced pilots) who were all present at the 2007 session. Among the 139 retakers in Sample 1, 31 took the selection tests again in 2008 but were not included in the sample of 835.

Materials

This study focused on three "old" and three "new" cognitive ability tests, described below using Carroll's terminology (1993). As mentioned above, old tests were determined based on their previous use in the selection process. Therefore they were known to the retakers. The three old tests consisted of:

1. *Visual perception test (S0).* This test is composed of three time-limited subtests (180 s, 300 s, 210 s) measuring the following abilities: (a) perceptual speed (an identical pictures test of 25 items), (b) spatial relations (a picture rotation test of 20 items), and (c) visualization (a block counting test of 15 items). Total scores vary from 0 to 85 (60 items

scored 0, 1 or 2 depending on the number of correct responses).

2. *Mechanical movement test (M0).* This test presents 36 situations to evaluate, from a mechanical point of view, with a choice of 4 possible answers for each situation. The test is time limited (25 min) and scores range from 0 to 42 (number of correct answers).

3. *Numerical ability test (N0).* This test comprises 30 short problem-solving items (involving secondary school level principles as proportionality, conversions, area computations...). The test is time limited (30 min) and scores range from 0 to 30 (number of correct answers).

The three new tests were introduced into the battery in 2007:

1. *Visual perception test (S1).* This test is the space relations test from the Differential Aptitude Tests, or DAT, battery (Bennett, Seashore, & Wesman, 1974), typically a test of visualization in three dimensions. This test is time limited (15 min) and scores vary from 0 to 30 (number of correct answers).

2. *Mechanical movement test (M1).* This test is the mechanical reasoning test from the DAT battery. This test presents 30 situations to evaluate from a mechanical point of view, with a choice of 3 possible answers for each situation. This test is time limited (15 min) and scores range from 0 to 30 (number of correct answers).

3. *Abstract reasoning test (R1).* This test is the French adaptation of the abstract reasoning subtest of the Graduate Managerial Assessment battery, or GMA (Blinkhorn, 1985). This test comprises 105 items and is time limited (30 min). Scores range from 0 to 105 (number of correct answers).

The tests S1, M1, and R1 were also used in the 2008 session; therefore, they became "old" tests for the 2008 retakers who were present in 2007.

Analyses

To test Hypotheses 1 and 2, the mean scores of the two groups of applicants were compared using t tests for independent samples, and effect sizes were estimated using Cohen's d and associated conventions (Cohen, 1988). To test Hypothesis 3 we also analyzed the 2008 session results by comparing the 2008 first-time test takers to the 2008 retakers on tests S1, M1, and R1.

Results

Session 1 Intergroup Mean Effects

Comparison of mean scores is displayed in Table 1. First, the mean scores of the retakers were significantly higher than those of the first-time test takers for the three old tests. Moreover, the standardized effect sizes were medium for tests S0 and M0, and large for test N0. Therefore Hypothesis 1 was supported by the data. Second, there was no significant difference between retakers and first-time test takers for the three new tests. Moreover, all effect sizes were below the 0.20 conventional threshold for small effects. Thus, Hypothesis 2 was also supported. Consequently, the retakers' advantage was specifically limited to testing material that they previously encountered.

Intercorrelations within Session 1: Convergent validity

The reader could interpret negligible or small effects by casting doubts on the convergent validity of the paired old and new tests. However, the paired old and new tests implement similar although distinct tasks. Table 2 allows to check that the pattern of intercorrelations conforms to the following expectation: as the tests S0 and S1 on the one hand, and M0 and M1 on the other hand, rest on the same kind of tasks, the respective correlations should reflect convergent validity. So a decreasing ranking of the observed correlations corroborates this expectation, for both the subsamples of first-time test takers

and retakers. Indeed, the strongest intercorrelations were observed for the two visual perception tests, S0 and S1 ($r(516) = .49$ for first-time test takers and $r(138) = .54$ for retakers), and between the two mechanical comprehension tests, M0 and M1 ($r(516) = .61$ for first-time test takers and $r(138) = .45$ for retakers).

