2007

Practice Effects on Test-Takers' Performance and Quality of Cognitive Domains Should Be Part of Every Pilot Assessment. Investigation of the Mechanisms of Practice Effects Ability Tests in Pilot Selection: A Spatial Ability Test as an Example

Frank Albers

Stefan Hoeft

PRACTICE EFFECTS ON TEST-TAKERS' PERFORMANCE AND QUALITY OF COGNITIVE ABILITY TESTS IN PILOT SELECTION: A SPATIAL ABILITY TEST AS AN EXAMPLE

Frank Albers
German Aerospace Center (DLR), Aviation and Space Psychology
Hamburg, Germany

Stefan Hoeft
German Aerospace Center (DLR), Aviation and Space Psychology
Hamburg, Germany

This study deals with the problem of retaking identical or parallel mental ability tests. This can lead to difficulties in the assessment for prestigious jobs like pilot or ab initio pilot candidate positions, where test preparation is common and a large training industry has been established. We investigated practice effects on test-takers' performance and reliability as well as validity of a spatial ability task. The task was administered ten times, five minutes each, in a sample of 156 ab initio pilot applicants. A performance plateau was reached after the fifth trial, reliability and validity were not affected negatively, they even tend to rise. Consequences for diagnostics are discussed and a brief outlook on the incorporation of the spatial ability task in a multiple task performance test battery is given.

Introduction

There is a long tradition of applying mental ability tests in the personnel selection for flight deck jobs. The fact that the pilot's job is very prestigious and well-paid leads motivated applicants to prepare themselves as intensively as possible for the assessment for flight deck positions or ab initio trainee programs for these positions. In recent years a growing training-industry has become established in Germany for these assessments which provides practice materials in the form of books, training courses and, above all, training software for computerized cognitive tests. Combined with the fact that in the information age nothing, including tests or testing-principles can be kept secret, there is a general problem with preparation in the assessment of cognitive abilities in this field. Test fairness is not guaranteed when different applicants have different experience with the tests applied. In terms of reliability and validity of the tests it is unclear whether a test produces stable measures after extensive practice of test-takers and whether this test still measures the intended construct. We believe that these problems tend to be ignored by practitioners and not much systematic research has been done in this problem-field.

Theoretical background for test practice

The term "test practice" must be distinguished from "test coaching". Practice simply means repeatedly taking the same test or working on the same material, respectively. Coaching on the other hand involves systematic intervention between trials of test-taking in the form of detailed feedback, teaching of item-solving strategies and so on (Sackett, Buris & Ryan, 1989). Both forms of training can be seen as the ends of a continuum of training forms (Messick, 1981). Our research reported here clearly focuses on the effects of practice in the context of computer-based mental ability tests.

Kulik, Kulik & Bangert (1984) conducted a meta-analysis in which they examined 40 studies dealing with practice effects on different intelligence and university entry tests. They measured the effect size of practice (d) by subtracting the first test result from the second and dividing this score by the standard deviation of the first measure. They showed that performance increased up to the sixth application of a test. The effect sizes for the first repetition were d=.42 for identical and d=.23 for parallel test forms, the effect sizes for the sixth realization were d=1.94 (identical test forms) and d=.73 (parallel test forms). After the sixth testing performance scores stabilized and no further performance gain was observed. In addition to these results Sackett et al. (1989) report that while practice effects are prevalent in virtually all types of mental ability tests they are especially large for tests of psychomotor-coordination and spatial orientation. These are two of the most important basic abilities for flight deck jobs (Goeters, Maschke & Eissfeldt, 2004) and tests for these domains should be part of every pilot assessment.

Throughout the scientific debate concerning investigation of the mechanisms of practice effects three different reasons for performance gain have been discussed (cf. Lievens, Buyse & Sackett, 2005):

1. Practice leads to a reduction of test-irrelevant inhibitory influences (e.g. test anxiety or growing familiarization with test setting).
2. The test score increases due to influences which are construct-irrelevant, e.g. memory

effects for identical test forms or discovery of tricks which invalidate the test principle.

