Intelligence 50 (2015) 93–99

Contents lists available at ScienceDirect

Intelligence

A general factor of intelligence fails to account for changes in tests' scores after cognitive practice: A longitudinal multi-group latent-variable study

Eduardo Estrada [a], Emilio Ferrer [b], Francisco J. Abad [a], Francisco J. Román [a], Roberto Colom [a,*]

[a] Facultad de Psicología, Universidad Autónoma de Madrid, Spain
[b] University of California at Davis, USA

ARTICLE INFO

Article history:
Received 23 December 2014
Received in revised form 4 February 2015
Accepted 20 February 2015
Available online xxxx

Keywords:
General cognitive ability
Practice effect
Working memory span
Processing speed

ABSTRACT

As a general rule, the repeated administration of tests measuring a given cognitive ability in the same participants reveals increased scores. This brings to life the well-known practice effect and it must be taken into account in research aimed at the proper assessment of changes after the completion of cognitive training programs. Here we focus in one specific research question: Are changes in test scores accounted for by the tapped underlying cognitive construct/factor? The evaluation of the factor of interest by several measures is required for that purpose. 477 university students completed twice a battery of four heterogeneous standardized intelligence tests within a time lapse of four weeks. Between the pre-test and the post-test sessions, some participants completed eighteen practice sessions based on memory span tasks, other participants completed eighteen practice sessions based on processing speed tasks, and a third group of participants did nothing between testing sessions. The three groups showed remarkable changes in test scores from the pre-test to the post-test intelligence session. However, results from multi-group longitudinal latent variable analyses revealed that the identified latent factor tapped by the specific intelligence measures fails to account for the observed changes.

1. Introduction

Practice effects are broadly acknowledged in the cognitive abilities literature (Anastasi, 1934; Colom et al., 2010; Hunt, 2011; Jensen, 1980; Reeve & Lam, 2005). When the same individuals complete the same (or parallel) standardized tests, their scores show remarkable improvements. However, as discussed by Jensen (1998) among others (Colom, Abad, García, & Juan-Espinosa, 2002; Colom, Jung, & Haier, 2006; te Nijenhuis, van Vianen, & van der Flier, 2007), specific measures tap cognitive abilities at three levels: general ability (such as the general factor of intelligence, or *g*), group abilities (such as verbal or spatial ability), and concrete skills required by the measure (such as vocabulary or mental rotation of 2D objects).

Within this general framework, recent research aimed at testing changes after the completion of cognitive training programs has produced heated discussions regarding the nature of the changes observed in the measures administered before and after the training regime (Buschkuehl & Jaeggi, 2010; Conway & Getz, 2010; Haier, 2014; Moody, 2009; Shipstead, Redick, & Engle, 2010, 2012; Tidwell, Dougherty, Chrabaszcz, Thomas, & Mendoza, 2013). The changes may or may not be accounted for by the underlying construct of interest. Thus, for instance, the pioneering work by Jaeggi, Buschkuehl, Jonides, and Perrig (2008) observed changes in fluid intelligence measures after completion of a challenging cognitive training program based on the dual n-back task. This report stimulated a number of investigations aimed at replicating the finding (Buschkuehl, Hernandez-Garcia, Jaeggi, Bernard, & Jonides, 2014; Colom et al.,

* Corresponding author at: Facultad de Psicología, Universidad Autónoma de Madrid, 28049 Madrid, Spain. Tel.: +34 91 497 41 14 (Voice).
E-mail address: roberto.colom@uam.es (R. Colom).

http://dx.doi.org/10.1016/j.intell.2015.02.004
0160-2896/

2013; Chooi & Thompson, 2012; Harrison et al., 2013; Jaušovec & Jaušovec, 2012; Redick et al., 2012; Rudebeck, Bor, Ormond, O'Reilly, & Lee, 2012; Shipstead et al., 2012; Stephenson & Halpern, 2013; von Bastian & Oberauer, 2013).