Session 2 Intergroup Mean Effects

Analyses for Hypothesis 3 consisted of replicating the analyses from Hypothesis 1 and assessing the stability or change of the previously observed mean effects. Table 3 summarizes the comparison between mean scores observed for first-time test takers and retakers at the 2008 selection session, where the tests S1, M1, and R1, which were new for retakers at the 2007 session, became old for retakers at the 2008 session. The mean differences between retakers and first-time test takers were all significant ($p < .001$) and corresponded to medium size effects. Thus, for the same tests, (i) there was no significant advantage for the retakers when the tests were unknown, and (ii) after only one exposure to the testing material, retakers scored significantly higher compared to first-time test takers. Thus, Hypothesis 3 was supported by our data. Moreover, this result is evidence that the similarity of mean scores between first-time test takers and retakers at the 2007 session for the new tests was not due to a particular characteristic of these tests (e.g., tests that were globally easier or more difficult for the population).

Controlling for sample effects

As all of the 2008 retakers were 2007 first-time test takers, we had the opportunity to check whether their advantage in 2008 was not due to an actual superiority in the targeted abilities that would have already been present in 2007–although failing at this step of the selection process does not imply bad performances on the tests, as only 10% of applicants are eventually selected at each session. Our data showed that, in 2007, the subgroup of first-time test takers who became 2008-retakers did not perform better than

the rest of the 2007 first-time test takers. Their mean scores were not significantly different from those of the other first-time test takers (see Table 4). Therefore, the higher mean scores of 2008 retakers over 2008 first-time test takers cannot be explained out by the hypothesis that the 2008 retakers were simply better performers.

Comparing two isomorphic versions

The material used during the 2007 session allowed for testing the effect of using an isomorphic version of an old test, that is, a version that differed from the old test only regarding superficial features. The question was whether a test based on an identical principle as an old test but using different figures would work as new or old material. The battery of tests comprised also a sustained attention test. On this test, applicants had to detect target signs among distractors on a page containing 1560 signs within a time limit (10 min). A version #0 had been used until the 2006 selection session. In the 2007 selection session, an isomorphic version, labeled version #1, was used, with only format changes of the signs.

The question was whether the 2007 retakers, who were familiar with version #0, would have an advantage on version #1 compared to the control group of the 2007 first-time test takers. Results showed a significantly superior mean score for retakers, $t(655) = 5.36$, $p < .001$, with a medium effect size ($d = 0.51$). Thus, only superficial changes in the testing material are not sufficient to prevent the familiarity effect. In this case, transfer of skills could occur.

Discussion

In this paper we investigated whether the mean superiority in test scores for retakers compared to first-time test takers, which has been robustly found in cognitive ability testing, should be interpreted as an improvement in the target ability or as a test-specific improvement. Using data from the operational selection setting for admission into the

French national air transport pilot training system, we provided evidence for the lack of transfer of skills from one test to another. Specifically, applicants who were already experienced with some testing material did show a mean advantage on old testing material, but did not show any advantage on new testing material when compared to "novice" applicants, except on purely isomorphic tests. Moreover, in the next selection session, the new group of retakers scored higher than the first-time test takers on the tests that were first introduced in the previous selection session.

Interpreting the test-specificity of mean higher scores

Our study provides empirical support for the idea that gains in scores between test and retest are not necessarily related to an actual improvement in the target abilities (e.g., Carroll, 1993; Coyle, 2006; Jensen, 1998; Lubinski, 2000). Moreover, the results show (i) the lack of transfer of one test to another test that is supposed to assess a similar target ability, and (ii) the lack of generalization of a test-taking skill that could be applied to any test (e.g., general strategies for responding to multiple choice questions, like reading the whole test first or skipping the difficult items, could have helped retakers on all testing materials).

Compared to other effect sizes in the literature, the present effect sizes are much larger (minimum 0.60 in our paper vs. a mean effect of 0.26 in the meta-analysis of Hausknecht et al., 2007). One interpretation of this result could be that besides the individual familiarity effect, some applicants may have memorized questions and worked together to improve their performances on their second presentation. Nevertheless this factor seems not to be the main factor that explains the retakers' advantage, as in this case the expected effect sizes would be much larger.

