Situational effects may account for gain scores in cognitive ability testing: a longitudinal SEM approach

Nadine Matton, Stéphane Vautier, Éric Raufaste

Running head: SITUATIONAL EFFECTS IN ABILITY TESTING

Situational Effects May Account for Gain Scores in Cognitive Ability Testing: A Longitudinal SEM Approach

Nadine Matton, Stéphane Vautier and Éric Raufaste

Université de Toulouse

Abstract

Mean gain scores for cognitive ability tests between two sessions in a selection setting are now a robust finding, yet not fully understood. Many authors do not attribute such gain scores to an increase in the target abilities. Our approach consists of testing a longitudinal SEM model suitable to this view. We propose to model the scores' changes of a battery of tests between two sessions with a single factor, namely the change in the *situational* component of the scores. The situational component encompasses all effects due to the specificity of the state of the person in the current situation (e.g., anxiety level, tiredness, test-taking practice) and is allowed to vary from one session to another. By definition, this single component is supposed to influence all tests at a given session. In particular cases such as high-stake selection settings, where applicants are likely to train themselves before retaking the tests, situational factors might even suffice to explain mean score increases. Empirically, our latent change model closely fitted the scores of 752 applicants for entry into the French Aircraft Pilot Training, gathered on a set of three tests (visual perception, mechanical comprehension, and selective attention). Gain scores of moderate to strong effect sizes could be explained by common situational effects, with no need for admitting change on ability components. Therefore, gain scores may be understood as construct-irrelevant changes.

Situational Effects May Account for Gain Scores in Cognitive Ability Testing: A Longitudinal SEM Approach

The Issue of Gain Scores

Since the beginning of the 20th century, cognitive ability tests are widely used in military and civil organizations for the selection of adult applicants (Domino & Domino, 2006). As selection rates are usually low (far more applicants than selected people), in case of failure on a first testing session, many organizations allow applicants to retake the tests (Lievens, Reeve, & Heggestad, 2007; Lievens, Buyse, & Sackett, 2005). As shown by numerous test-retest studies in the field, substantive mean gain scores can be expected (e.g., Hausknecht, Trevor, & Farr, 2002; Kulik, Kulik, & Bangert-Drowns, 1984). A recent meta-analysis (Hausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007) revealed a mean effect size of +0.26 standard deviation from one test administration to the next. Finding a proper interpretation of gain scores remains an issue.

Investigating Gain Scores

We will review current approaches to gain scores before presenting our approach based on longitudinal true-score modeling. First of all, we put forward one of the main assumptions of this paper, namely the temporal stability of broad cognitive abilities within a short time interval like one year.

Broad cognitive abilities stability. Concerning human cognitive abilities, there is a relative consensus about Carroll (1993)'s three-stratum hierarchical model:

1. general factor *g*, defined as the first factor extracted after factor analysis conducted on a battery of mental ability tests,

2. broad abilities (also named "stratum II"), defined as very general abilities like fluid intelligence, crystallized intelligence, general memory and learning, broad visual perception, broad auditory perception, broad retrieval ability, broad cognitive speediness and processing speed,

3. narrow abilities (also named "stratum I"), each being a specific component of a broad ability (e.g., sequential reasoning, induction, quantitative reasoning, and Piagetian reasoning are related to fluid intelligence).

It is generally assumed that, in the absence of any specific intervention (e.g., learning, practice, coaching, etc.), broad cognitive ability rankings are relatively stable from one age to another (Carroll, 1993, p. 662). Second, inter-individual differences in mental abilities are supposed to reach a fairly high level of stabilization by early adolescence (Dixon, Kramer, & Baltes, 1985; Jensen, 1998). Therefore cognitive abilities of young adults are supposed to have reached a plateau. In this case, or when the time interval between two assessments is small compared to life-span (e.g., one year), the problem is to explain the mean gain scores observed after retesting (Hausknecht et al., 2007).

Current approaches to gain scores. Many authors consider that gain scores in cognitive ability testing, even after coaching or practice, are not necessarily related to an increase of the assessed abilities (Anastasi, 1981; Cole, 1982; Jensen, 1998; Snow, 1982; Sternberg, Ketron, & Powell, 1982; Lubinski, 2000). The gain scores could be due to an increase of “test-wiseness”, test-familiarity, motivation, self-confidence, or sheer test-taking practice and/or a reduction of stress or anxiety. In reasoning tasks for instance, Roberts and Newton (2003) identified *task-specific short-cut strategies* that could reduce effort and stress thus increasing performances.

Considering a variety of tests, Jensen (1998) investigated how the mean gain score observed on a test relates to the *g*-loading of that test, and found a negative correlation.

Thus, the more a test is *g*-loaded, the less it is likely to allow for gain scores. In a recent meta-analysis of 64 test-retest studies using IQ batteries, te Nijenhuis, van Vianen, and van der Flier (2007) replicated this finding. Moreover, Jensen discarded the possibility that variability in mean gain scores observed through a range of various tests over various time lags would be related to the *g*-loadings of these tests, which suggests that mean gain scores are not associated with *g*.

Recently, other authors have studied the test-retest effects obtained with a battery of tests, through the test of measurement invariance hypotheses (Reeve & Lam, 2005; Lievens et al., 2007). Their conclusions were mixed, supporting both structural invariance (Reeve & Lam, 2005) or non-invariance (Lievens et al., 2007). Neither Jensen's nor Reeves et al.'s and Lievens et al.'s approaches focus on the structure of individual change effects, however. Our own approach is an attempt to elaborate and test a longitudinal model that specifies the structure of individual test-retest increases on a battery of tests.