3. Practice leads to a "true" increase of the tested ability (e.g. caused by automation and speeding-up of construct relevant processes).

If reasons 1 and 3 are responsible for the increase of test scores the construct validity of the test is preserved or even increased as well. If reason 2 is responsible the validity of test scores is diminished. In general it must be assumed that all three reasons contribute to the increase, although Reeve and Lam (2005) recently showed that the structure of the latent ability-related variables remained constant in an analysis of a test repetition (3 test applications). But as their methodology raises doubts (and three repetitions cannot be seen as extensive practice), the answers to the questions "what leads to increased test scores?" and "what do mental ability tests measure after repeated applications?" remain ambivalent and unclear.

Objectives and Hypotheses

The department of aviation and space psychology at the German Aerospace Center (DLR) conducts assessments for flight deck positions (ready entry and ab initio pilot applicants) of several airlines. The problem of test preparation and test practice is very prevalent here and has to be dealt with. The investigation of practice and coaching influences on tests and other diagnostical methods is part of the regular scientific evaluation and an element of the quality management system.

In this study this evaluation is described exemplarily with a new spatial orientation task. This task will be part of a test battery for the assessment of multiple task capacity but can be applied as a single test for spatial orientation as well.

Two hypotheses must be corroborated:
Hypothesis 1: Test score gains for a repetitive application of the test will decline over the course of applications. After that test scores remain constant and cannot be further increased.
Hypothesis 2: The test's reliability and validity are not affected by extensive practice.

Method

Subjects

Subjects were 157 applicants for ab initio pilot trainee positions at the German Lufthansa AG. Data was obtained during the first selection phase, where basic abilities are assessed via computer based tests. One subject's data set showed that he obviously had not understood the test instructions, so this data set was not evaluated. N=156 data sets were analyzed accordingly. The sample consisted of n=134 male (85.9%) and n=22 female (14.1%) subjects with an average age of 21.36 years (SD=2.08).

Materials

The task in question is called "Relative Position" (REP) and it is intended to measure two aspects of spatial abilities: spatial orientation and visualization (cf. Fleishman, 1992).

One Item of the REP consists of a pictogram of an aircraft, which can be turned in one of 12 positions, comparable to the 12 positions of a clock face. In relation to this aircraft a small object, a point, is displayed. This point can also be placed in one of the 12 clock face positions. Figure 1 displays one example item.

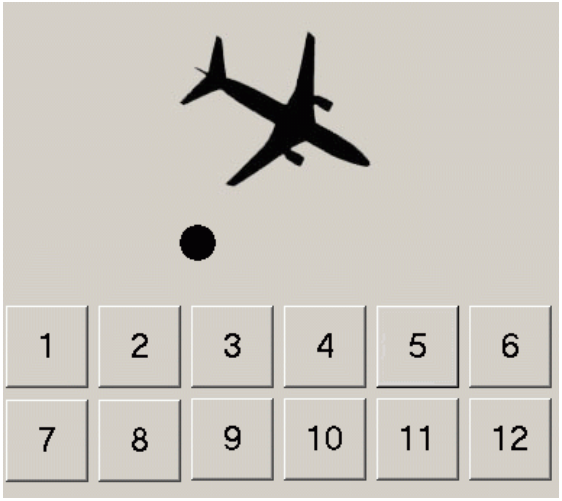

Figure 1. REP example item with keyboard for solution input (solution: 3).

The subject's task is to give the position of the point in relation to the center of the aircraft using one of 12 clock positions. The solution is to be entered on a displayed keyboard on a touch sensitive computer monitor. The task is self-paced and subjects are instructed to solve as many items as possible in the 5-minute testing-time. The subjects had no possibility to practice on the REP in advance, as the test was wholly new.