At first glance, this test-specificity may appear to contradict Matton et al. (2009). Indeed,in their study of the gain scores for a given sample between two testing sessions,

they concluded in favor of a latent change factor common to different tests. Nevertheless, nothing prevents the effects of a single latent factor to be conditional on specific aspects of the situation. For example, practice can improve performance on a variety of tests, but its effects do not necessarily generalize to new tests.

For Anastasi (1981), increases in scores between two testing sessions could be explained by general factors like a reduction of anxiety or stress or an increase in motivation. Our results are not in favor of such hypotheses. Indeed, if a reduction in anxiety or an increase in motivation had been the main factors, then there should have been a homogeneous effect on all tests, which was not the case. Moreover, retakers of the ENAC airline pilot selection tests are known to be more motivated than first-time test takers because some of the latter only take the tests to experience the selection process. When applicants retake, they are more likely to strive for success. Nevertheless, the test-specificity of the retakers' superiority showed that even for motivated applicants, the skills acquired during test preparation are limited to the specific tests.

Non-invariance of measurement for first-time test takers and retakers

As stated in the Introduction, studies that have investigated the measurement invariance of the test scores at first and second test sessions for the same applicants found results supporting both measurement invariance (Reeve & Lam, 2005) and non-invariance (Lievens et al., 2007). The present approach is totally different in the sense that it is a between-groups study. Nevertheless, our results indirectly support the hypothesis of non-invariance of measurement. Indeed, strict measurement invariance implies equality of mean structures (Byrne, Shavelson, & Muthén, 1989; Vandenberg & Lance, 2000). Thus, the observed difference between the mean scores of the two groups does not support strict measurement invariance between first-test takers' and retakers' scores. Therefore, retakers' scores do not assess the target ability in the same way that first-time test takers' scores do.

Limitations

One question that arises in the light of the present findings concerns the degree to which they are generalizable to other tests or to other populations. First, the lack of transfer of skills observed on three cognitive ability tests does not guarantee a lack of transferability for any test. One could argue that the *g*-loading of the test could be an important moderator. For example, on one hand, a highly *g*-loaded test could be resistant to practice or familiarity effects (Jensen, 1998; te Nijenhuis et al., 2007), and thus lead to similar scores between first-time test takers and retakers whether the test is new or old. On the other hand, a low *g*-loaded test could be more sensitive to practice or familiarity effects, and thus lead to higher scores for retakers whether the test is new or old. Nevertheless, the present study provides evidence for tests that can be seen both as resistant to the transfer of skill acquisition and as sensitive to the specific exposure. Therefore, the link between the likelihood of transfer of skills and the *g*-loading of a test is not straightforward.

Second, these findings about the population of applicants to the French airline pilot selection process could be specific to that population. Indeed, these applicants are well organized (e.g., through Internet forums) to promote the diffusion of any information concerning the tests already used in the selection process. Moreover, the present empirical data showed a case where only one exposure to the testing material was sufficient to induce a familiarity effect. One could also question whether different tests could have led to another conclusion. To answer these questions, large scale empirical studies should be conducted including a large sample of tests and different populations of applicants. Our purpose is not to over-generalize the findings, but only to warn practitioners that such effects can occur.

Practical Implications

First, the difference between first-time test takers and retakers can be used as an indicator of test-specific measurement bias. Indeed, if for the same population the magnitude of the difference is large on certain tests and small or null on others, this could reveal a differential impact of previous experience with the test. Second, the present empirical study shows that, in high-stake selection settings, where people are likely to train themselves before taking the tests, a given test has a "limited shelf-life" and that superfluous changes are not sufficient for renewing the testing material. Third, our findings have implications for the predictive validity of studies. If a test does not involve the same processes when it is taken for the first or the second time, then the corresponding test scores are not supposed to have the same relationship with external criteria. Thus the predictive validity of test scores for a given test and a given population can change depending on the degree of the diffusion of the test materials.

Perspectives

The present article focused on mean differences of composite test scores. Further research could be conducted at the item level to investigate the possible interaction between the group (first-time test takers or retakers) and the test material (old or new) for the item characteristics. In other words, one could test the following hypotheses:

1. the item characteristics are similar between the group of first-time test takers and retakers for new testing material;

2. the item characteristics are different between the group of first-time test takers and retakers for testing material known by the retakers only.