A Longitudinal True-Score Approach to Gain Scores

Assuming the Classical Test Theory decomposition of the observed-score variable into a true-score variable and an error variable (Lord & Novick, 1968), if the regression-to-the-mean effect is negligible, one cannot explain gain scores by measurement error (e.g. Lievens et al., 2005), since the error variable has a null expectation (e.g., Lord & Novick, 1968; Steyer, 1989; Zimmerman, 1975).[1] Thus, gain scores are true gain scores. The true score refers substantively to a construct-relevant component (linked with the target ability) and to a construct-irrelevant component (with no link with the target ability level), i.e., 'true' does not necessarily mean 'construct relevant' (Borsboom & Mellenbergh, 2002; Zumbo & Rupp, 2004).

The key idea of the present approach is to model the construct-irrelevant component of the true score of a test (re)taker as an individual *situational effect*. Situational effects

include all effects due to the specificity of the situation in which the person takes the test and which induces a given *state* of the person. To illustrate how this notion applies to true scores, let us take Jensen's example of the indirect height measure through the shadow length of a person in the sunlight (1998, p. 312). Depending on the time of day, for a given height, the shadow will have a different length. Thus the time of the measurement is a situational factor. Measuring the shadow height of a person twice, but at different times of day, different shadow lengths will be found, although the person's height did not change. Analogously, test-retest gain scores may be interpreted as changes in situational effects on the person's states, while the cognitive traits remain stable.

Situational effects were first formalized within the SEM framework in the Latent State-Trait Theory (Steyer, Ferring, & Schmitt, 1992; Steyer, Schmitt, & Eid, 1999). This theory states that any test score measures characteristics of the person (*traits*), but also measures characteristics of the situation and characteristics of the interaction between person and situation. Taken together these factors create a psychological state specific to the situation to which the person is exposed. Following this theory, a test never measures trait differences only but also individual differences due to situational effects.

Within the context of testing, there could be some individual situational features common to every test in Session 1: anxiety, stress, self-confidence, test familiarity level, general test-taking practice, and the like. Those common features may change to some extent from Session 1 to Session 2. This temporal change would be common to all tests because it would be part of the situation to which participants are exposed to rather than specific to any single test.[2] Following the preceding "shadow-illustration", let us imagine a battery of height measures that are taken at a given time of day (e.g., height of the head, of the body, of the arm). Taking the same measures six hours later the next day, change observed for the head height is linearly related to change observed for body height.

To our knowledge, the possibility that change in situational effects could account for

individual gain scores has not been investigated in previous research. In the sequel, we investigate how far one can go using the situational effects view as an account for gain scores effects.

Elaborating such a view requires some mathematical developments. We begin by formally analyzing the test scores supplied by a battery of tests in a test-retest design, which allows for the definition of the concept of a reliable measurement artifact, in contrast to the reliable construct relevant component of the true score. Then, we identify sufficient conditions in the structure of the situational effects for deducing testable statistical consequences through SEM. Finally, we test the modeling on data from a sample of test retakers, applicants for entry into French Aircraft Pilot Training.

Decomposing Cognitive Ability Test Scores

We now present a SEM decomposition of cognitive ability test scores. This decomposition is a necessary step for the purpose of testing the hypothesis. However, the reader who is not familiar with SEM may skip to the next section.

Considering Y_{it} the test-score variable observed with test i at time t, where i indexes a number of cognitive ability tests and $t \in \{1,\ 2\}$ denotes the test and retest sessions, the usual true-score decomposition is

$$Y_{it} = \tau_{it} + \varepsilon_{it}, \tag{1}$$

where τ_{it} and ε_{it} denote the true-score and measurement error variables, respectively. Let the true scores associated with scores on the test i at time t be defined as the expectations of the test scores when subjects are measured with the test i at time t. Thus, the test score is thought of as a probabilistic event, and measurement error represents the stochastic component of the measurement experience.

Firstly, we consider two construct relevant and reliable sources of inter-individual variations, namely the standing on g, the general cognitive ability as defined in the g factor

literature, and the standing on a_i, the cognitive ability which is specifically measured by the test i (Carroll, 1993).

Secondly, we define a third component s_{it}, representing all reliable inter-individual variations that do not pertain to the ability the test i is purported to measure. Test anxiety, the degree of familiarity with the test, emotional or attentional transient dispositions due to real-life events around the testing session, etc., are substantive factors entering the third component s_{it}. Tautologically, the situational effects express as follows:

$$s_{it} \equiv \tau_{it} - [(\alpha_i + \beta_i \cdot g) + a_i], \tag{2}$$

where $\equiv$ denotes equality by definition, and the coefficients α_i and β_i are used to allow g to predict scores on the tests of the battery specifically. As the paper's purpose is to test the plausibility of retest gain scores' explanation through latent change of situational factors only, we assume the variances on the g and the a_i components are constant across situations; hence, the g-loadings β_i are assumed to be constant across situations.

Consequences of the Test Scores Decomposition

A first consequence of the true-score decomposition is that the situational component is allowed to have a non-null mean, that is, to act as a systematic bias with respect to the construct relevant component $(\alpha_i + \beta_i \cdot g) + a_i$. Thus, the observed score y_{it} of any person suffers from a random bias, i.e., measurement error, and from a systematic bias, i.e., the situational effect.

A second consequence of the true-score decomposition is that the true variance $\sigma^2(\tau_{it})$ of the test scores Y_{it} is a quantitative mix of variances and covariances, as detailed below:

$$\begin{aligned} \sigma^2(\tau_{it}) &= \beta_i^2 \cdot \sigma^2(g) + \sigma^2(a_i) + \sigma^2(s_{it}) \\ &+ 2 \cdot \{\beta_i \cdot [\sigma(g,\ a_i) + \sigma(g,\ s_{it})] \\ &+ \sigma(a_i,\ s_{it})\}, \end{aligned} \tag{3}$$

where $\sigma(\cdot, \cdot)$ denotes covariance. Note that we do not introduce the assumption that the true-score components are uncorrelated. As the situational component may change across sessions, even if the error variance does not change across sessions, the reliability of the test scores is not supposed to be invariant because the reliable variance is not supposed to be invariant across sessions.

