Minimizing Extraneous Factors That May Impact Computer-Based Test Validity

Michael C. Heil
Jillian Petro

Caliber Associates

Overview of Issues

Extraneous factors that may affect computer-based test validity include:

- Computer Experience
- Practice and Coaching
- Software and Hardware changes

Computer Experience

- There has been relatively little written in the scientific literature about the use of computerized tests by organizations.
- The effect of computer experience on computerized test performance has received more attention in the education literature.
- Two primary issues related to computerized testing:
  - **computer anxiety** (Bradley & Russell, 1997; Dimock & Cormier, 1991; Legg & Buhr, 1992; Levine, & Donista- Schmidt, 1998; Powers & O'Neill, 1993)
  - **equivalence of paper-and-pencil and computerized tests** (Finegan & Allen, 1994; Mead & Drasgow,1993).

Computer Experience

- Two findings have generally emerged:
  - People with more computer experience reported less computer anxiety, while people with higher levels of computer anxiety failed to perform as well on computerized tests
  - Differences in performance between paper-and-pencil and computerized tests were negligible.

Practice and Coaching

- The distinction between practice and coaching lies in whether or not there is an outside intervention.
- When the accuracy of a selection device is compromised, less informed hiring decisions are inevitably made.
- If practice and/or coaching increases an applicant's score on a selection test, then the predictive validity of that selection device is undermined (Sackett et al., 1989).

5

Practice and Coaching

- Mauer et al. (1998) illustrated that coaching is related to successful performance in structured interviews, which are a widely used method of selection.

- Alliger, Lillienfeld, and Mitchell (1996) reiterated that coaching on overt integrity tests is related to significant increases in test scores.

- According to Furhnam (1997), tests that measure personality or temperament, also used in some organizations for the purpose of selection, have been made available to the public so that people can practice.

Practice and Coaching

The implications for personnel selection:

- Practice and coaching may interfere with the ability of a test or test battery to accurately predict future job performance

- Test scores that improve due to practice and coaching effects might influence hiring decisions based on test performance.

Changes to Hardware & Software

- Limited published research on this topic.
- Most of this research is between 10-20 years old.
- Relevant issues include software upgrades, operating system, screen size, fonts, colors, and changes to "look and feel".
- Research literature addressed font size, display size, display color, and changes from paper-and-pencil to computer.

Changes to Hardware & Software

- Practical issues are readability and whether or not changes make the test easier.
- According to Wildstrom (1998), differences in monitor size can be offset by changing the resolution and object size.
- No differences in performance due to:
  - Display size (Duchnicky & Kolers, 1983)
  - Reading ease (Display had minimal impact on reading speed) (Gould et al., 1987)
  - Format (Paper-and-pencil vs. computer (Honaker et al., 1988)

Applied Example

The FAA's Air Traffic-Selection and Training (AT-SAT) test

The Air Traffic Selection and Training (AT-SAT) test battery is the Federal Aviation Administration's (FAA's) computerized selection test for Air Traffic Control Specialists (ATCSs).

It is comprised of seven tests of cognitive ability and one Non-cognitive measure.

AT-SAT Tests

- Applied Math
- Scan
- Angles
- Analogies
- Letter Factory
- Dials
- Air Traffic Scenario Test
- EQ

Computer Experience

12

Objectives of this study*:

- Determine if examinees with more computer experience perform better than examinees with less computer experience.

- Determine if the relationship between computer experience and computerized test performance remains consistent, regardless of the type of test being presented in the computerized format.

* Heil & Agnew (2000)

13

Correlation of AT-SAT Subtests with Computer Experience

Test	Computer Exp. Score
Dials	0.196
Angles	0.266**
Amath	0.372**
LF SA	0.447**
LF TP	0.283**
Analogies	0.339**
Scan	0.199
AT Eff.	0.365**
AT Safe	0.207*
AT PA	0.054
AT Total	0.318**
AT-SAT	0.380**
Composite	

** Correlation is significant at the 0.01 level (2-tailed)

* Correlation is significant at the 0.05 level (2-tailed)

Hierarchical Multiple Regression of Age, Education, and Computer Experience on Composite AT-SAT Score

	ATSAT Comp.	**Age**	**Educ.**	**B**	β	**ΔR^2**	R^2=.258*
Age	-.045			-.196	-.048	.002	**Adjusted R^2=.233**
Education	.456**	.129		4.642	.375	.217**	
Computer Experience	.380**	.205*	.389**	.317	.224	.039*	R=.508

*p<.05

**p<.01

15

Summary of Results:

- The results provide evidence of a relationship between prior computer experience and performance on a computerized personnel selection test

- Once education level was controlled for, computer experience was correlated with performance on dynamic tests that required the use of a mouse.