Another interesting approach would be to evaluate the differential predictive validity for the same test but under different conditions (i.e., when a test is new or when it is old). In other words, one could question the differential magnitude of the relationship between

the test scores and an external criterion (e.g., training success).

Finally, because our results suggest that first-time test takers and retakers are not treated on the same basis, research is needed on the possibility of applying correcting norms for retakers.

Conclusion

In short, we reported results showing that familiarity effects observed in a selection setting after a previous exposure to testing material could be limited only to the given tests. Therefore, even though applicants could practice between two selection sessions, this training can be seen as rather superficial, even though leading to substantive mean gain scores. Aptitude tests in a selection setting seem to be closer to achievement tests. Our findings may be related to the general decrease in predictive validity of tests observed, for example, in aircraft pilot training since the 1940s (Hunter & Burke, 1994). For sure, the diffusion of information concerning the content of cognitive ability tests has increased since the beginning of the use of such tests. Meanwhile, the processes involved in solving an item on a given test have evolved. In such a context, selecting applicants based on tests is a challenging issue that needs regular renewing of the test material.

References

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist*, *36*, 1086–1093.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1974). *DAT : Manuel des tests différentiels d'aptitudes - forme abrégée [DAT: Manual of Differential Aptitude Tests - abridged]*. Paris: Éditions du Centre de Psychologie Appliquée.

Blinkhorn, S. F. (1985). *Graduate and managerial assessment : Manual and user's guide*. Windsor: NFER-Nelson.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456–466.

Carroll, J. B. (1993). *Human cognitive abilities*. Cambridge: Cambridge University Press.

Chamorro-Premuzic, T., & Furnham, A. (2010). *The psychology of personnel selection*. New York: Cambridge University Press.

Cliffordson, C. (2004). Effects of practice and intellectual growth on performance on the Swedish Scholastic Aptitude test (SweSAT). *European Journal of Psychological Assessment*, *20*, 192–204.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cole, N. (1982). The implication of coaching for ability testing. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies (part II)*. Washington, DC: National Academy Press.

Coyle, T. R. (2006). Test-retest changes on scholastic aptitude tests are not related to g. *Intelligence*, *34*, 15–27.

Embretson, S. E. (1987). Improving the measurement of spatial aptitude by dynamic testing. *Intelligence*, *11*, 333–358.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*, 373–385.

Hausknecht, J. P., Trevor, C. O., & Farr, J. L. (2002). Retaking ability tests in a selection setting: implications for practice effects, training performance, and turnover. *Journal of Applied Psychology*, *87*, 243–254.

Hunter, D. R., & Burke, E. F. (1994). Predicting aircraft pilot- training success: A meta-analysis of published research. *International Journal of Aviation Psychology*, *4*, 297–313.

Jensen, A. R. (1998). *The* g *factor: The science of mental ability*. Westport, CT: Praeger.

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin*, *95*, 179–188.

Kulik, J. A., Kulik, C. C., & Bangert-Drowns, R. L. (1984). Effects on practice on aptitude and achievement test scores. *American Educational Research Journal*, *21*, 435–447.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retest effects in operational selection settings: developments and test of framework. *Personnel Psychology*, *58*, 981–1007.

Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology*, *92*, 1672–1682.

Lubinski, D. (2000). Scientific and social significance of assessing individual differences: "sinking shafts at a few critical points". *Annual Review of Psychology*, *51*, 405–444.

Matton, N., Vautier, S., & Raufaste, E. (2009). Situational effects may account for gain scores in cognitive ability testing: A longitudinal SEM approach. *Intelligence*, *37*, 412–421.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin*, *89*, 191–216.

Reeve, C. L., & Lam, H. (2005). The psychometric paradox of practice effects due to retesting: measurement invariance and stable ability estimates in the face of observed score changes. *Intelligence*, *33*, 535–549.

Reeve, C. L., & Lam, H. (2007). The relation between practice effects, test-taker characteristics and degree of g-saturation. *International Journal of Testing*, *7*, 225–242.

Sackett, P. R., Burris, L., & Ryan, A. M. (1989). Coaching and practice effects in personnel selection. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 145–183). Oxford: Wiley.