A third consequence of the true-score decomposition is that we obtain a clear definition of gain scores: The test-retest true-score difference $\delta_i \equiv \tau_{i2} - \tau_{i1}$ measures only the temporal variation of situational effects:

$$\delta_i = s_{i2} - s_{i1}. \tag{4}$$

Consequently, the mean retest improvement is the mean change of situational effects measured at both times of measurement (for a proof, see Appendix A):

$$\mu(Y_{i2}) - \mu(Y_{i1}) = \mu(s_{i2} - s_{i1}) \tag{5}$$

From the substantive point of view, a mean retest improvement means that the situational effects on Session 2 improved upon test performance measured at Session 1. We defined the s_{it} component as a between-subjects variable. If test-retest effects are associated with little between-subjects variability, the variance of the situational component will be reduced, which entails that the reliability of the difference $Y_{i2} - Y_{i1}$ will be small. At the limit, no between-subjects variability would mean that the retest effect is a constant–in which case the reliability of the gain test scores would be undefined (Raykov, 2001).

Operational hypothesis

Now we consider a battery of m tests. For technical reasons detailed in the Method section, we assume that the test-specificity of the situational effects, i.e., the fact that the situational effects measured by two tests do not correlate perfectly, can be neglected in the

statistical modeling. It is likely that such assumption would not hold in settings where each test was taken in very different situations. This assumption can be formulated as the linear relationship:

$$s_{jt} = \nu_j + \lambda_j \cdot s_{1t}, \tag{6}$$

where $j \in \{2, \ldots, m\}$. The coefficients ν_j and λ_j serve to model scaling differences due to the ways test j and test 1 capture situational effects.[3] It is noteworthy that the s_{it} of the m tests are linearly linked, but as they are allowed to vary across situations (by definition) they cannot be interpreted as part of g.

Coming back to the assumption that the construct relevant components of the true score are temporally stable it follows that true changes δ_1, δ_2, ..., δ_m are proportional (see Appendix A for a proof):

$$\delta_j = \lambda_j \cdot \delta_1 \tag{7}$$

Such a simple linear structure of true change has testable implications on the moment structure of the data. If a non-null retest improvement is observed, say $\mu(Y_{12}) - \mu(Y_{11}) \neq 0$, the coefficient λ_j is determined by the means of the test variables Y_{1t} and Y_{jt} (see Appendix A for detailed demonstration):

$$\lambda_j = \frac{\mu(Y_{j2}) - \mu(Y_{j1})}{\mu(Y_{12}) - \mu(Y_{11})}. \tag{8}$$

Therefore the experimental assumption of common situational effects on a set of cognitive ability tests enables the testing of the assumption of temporally stable ability components by fixing the loadings of a latent change factor in a longitudinal structural equation model, as detailed below.

Method

Rationale

As a complement to the classical approach, which assumes that gain scores reflect changes of test-specific abilities, we hypothesized that gain scores can be attributed to situational factors. In other words, there would be a construct-irrelevant component, common to all test scores in the same session, varying from one session to the other, and responsible for score increases. Contrary to g, which is assumed to be temporally constant, the common component would vary from one session to the other and would be responsible for score increases in a test-retest design. Two forms of this hypothesis can be developed, depending on whether construct-relevant components are assumed temporally stable or not. In the strong version, situational effects would be sufficient to explain all of the gain scores. A weaker version would explain retest effects through a combination of situational effects and some other factors (i.e., test-specific abilities). The present paper focuses on the strong version only.

If the strong version holds, provided that the situational effects measured by different tests can be approximated as perfectly correlated, the score increases on a battery of tests can be modeled through a unique latent factor, representing the latent change between both situations. Thus a way to test the strong version is to test the plausibility of a unifactorial latent change model. Figure 1 depicts the latent decomposition corresponding to the strong version. Figure 1 is only given for pedagogical purposes to illustrate the general hypothesis. Such a path diagram does not allow statistical testing because the corresponding model would be unidentified.

The assumptions of common situational effects along with temporally stable effects of the g and a_i components on scores from a battery of cognitive ability tests can be tested in the context of a high-stake selection setting, namely, admission into French Aircraft

Pilot Training (on average each year only 8% of the applicants are selected). As all tests in a single session are administered during a short period (one day), and as retest cannot occur before in under a year, each testing session can be defined as a situational unit.

Participants

The sample comprised 752 applicants to the École Nationale de l'Aviation Civile (French Air Transport Pilot Training) who took the battery of ability tests twice. Test administrations were one year apart for 90% of the sample. The sample was composed of 89% men, 97% aged between 19 and 23, and 97% coming from preparatory years. 95% of the participants had no aeronautic experience (ab initio), and 5% were experienced pilots.

Materials

This study focused on the three paper-and-pencil tests taken by all applicants between 1998 and 2005. Applicants were exposed to the exact same tests across both testing sessions. Tests are described below using Carroll's terminology.

1. *Visual Perception test (V).* This test is composed of three time-limited sub-tests (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 varied from 0 to 85 (number of correct answers).

2. *Mechanical movement test (M).* This test presents 36 situations to evaluate, from a mechanical point of view, with a choice of 4 possible answers for each situation. Test is time limited (25 min) and scores range from 0 to 42 (number of correct answers). As Carroll (1993) mentioned, this kind of test is loaded as well on Mechanical Knowledge and Visualization factors (p. 324).