The meta-analysis published by Melby-Lervåg and Hulme (2012) concluded that short-term cognitive training fails to improve performance on far-transfer measures. Nevertheless, their results support the conclusion that the completed programs might improve performance on near-transfer measures, meaning that specific cognitive skills seem sensitive to training. The meta-analysis by Au et al. (2014), focused on reports analyzing the effect of cognitive training programs based on the n-back task, showed a positive, albeit small, impact on fluid intelligence measures. The weighted average effect size was .24, which is equivalent to 3.6 IQ points. The authors suggested that even these small increments might impact performance on real-life settings (see also Herrnstein & Murray, 1994 for a similar argument).

Importantly, it has been proposed that the latter statement may be relevant if and only if observed increments in test scores are accounted for by the tapped latent factor representing the construct of interest. In this regard, te Nijenhuis et al. (2007) reported a meta-analysis of sixty-four studies using a test-retest design, finding a perfect negative correlation between the vectors defined by tests' scores changes and the *g* loadings of these tests. Their main conclusion was that observed improvements on scores were test-specific and unrelated to the general factor of intelligence (*g*). However, this study was based on the method of correlated vectors, which has been questioned on several grounds (Ashton & Lee, 2005). Latent-variable analyses are more robust and, therefore, may provide better answers to the question of interest (Dolan & Hamaker, 2001; Haier, 2014).

These latent-variable analyses require appropriate sample sizes. Published research reports analyzing changes after completion of cognitive training programs consider small samples on a regular basis. Thus, for instance, the meta-analysis by Au et al. (2014) is based on reports with sample sizes ranging from 3 to 30 participants (see their Table 1) and this precludes the application of the recommended latent-variable analyses.

To help fill this gap, here we report a study considering a large number of participants (477). All these participants completed screen versions of four heterogeneous standardized intelligence tests on two occasions, separated by four weeks. Participants were randomly assigned to three groups comprising more than one hundred participants each. Between the pre-test and post-test sessions, the first group completed 18 practice sessions based on memory span tasks, the second group completed 18 practice sessions based on processing speed tasks, whereas the third group did nothing. These three groups were systematically compared using multi-group longitudinal latent-variable analyses in order to examine the main research question, namely, are changes in tests' scores from a pre-test to a post-test intelligence session accounted for by the tapped latent trait?

2. Method

2.1. Participants

477 psychology undergraduates took part in the study (82% were females). The mean age was 20.13 (SD = 3.74). They participated to fulfill a course requirement. Participants were randomly assigned to three groups. The first group (memory span) comprised 170 students, the second group (processing speed) comprised 114 students, and the third group (passive control) comprised 193 students.[1]

2.2. Measures

Intelligence was measured by four tests: the Raven Advanced Progressive Matrices Test (RAPM), as well as the abstract reasoning (DAT-AR), verbal reasoning (DAT-VR), and spatial relations (DAT-SR) subtests from the Differential Aptitude Test Battery (Bennett, Seashore, & Wesman, 1990). Memory span tasks completed by participants of the first group included short-term memory (STM) and working memory (WMC). Participants on the second group completed verbal, quantitative, and spatial processing speed (PS) tasks along with verbal, numerical, and spatial attention (ATT) tasks between the pre-test and post-test intelligence session. Appendix 1 includes a detailed description of these measures.

2.3. Procedure

In the first week, participants were tested on the intelligence tests using the odd numbered items only. In the second, third, and fourth weeks the first group (memory span) completed the three verbal, numerical, and spatial short-term memory and the three verbal, numerical, and spatial working memory tasks, whereas the second group completed the three verbal, numerical, and spatial processing speed and the three verbal, numerical, and spatial attention tasks. Participants completed six different tasks on each session, resulting in a total of 18 sub-sessions. Order of administration was counterbalanced across weeks. Finally, on the fifth week participants were tested on the intelligence tests using the even numbered items only. Participants completed the five sessions exactly the same day at the same time every week.