Snow, R. E. (1982). The training of intellectual aptitude. In D. K. Detterman & R. J. Sternberg (Eds.), *How and how much can intelligence be increased* (pp. 1–37). Norwood, NJ: Ablex.

Sternberg, R. J., Ketron, J. L., & Powell, J. S. (1982). Componential approaches to the training of intelligence performance. In D. K. Detterman & R. J. Sternberg (Eds.), *How and how much can intelligence be increased* (pp. 155–172). Norwood, NJ: Ablex.

te Nijenhuis, J., van Vianen, A., & van der Flier, H. (2007). Score gains on *g*-loaded tests: No *g*. *Intelligence*, *35*, 283–300.

te Nijenhuis, J., Voskuijl, O., & Schijve, N. (2001). Practice and coaching on IQ tests: Quite a lot of *g*. *International Journal of Selection and Assessment*, *9*, 302–308.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–69.

Author Note

Nadine Matton, Université de Toulouse, UT2, CNRS, CLLE-LTC, ENAC, Toulouse, France. Stéphane Vautier, Université de Toulouse, UT2, OCTOGONE-CERPP, Toulouse, France. Éric Raufaste, Université de Toulouse, UT2, CNRS, CLLE-LTC, Toulouse, France.

Correspondence: Nadine Matton, CLLE-LTC, MDR-UTM, 5, Allées Machado 31058 Toulouse Cedex 9, France.

E-mails:{nadine.matton;vautier;raufaste}@univ-tlse2.fr

Notes

[1]Especially when it is known that retakers have greater chance to pass the tests. For instance, at the ENAC, circa 50% of the retakers pass the cognitive ability selection stage whereas only 30% of the first-time test takers do.

Table 1

Comparison of means between retakers and the control group of first-time test takers at the 2007 session (sample 1).

2007		Control group (N=517)		Retakers (N=139)		t test		Effect size
Test	Hist.†	Mean	SD	Mean	SD	t	p	d
S0	old	51.12	9.56	57.87	9.60	7.38	<.001	0.70
M0	old	26.27	3.95	28.65	3.60	6.42	<.001	0.61
N0	old	15.33	4.15	19.65	4.34	10.79	<.001	1.03
S1	new	18.08	5.20	18.22	5.12	0.27	.79	0.03
M1	new	23.15	3.59	23.60	3.64	1.30	.19	0.12
R1	new	65.68	12.04	67.84	10.59	1.92	.06	0.18

†Hist. = historical status of the test, that is, old or new for the retakers.

Table 2

Correlations among scores for first-time test takers only within Session 1 ($N = 517$, below diagonal). Correlations for retakers only within Session 1 ($N = 139$, above diagonal).

Variable	S0	M0	N0	S1	M1	R1
S0	-	.25	*.15*	**.54**	.36	.31
M0	.36	-	*.00*	.31	**.45**	*.13*
N0	.30	.23	-	.17	*.06*	*.10*
S1	**.49**	.39	.17	-	.37	.34
M1	.36	**.61**	.18	.42	-	*.10*
R1	.28	.20	.21	.30	.24	-

Note. All correlations are significant ($p < .01$) except those in italics.

Convergent validities are bolded.

Table 3

Comparison of means between retakers and the control group of first-time takers at the 2008 session (sample 2).

2008		Control group (N=674)		Retakers (N=161)		t test		effect size
Test	Hist.[†]	Mean	SD	Mean	SD	t	p	d
S1	old	17.61	5.27	21.41	5.14	8.26	<.001	0.72
M1	old	22.22	4.04	24.80	2.97	7.61	<.001	0.67
R1	old	65.15	12.17	74.30	9.84	8.87	<.001	0.78

Note. All the 2008 retakers were present at the 2007 session.

[†]Hist. = historical status of the test, that is, old or new for the retakers.

Table 4

Comparison of means between two subgroups of 2007 first-time test takers, future 2008 retakers, and those who did not retake.

2007 session	Future retakers ($N = 161$)		Others ($N = 356$)		t-test		effect size
Test	Mean	SD	Mean	SD	t	p	d
S1	18.23	5.15	18.02	5.23	-0.43	.67	.04
M1	22.96	3.33	23.23	3.70	0.79	.43	.07
R1	65.00	11.05	66.00	12.46	0.89	.38	.08