3. *Attentional ability test (A).* In this test applicants have to detect three target signs among eight in a page containing 1560 signs. Time is limited (10 min) and scores

range from −60 to 60 (number of correct answers minus number of omissions divided by 10). This test evaluates selective attention and concentration during a monotonous task (the "ability to attend" in Carroll's terminology). However, it also includes a perceptual speed component because it also measures the speed of locating given symbols in an extended visual field.

Analyzes

As a whole, the model depicted in Figure 1 is not testable due to the high number of parameters to be estimated. Therefore we will test an identified model that is implied by the hypotheses. To test the hypothesis of temporal stability of the g and a_i components jointly with common situational, transient effects, we specified a *latent change* structural equation model where true change was accounted for by a single change factor with fixed loadings. The change factor represents situational test-retest effects. The model is depicted in Figure 2, lower panel. To make the model algebraically identified, the loading λ_{V} was fixed at unity. The mean structure was completely specified by (i) fixing the means of the latent variables, as the values of parameters $\mu(\tau_{\mathrm{V1}})$, $\mu(\tau_{\mathrm{M1}})$, $\mu(\tau_{\mathrm{A1}})$, and $\mu(\delta_{\mathrm{V}})$ are those of estimates $m(Y_{\mathrm{V1}})$, $m(Y_{\mathrm{M1}})$, $m(Y_{\mathrm{A1}})$, and $m(Y_{\mathrm{V2}}) - m(Y_{\mathrm{V1}})$, respectively, where $m(\cdot)$ denotes the sample mean, and (ii) by fixing the loadings on the change factor as detailed in Equation 8.[4] The manifest variables had null intercepts. As the model has 11 degrees of freedom, we obtained sufficient statistical power to test "close fit" (i.e., null RMSEA = 0.05 and population RMSEA = 0.08), as $\pi = .81$ with $N = 752$ and $\alpha = .05$ (MacCallum, Browne, & Sugawara, 1996). As with the multitrait model, the latent change model does not constrain the error variances to be equal over time, and assumes uncorrelated errors.

A more restricted latent change model was specified, within which the error variances were constrained to be equal over time, yielding 14 degrees of freedom. Comparing the fits of both models allows testing for temporal invariance of the error

variances. The models were tested with the Mplus software (Muthén & Muthén, 2004), by using the robust maximum likelihood estimator (see Appendices B, C and D for the Mplus syntaxes of the three models).

Results

Mean Retest Improvements

The observed moment structure of the final sample of data (see below) is displayed in Table 1. The data exhibited large mean effect sizes, $\hat{d}_{\mathrm{V}} = 0.85$, $\hat{d}_{\mathrm{M}} = 0.64$ and $\hat{d}_{\mathrm{A}} = 1.02$ for the visual, mechanical and attentional tests, respectively. The effect sizes reported here are calculated with the d formula (Cohen, 1988).

Neglecting the Regression-to-the-Mean Effect

Because of the unperfect reliability of the tests, the mean gain score is overestimated, due to the regression-to-the-mean effect (Nesselroade, Stigler, & Baltes, 1980). Thus, the practical issue is to evaluate whether this effect may be neglected by a SEM account of the gain scores. According to Bobko (2001)'s computation formula of the expected gain score due to the regression-to-the-mean effect, gains of .04, .02, and .10 standard deviation are expected for tests V, M, and A, respectively. These values are rather small compared to the observed effect sizes in Table 1. Thus regression-to-the-mean alone cannot account for the data, and we will neglect it in further analyses.

Common Situational Effects

The latent change model with 11 degrees of freedom, i.e., without the constraint of constant error variance across time, fitted the data very closely, $\chi^2(N = 752,\ df = 11) = 7.00$, $p = .80$, RMSEA $= 0.000$. The more restricted model, i.e., with the constraint of constant error variance across time, exhibited a worse fit, $\chi^2(N = 752,\ df = 14) = 33.06$, $p = .003$, RMSEA $= 0.043$, $\chi^2_{\mathrm{diff}}(N = 752,\ df = 3) = 13.17$,

$p = .004$. The modification indices suggested that scores at test P could exhibit unstable error variances across sessions.

Therefore, to improve linearity, we replaced the variables $Y_{\mathrm{A}1}$ and $Y_{\mathrm{A}2}$ with a log-transformation (the best transformation to improve normality indices) after having removed 16 cases with extreme low scores ($N = 736$). The model with 11 degrees of freedom fitted the data very closely, $\chi^2(N = 736,\ df = 11) = 7.20$, $p = .78$, RMSEA = 0.000. The fit of the restricted model was not significantly decreased, $\chi^2(N = 736,\ df = 14) = 14.54$, $p = .41$, RMSEA = 0.007, $\chi^2_{\mathrm{diff}}(N = 736,\ df = 3) = 6.73$, $p = .08$, and the more parsimonious model was retained. Thus, assuming that the log-transformation of the scores from the test P is suitable, invariance of error variances was retained. The input covariance matrix, after log-transformation of test P scores, is given in Table 1.

In short, the goodness-of-fit summaries support the conjoint hypotheses of common situational effects as stated in Equation 6 and temporally stable ability components as stated in Equation 2.

Further Analyzes

Estimates of latent variances, covariances, and correlations are given in Table 2. Results show correlations less than .50 between the true-score variables. The estimated variance of the change factor δ_{V} was very small with respect to the variance of $\tau_{\mathrm{V}1}$, $5.52/61.54 \simeq 0.09$. Similarly, the estimated variance of the latent gain scores at test M was very small with respect to the variance of $\tau_{\mathrm{M}1}$, $5.52 \times 0.303^2/10.75 \simeq 0.05$. Test A seemed more sensitive to gain scores, $5.52 \times 0.034^2/0.03 \simeq 0.21$.