2.4. Analyses

Factorial invariance has been proposed as an analytic strategy for studying longitudinal changes. This technique can be used for assessing whether the same latent construct is measured by a set of indicators at different time points, and whether the relations between the latent variable and the indicators remain invariant across occasions (Meredith, 1993; Meredith & Horn, 2001; Widaman, Ferrer, & Conger, 2010). It is particularly applicable when a set of indicators is available for measuring the same latent construct, and the sample size is adequate. Our dataset meets both requirements.

Studying intelligence as a latent variable allows overcoming limitations related to tests' scores (Haier, 2014), thus helping to achieve more sound conclusions regarding the observed changes and their nature. For assessing factorial invariance, a set of nested structural equation models with increasing invariance restrictions are usually specified. First, the fit of the less restricted model (configural invariance) is examined. More constrained versions are considered in successive steps.

Here we analyze invariance from a multi-group approach. This procedure allows testing specific hypothesis regarding

[1] Please note that the first and second groups were analyzed in Colom et al.'s (2010) report for addressing another research goal.

factorial invariance and differences among the groups of interest, namely, memory span practice, processing speed practice, and no practice. We compared a series of nested models with increased levels of invariance across time points and groups. The analyses were carried out in Mplus 7 (Muthén & Muthén, 1998–2012) using maximum likelihood.

2.5. The baseline model

The starting SEM (Model 1) is presented in Fig. 1. This model was fitted separately for each group.

We use here the notation proposed by McArdle (1988) and Widaman et al. (2010). The observed variables are represented as squares, the latent variables as circles, and the fixed variable (constant unit for computing means and intercepts) as a triangle. The arrows from the constant to the latent variables (α_1 and α_2) represent latent variable means, whereas the arrows from the constant to the observed variables (τ_{11} to τ_{24}) represent intercepts. The arrows from the latent variables to their respective observed indicators (λ_{11} to λ_{24}) represent factor loadings. The factor metric is defined by fixing the DAT-SR factor loading to one at both measurement occasions ($\lambda_{14}=\lambda_{24}=1$), and fixing the latent means $\alpha_1=\alpha_2=0$. The unique variance for each observed indicator is represented by θ_{11} to θ_{24}. The variances of the pre and post latent variables are represented by σ^2_{11} and σ^2_{22}, and their covariance by σ_{12}. Covariances between observed indicators across pre and post measures (not depicted in Fig. 1) are allowed in order to account for correlations between test-specific variance over time.

2.6. Evaluating invariance across measures and groups

Starting with Model 1, we fit a series of models with increased restrictions for assessing specific hypothesis about the data. This series is consistent with the procedure for evaluating factorial invariance proposed by Widaman and Reise (1997) and Widaman et al. (2010), but it introduces some modifications for formally testing hypothesis regarding group differences.

In Model 2, the factor loadings are constrained to be equal *across measurements and groups* ($\lambda_{11}=\lambda_{21}$, $\lambda_{12}=\lambda_{22}$, $\lambda_{13}=\lambda_{23}$, $\lambda_{14}=\lambda_{24}=1$). If this model shows adequate fit, weak factorial invariance holds. Moreover, assuming equal factor loadings across groups entails that the different cognitive practice imposed by our design does not cause differences in the relative weight of the latent variables on each observed indicator.

In Model 3, the intercepts at *pre* measurement (τ_{11} to τ_{14}) are constrained to be equal across the three groups (maintaining $\alpha_1=0$). If this model shows adequate fit, the three groups can be assumed to be equivalent before cognitive practice, since they are described by an identical set of parameters.

In Model 4, the intercepts in the *post* measurement (τ_{21} to τ_{24}) are constrained to be equal across the three groups (maintaining $\alpha_2=0$). If we find adequate fit for this model, the groups can be assumed to be exactly equal both *before* and *after* cognitive practice. Assuming the latter implies that differential cognitive practice cause no differences in the means of the observed indicators, and hence the three groups show exactly the same change from the pretest to the posttest session.