Furthermore the subjects worked on 12 tests for diverse mental abilities and knowledge domains and on a personality inventory as regular parts of the

assessment. These tests were relevant for the selection decision regarding the next phase of the assessment process. The subjects had the possibility to practice these cognitive ability tests via computer based trainings (CBTs) in advance to prepare for the assessment. Some of the ability tests were used as references for the validation of the REP. The relevant tests are introduced in the results section.

Procedure

The assessment stage consisted of two half days in which the subjects completed the computer based tests. After the regular assessment program the REP was applied. The application consisted of a computerized instruction and 10 repetitive trials of an identical test form of the REP. This procedure guaranteed extensive and massed practice. Variables registered were sums of correct, false and total solutions (items processed) per trial and reaction times for each item.

The REP was introduced like a regular test and the subjects had to believe that this test was part of the regular assessment. In this way motivation for good results was kept high.

Results

The primary variable analyzed was the sum of correctly solved items per trial. The absolute amount of mistakes was very small (M=5.9 over all trials). The total sums of processed items were closely correlated with the sums of correctly solved items. Therefore, an analysis of these variables would have been redundant. Reaction times were used in parts of the validation analyses.

Practice effects

Figure 2 shows the averaged sums of correctly solved items. A clear increase can be observed and an analysis of variance with repeated measures was significant accordingly (*Pillai-Spur*=.931; F=221.18; p=.000).

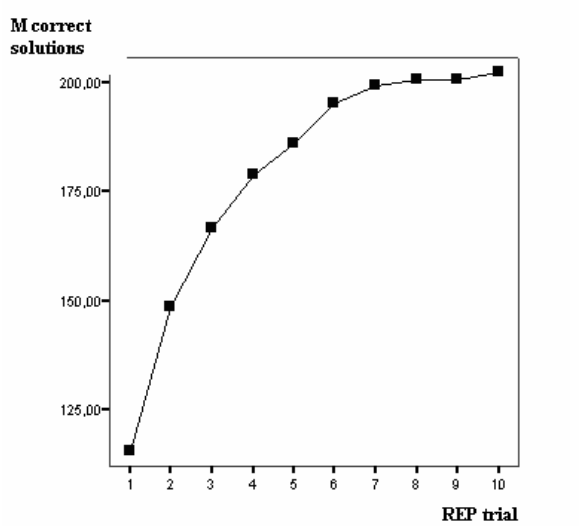

Figure 2. Developing of averaged sums of correct responses over the 10 REP repetitions.

According to Kulik et al.'s (1984) procedure effect size measures d were calculated. These effect sizes rise from d=1.161 (first and second trial) to d=2.815 (first and fifth trial). After that, d-measures do not rise considerably. This can be seen in figure 2 as well: The averaged sums of correct solutions form an asymptote. The performance reaches a plateau with no further practice effects. Although paired-sample T-Tests revealed significant differences between trials 6, 7, 8, 9 and 10, respectively (all p<.05) the effect sizes reveal that there are no changes of practical significance, with the largest d=.21 and the smallest d=.005 (cf. Cohen's, 1988, classification of effect sizes).

Reliability of test scores

Table 1 shows the correlation matrix for the test scores of the 10 REP trials. The correlations can be interpreted as retest reliability coefficients. The coefficients are altogether high and they stabilize at a very high level after the fifth trial with a minimum of r=.886 (fifth and tenth trial) and a maximum of r=.956 (ninth and tenth trial).

Table 1. Inter-correlations of REP trials. All correlations $p<.01$.

	1	2	3	4	5	6	7	8	9	10
1	1	.867	.782	.770	.721	.727	.685	.652	.635	.612
2		1	.916	.900	.861	.847	.804	.766	.760	.739
3			1	.945	.925	.888	.869	.845	.851	.834
4				1	.954	.933	.910	.881	.893	.876
5					1	.935	.915	.894	.899	.886
6						1	.940	.904	.902	.894
7							1	.937	.929	.920
8								1	.948	.939
9									1	.956
10										1

Validity of test scores

Two kinds of validity analyses were performed: At first a correlational analysis of convergent and discriminant validity of REP test scores using the regular tests as references was conducted. Thereafter, a "test-immanent" examination of construct validity, i.e. if mental rotation is still used by the subjects after practice, was performed.