The estimated residual variances of the manifest variables $Y_{\mathrm{V}t}$, $Y_{\mathrm{M}t}$, and $Y_{\mathrm{A}t}$, $t = 1,\ 2$, were 26.75 ($SE = 1.84$), 4.390 ($SE = 0.34$), and 0.02 ($SE = 0.00$), respectively. Table 3 displays the reliability estimates of the variables Y_{it}, and the reliability estimates

of the gain-score variables $Y_{i2} - Y_{i1}$. As expected, reliability estimates were not time-invariant although they ranged in the usual levels. Unsurprisingly, the gain-score reliability values were very small.

Discussion

In this paper, we investigated whether the error-free component of gain scores could be accounted for by temporal variations in situational effects, that is, reliable although construct-irrelevant effects. Situational effects refer to the person's state in a given testing session. Thus, change in situational effects could increase the observed scores significantly. We showed that such a view is testable as far as situational effects can be thought of as common to a battery of cognitive ability tests. Specifically, the gain scores can be formulated as the transient effects of a one latent change variable, in a highly constrained structural equation model. Using data from the operational selection setting for admission into the French Aircraft Pilot Training, we provided evidence for the plausibility of such a view.

Interpreting the Plausibility of Perfectly Correlated Situational Effects

As a rule, situational effects are defined at the level of a test. Thus, different tests measure different situational effects. However, in the present paper, we were successful in assuming perfectly correlated situational effects at both times. There are two possible interpretations of this result:

1. Situational effects were perfectly correlated.

2. Situational effects–defined as situational transient variations–were not perfectly correlated and the sizes of the involved variances, that is $\sigma^2(s_{it})$ and $\sigma^2(\delta_i)$, were small enough–with respect to the sizes of $\sigma^2(g)$ and $\sigma^2(a_i)$–to make that mispecification undetectable by the likelihood ratio test.

From a substantive point of view, in settings such as one-day-session testing, it is likely that situational effects were not very test-specific because the testing situation acts as a situational unit. At Session 1, test anxiety, other emotional factors and the way each applicant interacts with each test would work as a halo effect. At Session 2, such a halo could be enriched by a global effect of practice. For example, it is known that for entry to the French Air Transport Pilot Training, applicants are highly motivated, and organized with Internet forums where they exchange training programs and information about the tests. As a consequence, they are globally well-trained on all tests. Thus, it is likely that practice acts as a homogeneous situational factor.

The Situational Effect and the g*-Loading Effect*

In this paper, we tested the strong version of the situational effects hypothesis, that is, situational effects can explain all gain scores. Previous research (e.g., te Nijenhuis et al., 2007 for a meta-analysis) has shown that the less a test is g-loaded, the greater the score increase is after cognitive intervention. However, our approach to gain scores is not opposed to the g-loading effect which could be used to explain size effect differences for different cognitive ability tests. Our purpose was to propose a modeling that accounts for individual effects on different tests, while assuming stability of g. In other words, we investigated another interpretation of the gain scores on cognitive ability tests compatible with no change on g.

In our empirical study, the strongest retest effect was observed for test A, which is probably less g-loaded than test V for example, in accordance with the g-loading effect (estimated g-loadings of each test of this study were not available). Nevertheless, the difference of size effects for the different tests is included in the λ_i parameters of our model. Indeed this parameter corresponds to the size effect of one test compared to the size effect for the reference test (test V in our study). Thus the g-loading effect is included

in the modeling.

However, it remains an open question how to statistically disentangle the variance due to g and the variance due to common situational effects, as both sources of variance are defined as common variance. From this perspective, it can be suggested that common factors used in previous research that do not account for situational effects do not represent g with perfect validity, especially when participants are likely to have trained themselves.

Implications on Predictive Validity

Our true-score decomposition highlights in what sense test scores can exhibit different predictive validity for a criterion. The covariance between the situational component and the criterion is not supposed to be invariant across testing sessions, as the situational effects are transient effects. As the predictive validity of the true-score variable is a function of its covariance with the criterion, and because the covariance between the situational component and the criterion enters the covariance between the true-score variable and the criterion, the latter is not supposed to be invariant across testing sessions as well. Because test scores measure situational effects, their construct and predictive validity may change across testing sessions while their validity with respect to the target ability components remains unchanged.

Implication on the Measurement Invariance Question

As noted in the Introduction, Reeve and Lam (2005) and Lievens et al. (2007) investigated measurement invariance of cognitive ability test batteries and found evidence supporting both measurement invariance (Reeve & Lam, 2005) and non-invariance (Lievens et al., 2007), demonstrating the complexity of retest effects. Our findings are clearly in favor of metric non-invariance, as the true-score composition is supposed to change from one session to another.

Interpreting Reported Decrease of g*-Loadings after Practice*

Practice could be associated with a reduction of g-loadedness (te Nijenhuis, Voskuijl, & Schijve, 2001), although the authors warned against too strong conclusions in the presence of small effect sizes. In usual factor analytic models, the g-loadings are loadings of a common factor. Recognizing that this common factor may capture a combination of g and common situational effects, our framework could explain that the loadings on this factor decrease at the retest, provided that the variance of the situational effects at the retest decrease, as shown in Appendix E. In such a case, assuming scalar invariance and temporal invariance of the error variances, (i) the variance of the observed score decreases at Session 2 and (ii) the g-validity of test scores would increase after practice. Further research would be needed to properly test the null hypothesis of invariance of the g-loadings in longitudinal test-retest models.

Limits

In the present study we were led to assume that the retest effects were due to training between both sessions. But there was no available information concerning the type and number of training hours for each applicant. Compared to the meta-analytic mean effect size observed in practice effects studies (+0.26 SD, Hausknecht et al., 2007), our empirical results suggest many training hours. Moreover, our finding of small variance of the change factor suggests a homogeneous training across individuals but it would have been interesting to have empirical data on it.