In Model 5, the intercepts are constrained to be equal *for all groups, both in pre and post measures* ($\tau_{11}=\tau_{21}$, $\tau_{12}=\tau_{22}$, $\tau_{13}=\tau_{23}$, $\tau_{14}=\tau_{24}$) and, importantly, the latent variable mean in the post measurement (α_2) is estimated freely and separately for each group. Note that the indicators' mean scores are not supposed to be equal in pre and post measures if practice has any effect. Hence, by freely estimating α_2, the model allows testing whether the observed changes in the indicators are explained uniquely by changes in the latent variable. If this model shows adequate fit, we can conclude so. Besides, this model is intended to capture the extent to which the latent variable accounts for the observed change in each group. Since each mean α_2 is allowed to be different, we can detect latent differences not observable by studying only the tests' scores.

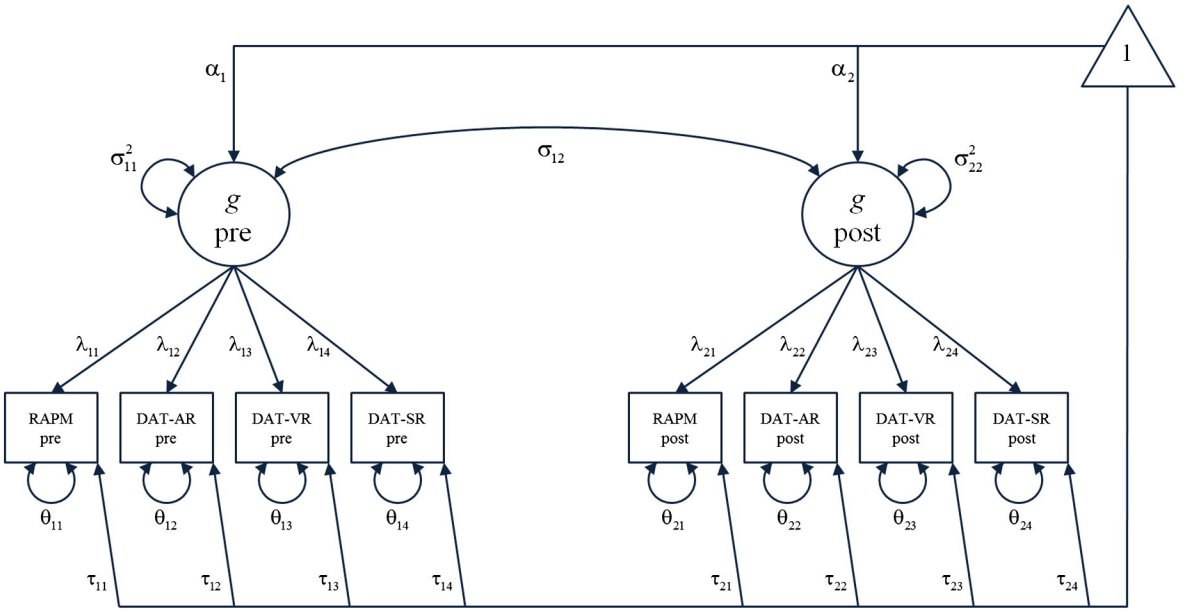

Fig. 1. Baseline model, fitted separately for each of the three groups.

Table 1
Summary of models' restrictions. Parameters in shaded cells are constrained to be equal across groups. Asterisks (*) represent parameters estimated independently for each group.