Convergent and discriminant validity. First correlations with the construct nearest reference test were examined. This is a spatial abilities test demanding mental rotation of a dice. Figure 3 shows the correlations of the two test scores over the ten REP trials. The coefficients are not only stable but rather tend to rise from $r=.233$ in the first trial to $r=.379$ in the tenth trial. Although the minimum and maximum correlation are not significantly different ($p=.074$) the trend is obvious.

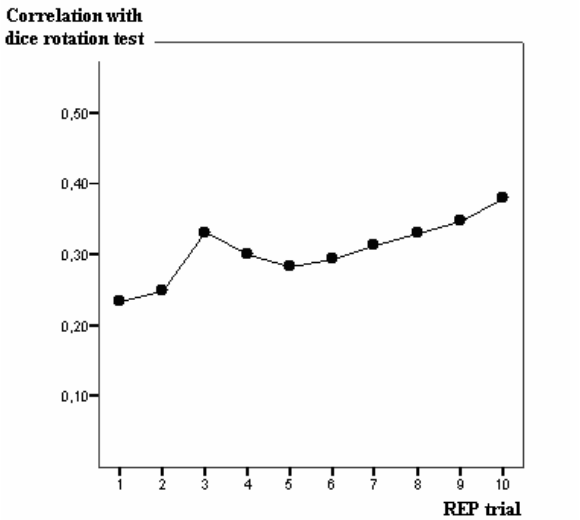

Figure 3. Correlation of REP test score with reference test "Dice Rotation" over the 10 REP repetitions.

Correlations with a second spatial abilities test were investigated as well. This test involves mental manipulation and comparison of unfolded dice. The correlations with the ten REP scores were minimally $r=.283$ (tenth trial) and maximally $r=.346$ (third trial) with no significant difference between these extreme correlations ($p=.271$).

Relations to other measures of mental abilities were investigated as well: Correlation with a measure of concentration rises over practice to a maximum of $r=.353$ in the tenth trial. Other significant correlations with the test score of the tenth REP trial were: Perceptual speed $r=.295$; mental arithmetic $r=.255$ and acoustical working memory $r=.213$.

Correlations with diverse knowledge domains (English language, technical knowledge and mathematical knowledge) were not significant and around $r=0$.

Additionally conducted factor analyses (PCA with varimax rotation) with all reference tests and single REP trials supported the previous findings: When the REP score of the last practice trial was included in the analysis a simple structure with all cognitive ability measures as first factor showed up.

Mental Rotation. This test-immanent analysis of validity explored if the task-intended strategy of working on the REP (i.e. first mental rotation of the pictogram to the 12 o'clock position and then determination of object position) was used even after extensive practice on the task. If this was the case then items with the pictogram further away from the 12 o'clock position should take the subjects more time to solve than those items with the pictogram nearer to the 12 o'clock position (Shepard & Metzler, 1971). For this analysis all reaction times for all items with the same position of the aircraft pictogram were aggregated for all subjects for each trial, regardless of the relative object position.

Figure 4 shows the means of the reaction times (RT in ms, and the 95% CIs) for the 12 clock positions of the pictogram for the first and last REP trial. In both trials the RTs for the 6 o'clock position, which is the position furthest away from 12, are the largest. And in trial 10 the RTs are graded over the positions: The bigger the distance from 12 the larger the RT.