For the purpose of the paper it was sufficient to test the situational effect hypothesis on three tests. However it would have been even more convincing with a large battery of tests covering a variety of cognitive abilities. To assess the degree of generalization of our findings, one could test our hypothesis with a greater number of tests, including more highly g-loaded tests like Raven's matrices and tests measuring very disparate abilities

(e.g., verbal ability tests, quantitative ability tests, psychomotor tests, etc.).

In our study, we argue in favor of construct-irrelevant changes after practice because these changes can be modeled by a single factor, common to a battery of tests. The idea is that if the factor is common to different tests and if it changes from one session to another, then it is not likely to represent construct relevant change on each test. Nevertheless, an alternative would be homogeneous construct-relevant changes. Coming back to the "shadow-illustration", such a situational effect could not be disentangled with a homogeneous and real growth of the person between two measurements. From this point of view, the only way to investigate this question is to assess *far transfer* or *broad generalizability* (Jensen, 1998, p. 333).

Conclusion

In short, we proposed a new interpretation of gain scores observed between two administrations of cognitive ability tests, especially in a selection setting where applicants are likely to train themselves. Our empirical results are compatible with a situational effect hypothesis which assumes the stability of the construct-relevant component of the score. Thus our results are compatible with a non-g increase of scores. One practical implication is that first-time takers' and retakers' scores should not be treated identically. Indeed, all scores embed a construct-irrelevant component, but that component looms larger in retakers' scores than in first takers'.

References

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

Bobko, P. (2001). *Correlation and regression: Applications for industrial/organizational psychology and management*. Thousand Oaks, CA: Sage.

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence*, *30*, 505–514.

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

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.

Dixon, R., Kramer, D., & Baltes, P. (1985). Intelligence: A life-span developmental perspective. In B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements and applications* (pp. 301–350). New-York: Wiley.

Domino, G., & Domino, M. L. (2006). *Psychological testing*. Cambridge: Cambridge University Press.

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.

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

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.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure analysis. *Psychological Methods*, *1*, 130–149.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus 3.11.* Computer Program. Los Angeles: Authors.

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression toward the mean and the study of change. *Psychological Bulletin*, *88*, 622–637.

Raykov, T. (2001). On the use and utility of the reliability coefficient in social and behavioral sciences. *Quality and Quantity*, *35*, 253–263.

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.

Roberts, M. J., & Newton, E. J. (2003). Individual differences in the development of reasoning strategies. In D. Hardman & L. Macci (Eds.), *Thinking: Psychological perspectives on reasoning, judgment, and decision making.* Chichester: John 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.* 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.* Norwood, NJ: Ablex.

Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability, and testability. *Methodika*, *3*, 25–60.

Steyer, R., Ferring, D., & Schmitt, M. J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment*, *8*, 79–98.

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality*, *13*, 389–408.

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.

Zimmerman, D. W. (1975). Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika*, *40*, 395–412.

Zumbo, B. D., & Rupp, A. A. (2004). Responsible modeling of measurement data for appropriate inferences: Important advances in reliability and validity theory. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 73–92). Thousand Oaks, CA: Sage.

Appendix A

Proofs for Equations 5, 7, and 8

Equation (5) :

$$\begin{aligned}\mu(Y_{i2}) - \mu(Y_{i1}) &= \mu(\tau_{i2}) - \mu(\tau_{i1}) \\ &= \mu(\tau_{i2} - \tau_{i1}) \\ &= \mu(s_{i2} - s_{i1}).\end{aligned}$$

Equation (7) :

$$\begin{aligned}\delta_j &= s_{j2} - s_{j1} \\ &= \nu_j + \lambda_j \cdot s_{12} - (\nu_j + \lambda_j \cdot s_{11}) \\ &= \lambda_j \cdot (s_{12} - s_{11}) \\ &= \lambda_j \cdot \delta_1.\end{aligned}$$

Equation (8) :

$$\begin{aligned}\mu(Y_{j2}) - \mu(Y_{j1}) &= \mu(\delta_j) \\ &= \lambda_j \cdot \mu(\delta_1) \\ &= \lambda_j \cdot [\mu(Y_{12}) - \mu(Y_{11})],\end{aligned}$$

$$\Longleftrightarrow \lambda_j = \frac{\mu(Y_{j2}) - \mu(Y_{j1})}{\mu(Y_{12}) - \mu(Y_{11})}.$$

Appendix B

Mplus syntax of the multitrait model

```
VARIABLE:   NAMES ARE V1-V6;
!V for observed variables,
!V1=test V at Session~1, V2=test V at Session~2,
!V3=test M at Session~1, V4=test M at Session~2,
!V5=test A at Session~1, V6=test A at Session~2.

          USEOBS ARE V5>=100 AND V6>=100;
          USEVARIABLES V1-V4 V7 V8;
DEFINE:     V7 = LOG(V5); V8 = LOG(V6);
ANALYSIS: TYPE= MEAN;
          ESTIMATOR = MLR;

MODEL:
!T for true scores
TV BY V1 V2@1;
TM BY V3 V4@1;
TA BY V7 V8@1;

[TV TM TA];
[V1@0];
[V3@0];
```

```
[V7@0];

[V2 V4 V8];

OUTPUT: STANDARDIZED;RESIDUAL;TECH4;SAMPSTAT;MODINDICES;
```