Parameter	Model 1 Conf. invariance	Model 2 Weak FI. Equal loadings across groups and measures	Model 3 Equal intercepts in t1 across groups	Model 4 Equal intercepts in t1 and t2 across groups	Model 5 Equal intercepts across groups and measures
λ_{11} RAPM pre	*	.609	.606	.608	.542
λ_{12} DAT-AR pre	*	.932	.925	.927	.645
λ_{13} DAT-VR pre	*	.581	.577	.580	.553
λ_{14} DAT-SR pre	1	1	1	1	1
λ_{21} RAPM post	*	.609	.606	.608	.542
λ_{22} DAT-AR post	*	.932	.925	.927	.645
λ_{23} DAT-VR post	*	.581	.577	.580	.553
λ_{24} DAT-SR post	1	1	1	1	1
τ_{11} RAPM pre	*	*	10.46	10.47	10.52
τ_{12} DAT-AR pre	*	*	12.05	12.04	11.28
τ_{13} DAT-VR pre	*	*	11.71	11.71	11.86
τ_{14} DAT-SR pre	*	*	12.11	12.10	12.32
τ_{21} RAPM post	*	*	*	11.88	10.52
τ_{22} DAT-AR post	*	*	*	12.25	11.28
τ_{23} DAT-VR post	*	*	*	13.29	11.86
τ_{24} DAT-SR post	*	*	*	15.03	12.32
α_1 g pre	0	0	0	0	0
α_2 g post	0	0	0	0	*

2.7. Model fit

Each model is a constrained version of the previous one. Hence, they are nested models and thus can be evaluated with likelihood ratio χ^2 difference tests for comparing fit differences across models. We also computed the root mean square of approximation (*RMSEA*; Browne & Cudeck, 1993), the comparative fit index (*CFI*; Bentler, 1990), and the Tucker-Lewis index (*TLI*; Tucker & Lewis, 1973).

For *RMSEA*, values between 0 and .06 indicate very good fit, values between .06 and .08 indicate reasonable fit, and values greater than .10 indicate poor fit. For *CFI* and *TLI* indices, acceptable values must be larger than .90 and excellent values must be above .95 For, the χ^2/df ratio, regarding both individual models and comparison of two nested models, values showing a good fit must not be greater than 2 (Schreiber, Nora, Stage, Barlow, & King, 2006; Schweizer, 2010).

3. Results

Table 1 summarizes the constraints across all models. When the parameters are fixed, the values 0 (factor means) and 1

Table 2
Correlations between *g pre* and *g post* across models and groups.

Group	Model 1	Model 2	Model 3	Model 4	Model 5
Memory span ($n = 170$)	.975	.976	.977	.975	1
Processing speed ($n = 114$)	1	1	1	1	>1
Passive control ($n = 193$)	.984	.977	.978	.978	.993

(DAT-SR loadings) are shown. When they are freely estimated (i.e., same value across groups), the parameter estimates are shown.

The differences between the five models are determined by the parameters constrained to be equal for the three groups in each model. Those parameters are shaded in Table 1. The increased restrictions entail greater invariance at each step. None of the models presented improper estimates (such as negative variances) for any parameter. The latent variables *g pre* and *g post* were properly identified for all models. All the estimated factor loadings λ fell within the expected range and were coherent among models. All the factor loadings were > .5 showing a strong relationship with the latent construct *g*. Among the four observed variables, DAT-SR showed the strongest relationship with *g*, followed by DAT-AR, RAPM, and, finally, DAT-VR. All the intercepts were coherent with the metric and mean of their respective observed variables.

Importantly, the estimated correlations between *pre* and *post g* were very high across groups and models. These correlations are reported in Table 2. Such high correlations indicate that the relative order of the participants, and the distances between them, remained fairly constant. They did not change after practice. Participants showing higher *g* values in *t*1 also have higher *g* values in *t*2.

Table 3 reports the fit from each tested model and includes the likelihood ratio χ^2 difference tests between successive models.

The excellent fit for Model 1 indicates that the hypothesized model adequately described the data.