Practice and Coaching

17

Objectives of this study*:

- Determine if repeated test taking improves performance
- Determine if coaching improves performance
- Identify specific tests within the AT-SAT battery that are most susceptible to practice and coaching effects
- Determine the extent to which practice and coaching effects potentially impact hiring decisions.

* Heil, Detwiler, Agen, Williams, Agnew, and King (2002)

18

Testing Schedule

Group	**Treatment**
Coaching Prior to Testing	Coach → Test 1 → Test 2 → Test 3
Coaching After Testing	Test 1 → Coach → Test 2 → Test 3
Control	Test 1 → Test 2 → Test 3

Notes: 1. Coaching always occurred the day before the next testing session.

2. All test sessions were 3 weeks apart.

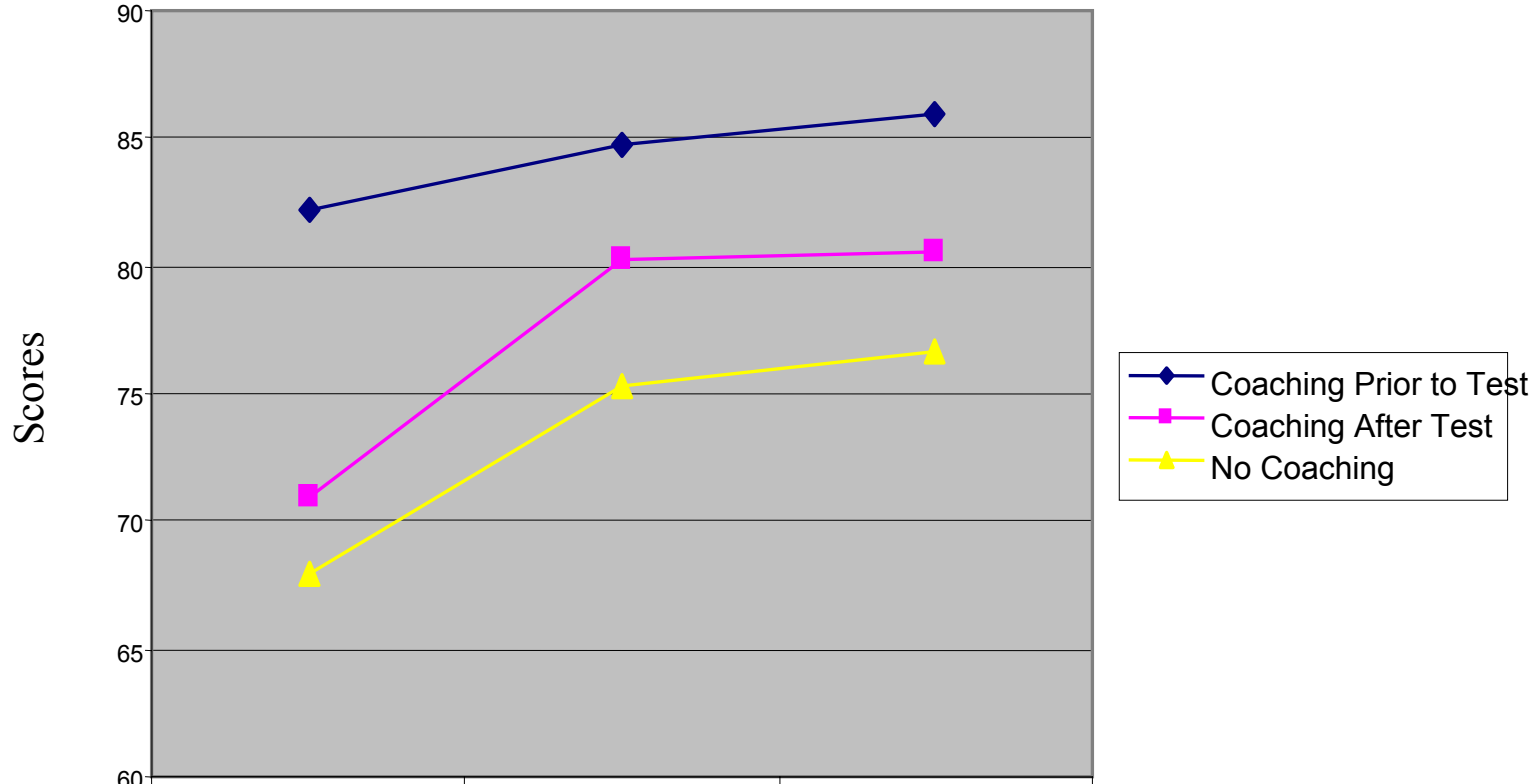
19

Mean AT-SAT Score for Each Experimental Group

Group	1	2	3
Control	68.0*	75.3*	76.6*
Coaching After Testing	71.0*	80.2	80.6
Coaching Prior to Testing	82.2	84.7	86.0

* Mean scores that are significantly different than the *coaching prior to testing* group mean ($p<.05$).