Appendix C

Mplus syntax of the latent change model with fixed loadings

```
VARIABLE:   NAMES ARE V1-V6;
!V for observed variables,
!V1=test V at Session~1, V2=test V at Session~2,
!V3=test M at Session~1, V4=test M at Session~2,
!V5=test A at Session~1, V6=test A at Session~2.

          USEOBS ARE V5>=100 AND V6>=100;
          USEVARIABLES V1-V4 V7 V8;
DEFINE:     V7 = LOG(V5/10); V8 = LOG(V6/10);
ANALYSIS:        TYPE= MEAN;
            ESTIMATOR = MLR;
MODEL:
!T for true scores for tests V, M, and A, at Session~1
!DELTA for true change on test V.

TV1 BY V1 V2@1;
TM1 BY V3 V4@1;
TA1 BY V7 V8@1;

DELTA BY V2
         V4@.303
```

```
         V8@.034;

[TV1@49.751];

[TM1@25.947];

[TA1@3.216];

[V1-V8@0];

[DELTA@6.442];

OUTPUT: STANDARDIZED;RESIDUAL;TECH4;SAMPSTAT;MODINDICES;
```

Appendix D

Mplus syntax of the restricted latent change model with fixed loadings

```
VARIABLE:   NAMES ARE V1-V6;
!V for observed variables,
!V1=test V at Session~1, V2=test V at Session~2,
!V3=test M at Session~1, V4=test M at Session~2,
!V5=test A at Session~1, V6=test A at Session~2.

          USEOBS ARE V5>=100 AND V6>=100;
          USEVARIABLES V1-V4 V7 V8;
DEFINE:     V7 = LOG(V5/10); V8 = LOG(V6/10);
ANALYSIS:        TYPE= MEAN;
            ESTIMATOR = MLR;
MODEL:
!T for true scores for tests V, M, and A, at Session~1
!DELTA for true change on test V.

TS1 BY V1 V2@1;
TM1 BY V3 V4@1;
TP1 BY V7 V8@1;

DELTA BY V2
         V4@.303
```

```
        V8@.034;

!constraint on equality of variances
V1 V2 (a);
V3 V4 (b);
V7 V8 (c);

[TV1@49.751];
[TM1@25.947];
[TA1@3.216];
[V1-V8@0];
[DELTA@6.442];
OUTPUT: STANDARDIZED;RESIDUAL;TECH4;SAMPSTAT;MODINDICES;
```

Appendix E

Revisiting the *g*-Loadings Decrease History

Let the scores Y_{it} at test i from a battery of tests taken at time t be a linear function of a centered common factor f_t and a residual variable k_{it}:

$$Y_{it} = b_i \cdot f_t + k_{it},$$

where b_i is a scaling constant which is measurement invariant across time for meaningfulness in the longitudinal perspective. The so-called g-loading β_{it} in a usual factor analytic model expresses as

$$\beta_{it} = b_i \cdot \sigma(f_t),$$

where $\sigma(f_t)$ denotes the standard deviation of f_t.

Assuming:

1. $f_t = g + s_t$, where s_t denotes common situational effects,

2. $\sigma(g,\ s_t) = 0$ for simplicity,

it follows that

$$\begin{aligned}
\sigma^2(s_2) < \sigma^2(s_1) &\implies \sigma^2(g) + \sigma^2(s_2) < \sigma^2(g) + \sigma^2(s_1) \\
&\implies \sigma^2(g + s_2) < \sigma^2(g + s_1) \\
&\implies \sigma^2(f_2) < \sigma^2(f_1) \\
&\implies b_i^2 \cdot \sigma^2(f_2) < b_i^2 \cdot \sigma^2(f_1) \\
&\implies \beta_{i2}^2 < \beta_{i1}^2 \\
&\implies \beta_{i2} < \beta_{i1}.
\end{aligned}$$

Assuming that

$$\sigma^2(k_{i1}) = \sigma^2(k_{i2}),$$

it follows that

$$\sigma^2(Y_{i1}) > \sigma^2(Y_{i2}).$$

Thus, defining the g-validity of the test-score variable Y_{it} as

$$\rho_g(Y_{it}) = \frac{b_i^2 \cdot \sigma^2(g)}{\sigma^2(Y_{it})},$$

it follows that

$$\rho_g(Y_{i2}) > \rho_g(Y_{i1}).$$

Author Note

Nadine Matton, CLLE-LTC (CNRS UMR5263, Université de Toulouse, EPHE), Toulouse, France. Stéphane Vautier, OCTOGONE-CERPP, Université de Toulouse, France. Éric Raufaste, CLLE-LTC (CNRS UMR5263, Université de Toulouse, EPHE), Toulouse, France.

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

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

Notes

[1]In selection settings, regression to the mean can occur due the sampling bias at Session 1, and the issue arises whether this effect is negligible or not, as will be discussed in our empirical study.

[2]Change associated with a given test is not supposed to be exactly equal to change associated with another test, but linearly related, as detailed in the Operational Hypothesis section.

[3]We assume that scaling differences of situational effects between measurement tools depend only on the interaction between the measurement tools and the population of testees, whatever the time they are used; thus, ν_j and λ_j are specified as scaling factors without reference to the number of the testing session.

[4] $\lambda_{\mathrm{M}} = (27.899 - 25.947)/(56.193 - 49.751) = 0.303$ and
$\lambda_{\mathrm{A}} = (3.433 - 3.216)/(56.193 - 49.751) = 0.034$.