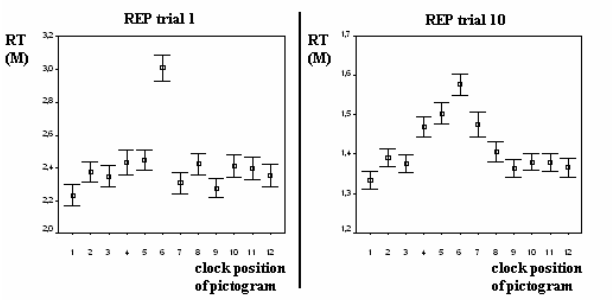

Figure 4. RTs (ms) on items with same clock position of pictogram, first and last REP trial.

Discussion

Our two hypotheses concerning practice effects were tested in a sample of 156 ab initio pilot training applicants using a new spatial ability test (REP).

The performance reaches an asymptote after the fifth trial and from there on the increases in performance have no practical significance.

The reliability of test scores rises with practice and reaches high values. The test validity is given even after extensive practice of 10 test trials: Test scores show correlations to similar construct measures. Furthermore, reaction times increase with rising rotational angle of item material. This is consistent with the theory of mental rotation (Shepard & Metzler, 1971), and this effect is even more accentuated *after* extensive practice.

Interpretation

The results show that subjects taking a cognitive ability test, in our case a spatial ability test, can be trained (under controlled conditions) to a level where no further increase in performance can be obtained. At the same time this test retains its reliability and validity. Therefore it can be assumed that increases in performance on the REP result from cognitive processes described by mechanisms 1 and 3 (cf. introduction part of this contribution), i.e. reduction of test-irrelevant influences and increase of the "true" ability in question.

Limitations

The results were obtained with a test for spatial abilities and have to be replicated with tests of other domains to make general statements on cognitive ability tests. Furthermore, practice was totally controlled and massed in our setting. Whether the plateau is stable over longer time periods or whether distributed training has different effects remains unclear.

Prospect

Uncontrolled test practice before assessment is a serious problem for diagnostics, especially in the field of prestigious jobs like flight deck positions in an airline. In our view this problem can be controlled by proper development of tests and evaluation of test quality, especially including examination of practice effects on performance, reliability and validity as in this study. With knowledge about practice effects on performance on a given test it is possible to re-establish test fairness by providing material for preparation. This can be done by making CBTs for computerized tests available and giving guidelines for proper practice.

Our research has shown that REP is a reliable and valid test for spatial abilities with good quality even after practice. Test-takers reach a plateau of performance after the fifth trial. With these results at hand a general practice recommendation can be stated: Applicants should practice this test's CBT at least five times. The practitioner using this test can be quite sure that practice beyond this guideline does no harm to the test's quality.

The future prospect of the REP is as follows: The REP is part of the development of a new test battery for multiple task performance abilities. This battery will be modular, which means that the multiple task ability test will consist of several modules (up to three) that have to be worked on simultaneously. The aim is to construct different modules which for themselves are reliable, valid and practice-resistant tests for basic requirements for (ab initio) pilots. So far two more modules besides the REP have been developed, one for psychomotor-coordination and one for perceptual speed and they have been evaluated in the same way with good results.

References

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

Fleishman, E. A. (1992). *The Fleishman Job Analysis Survey (F-JAS)*. Palo Alto: Consulting Psychologists Press, Inc.

Goeters, K.-M., Maschke, P. & Eissfeldt, H. (2004). Ability requirements in core aviation professions: Job analyses of airline pilots and air traffic controllers. In K.-M. Goeters (Ed.), *Aviation psychology: Practice and research* (pp. 99-122). Aldershot, UK: Ashgate.

Kulik, J. A., Kulik, C. C. & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21,* 433-447.

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

Messick, S. (1981). The controversy over coaching: Issues of effectiveness and equity. In B.F. Grenn (Ed.), *New directions for testing and measurement: Issues in testing - coaching, disclosure, and ethnic bias, No. 11* (pp. 21-53). San Francisco: Jossey-Bass.

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.

Sackett, P. R., Burris, L. R. & 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). Chichester: Wiley.

Shepard, R. N. & Metzler, J. (1971). Mental Rotation of Three-Dimensional Objects. *Science, 171,* 701-703.