Plot of AT-SAT Weighted Composite Score



	Testing 1	Testing 2	Testing 3
Coaching Prior to Test	82.2	84.7	86
Coaching After Test	71	80.2	80.6
No Coaching	68	75.3	76.6

AT-SAT Practice Index

AT-SAT (Unweighted Composite)	**0.4336**
AT-SAT (Weighted Composite)	**0.4423**
AT Efficiency	0.8103
LF – Situational Awareness	0.5430
Dial Reading	0.5030
AT Safety	0.4921
AT Procedural Accuracy	0.4064
Scan	0.3540
Analogies	0.2082
Angles	0.2002
Applied Math	0.1819
EQ Composure	-0.0085
EQ Self-confidence	-0.0096
EQ Interpersonal Tolerance	-0.0774
EQ Working Cooperatively	-0.0995
LF – Planning and Thinking Ahead	-0.1114
EQ Consistency of Work Behavior	-0.1591
EQ Decisiveness	-0.1638

AT-SAT Coaching Index

AT-SAT (Unweighted Composite)	**1.0610**
AT-SAT (Weighted Composite)	**0.7200**
EQ Consistency of Work Behavior	1.2010
EQ Interpersonal Tolerance	1.0470
EQ Working Cooperatively	1.0185
EQ Decisiveness	0.9553
EQ Composure	0.8372
EQ Self-confidence	0.6164
AT Efficiency	0.5696
LF – Situational Awareness	0.5370
Angles	0.4819
Scan	0.4763
LF – Planning and Thinking Ahead	0.4615
Applied Math	0.3340
Analogies	0.3157
Dial Reading	0.3059
AT Safety	0.2373
AT Procedural Accuracy	0.1728

Change in Rank Order of Top Ten Candidates Following Coaching

Hiring Decisions			
Treatment	Time 1	Time 2	Treatment
Control	**A**	E	Coached
Coached	B	B	Coached
Control	**C**	H	Coached
Control	**D**	I	Coached
Coached	E	**C**	**Control**
Coached	F	**A**	**Control**
Coached	G	**D**	**Control**
Coached	H	F	Coached
Coached	I	G	Coached
Control	**J**	**J**	**Control**

Summary of Results:

- The results suggest that performance on the AT-SAT battery may indeed be influenced by both practice and coaching effects.

- The composite AT-SAT score that is used for hiring decisions increases with repeated administrations, although the greatest increase occurs following coaching.

25

Summary of Results:

- Performance-based tests were affected more by practice than were tests that required knowledge or abilities not measured by computer simulations.
- The non-cognitive test, the EQ, was most susceptible to coaching effects.
- The large increase in score on the EQ scales following coaching suggests that the participants were able to easily learn how to fake well on this measure of personality in the workplace.

26

Changes to Hardware & Software

- AT-SAT was originally created on a DOS platform. This platform must be run on an older operating system (Windows 95).
- This platform was recently upgraded to a Windows platform.
- The new platform can run on a laptop, as well as newer versions of window.
- Changes included improved “look-and-feel”

28

- Although the items have not changed, has the test changed enough that the validity results don't generalize?
- Does the change in monitor size affect performance on the dynamic scenario-based tests?
- FAA plans to conduct an equating study of the old and new platforms before using the upgraded version operationally.

29

- The impact of hardware and software changes on AT-SAT performance is still under investigation.
- Results of computer experience and practice and coaching studies suggest that these extraneous factors could have an impact on test performance.
- The impact of these factors varies based on the type of test.
  - Performance on dynamic scenario tests most likely to be affected by practice effects.
  - Non-cognitive measure of personal characteristics most coachable

30

Conclusions

- These research results suggest that computer experience, practice, and coaching can all negatively effect selection test validity.

- Limited published research literature indicates that software and hardware upgrades are likely to have minimal impact on test validity.

31

Conclusions

Extraneous factors to consider:

- Pre-screen computer (mouse) training
- Mouse exercises prior to scored portion of the test
- Alternate forms of the test
- Test security
- Non-transparent items