Table 1

Input Variances (Diagonal), Covariances (Below Diagonal), Correlations (Above Diagonal and in Italics), Means of the Observed Variables

	Y_{V1}	Y_{V2}	Y_{M1}	Y_{M2}	Y_{A1}	Y_{A2}
Y_{V1}	83.21	*.66*	*.31*	*.26*	*.34*	*.19*
Y_{V2}	55.31	85.10	*.25*	*.21*	*.31*	*.26*
Y_{M1}	11.25	9.23	15.51	*.67*	*.10*	*−.08*
Y_{M2}	8.55	6.80	9.46	12.71	*.06*	*−.07*
Y_{A1}	0.68	0.62	0.09	0.04	0.05	*.55*
Y_{A2}	0.39	0.55	−0.09	−0.06	0.03	0.05
Means	49.75	56.19	25.95	27.90	3.22	3.43

Table 2

Estimates of the Latent Variances, Covariances (Below Diagonal), and Correlations (Above Diagonal and in Italics)

	τ_{V1}	τ_{M1}	τ_{A1}	δ_V
τ_{V1}	61.54 (4.34)	*.45*	*.48*	*−.36*
τ_{M1}	11.66 (1.32)	10.75 (0.70)	*.11*	*−.55*
τ_{A1}	0.64 (0.08)	0.06 (0.03)	0.03 (0.00)	*−.07*
δ_V	−6.57 (1.93)	−4.26 (0.71)	−0.03 (0.04)	5.52 (1.54)

Note. Robust standard errors in parentheses. The loading values can be computed using Equation (5).

Table 3

Reliability Estimates of the Test, Retest and Gain Score Variables

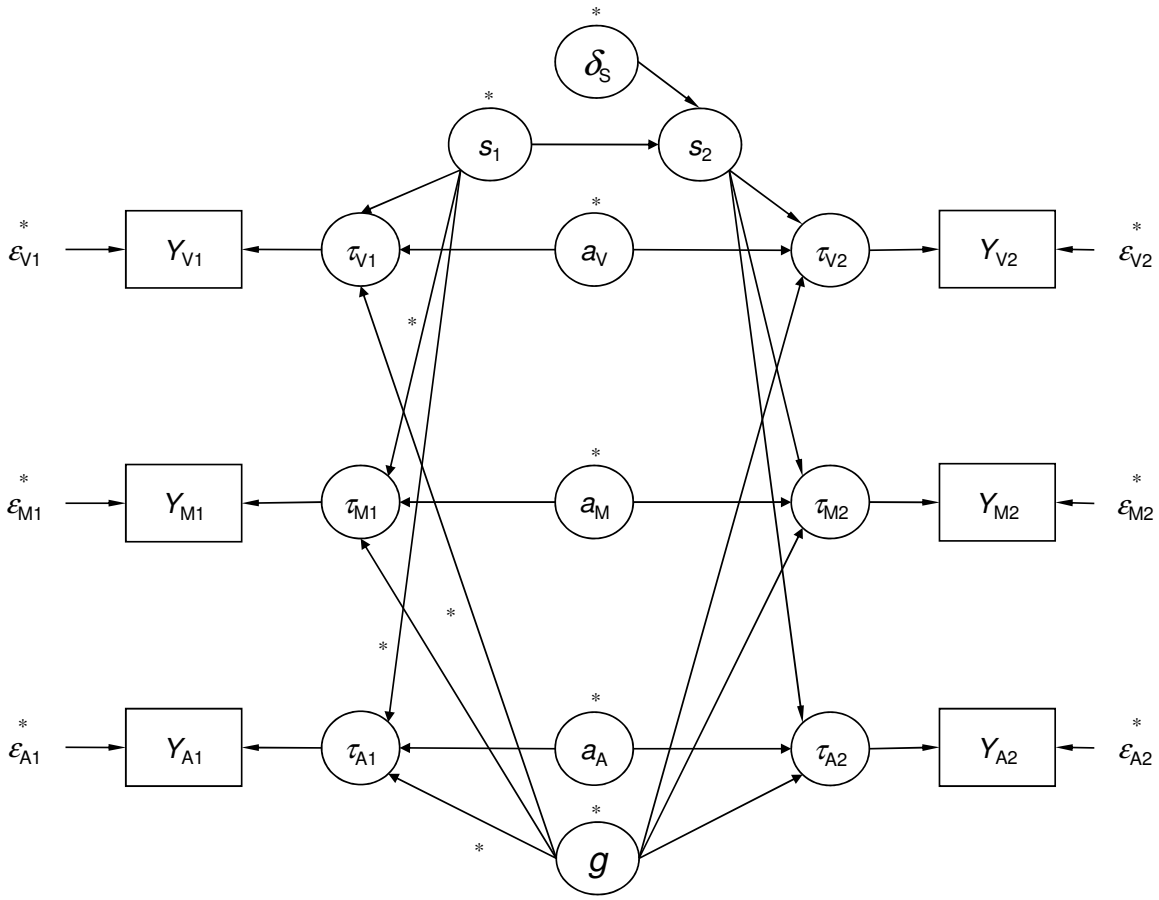
	Test	Retest	Gain
V	.70	.67	.09
M	.71	.66	.05
A	.61	.64	.14

Figure Captions

Figure 1. The latent decomposition corresponding to the hypothesis of no temporal stability of g and test-specific ability components, and common, test-unspecific, transient situational effects. Y and τ represent observed and true-score variables, respectively, associated with tests V, M, and A used at times 1 and 2. a_i represents ability specific to test i, g the general factor, s_t the situational effects at time t and δ_{s} represents the difference $s_2 - s_1$. The latent covariance structure cannot be identified because there are 21 manifest variances and covariances against 31 parameters to be estimated–the 4 estimated loadings and 12 estimated variances are noted with a star (*); the 15 estimated covariances are not represented due to a question of readability.

Figure 2. Path diagram of the latent change SEM. Y and τ represent observed and true-score variables, respectively, associated with tests V, M, and A used at times 1 and 2. δ_{V} represents the difference $\tau_{\mathrm{V}2} - \tau_{\mathrm{V}1}$. Parameters λ_M and λ_A are fixed–see Equation 8.

Situational Effects in Ability Testing, Figure 1



Situational Effects in Ability Testing, Figure 2