When factor loadings were restricted to be equal across groups and measures (Model 2), the misfit did not increase statistically ($\Delta\chi^2/\Delta df = 23.04/15$, p = .083), indicating that weak factorial invariance can be assumed. This condition implies that the relative weight of the latent factor on each test did not change after practice, and it was the same for the three groups. When the intercepts of the observed indicators were restricted to be equal in the first measure for the three groups (Model 3), the misfit did not increase statistically ($\Delta\chi^2/\Delta df = 9.67/8$, p = .289). Consequently, the three groups can be considered identical before practice.

When the intercepts of the observed indicators were constrained to be equal also in the second measure (Model 4), the misfit did not increase statistically ($\Delta\chi^2/\Delta df = 5.64/8$, p = .687). The good fit shown by this model implies that the three groups were equal before and after practice, meaning that differential practice was not associated with neither any differential effect in the latent variable weight on the test, nor the tests' means. Note that in Model 4 the intercepts are allowed to change from pre to post, though each parameter set is identical for the three groups. When the intercepts were constrained to be invariant across both groups *and measures* (Model 5), a substantial increase in the misfit was found ($\Delta\chi^2/\Delta df = 112.25/1$, p = .001). Therefore, Model 5 was not as tenable as previous models. Its fit indices indicate that the changes at the tests' level were not properly explained by the net change in the underlying latent variable. In other words, each test changed to a different extent, and that change was not properly explained by the common factor.

Since α_1 was fixed to zero in Model 5, the parameter α_2, which was freely estimated for each group, was intended to inform about the change in the latent factor. However, the poor fit implies that the α_2 values were not tenable given the data. Nevertheless, it should be noted that its estimated value was fairly similar across groups: memory span group $\alpha_2 = 2.213$, *CI*95% = [1.71, 2.53]; processing speed group $\alpha_2 = 2.399$, *CI*95% = [2.01, 2.79]; no practice group $\alpha_2 = 2.454$, *CI*95% = [2.08, 2.83] (for the descriptive statistics regarding tests' raw scores, see the *Appendix 2*).

4. Discussion

In this report we examined whether changes across testing sessions in a set of four standardized intelligence measures can be accounted for by a common latent factor representing general intelligence. This was evaluated considering three groups of participants who completed cognitive practice sessions (or did nothing, as in a passive control group) between pre-test and post-test intelligence sessions. The three groups showed generalized improvements in the tests' scores (see Appendix 2), but the longitudinal multi-group analyses revealed that a latent factor common to the specific intelligence measures failed to account for the observed changes in test's performance across testing sessions.

Table 3
Fit indices for models of longitudinal invariance.

Fit index	Model 1 Conf. Invariance	Model 2 Weak FI. Equal loadings across groups and measures	Model 3 Equal intercepts in *t*1 across groups	Model 4 Equal intercepts in *t*1 and *t*2 across groups	Model 5 Equal intercepts across groups and measures
χ^2	57.26	80.30	89.97	95.61	207.86
df	45	60	68	76	77
$\Delta\chi^2/\Delta df$		23.04 / 15	9.67 / 8	5.64 / 8	112.25 / 1
CFI	.991	.985	.984	.986	.905
TLI	.983	.979	.980	.984	.896
RMSEA	.041	.046	.045	.040	.103
RMSEA lower 95% *CI*	.000	.010	.011	.000	.087
RMSEA upper 95% *CI*	.071	.071	.068	.063	.120

Specifically, a single latent factor adequately explained the variance-covariance structure across measures and groups. The three groups were equal in the pretest and posttest intelligence sessions, which implies that the factor loadings for each test remained fairly constant. Besides, the factor loadings were the same for the three groups. The average scores on the tests can be considered equivalent for the three groups in the pretest session. Following the practice period, we observed a change in average tests' scores. However, this change was identical for three groups (including the passive group), and, therefore, the observed change can be implied as not related to cognitive practice.

The change at the tests' level was not explained by the latent factor because each test showed differential changes, and this was the case for the three groups. Lack of intercept invariance can be observed in the pattern of gains for different subtests. Tests with very high factor loadings (DAT-SR, $\lambda = 1$, and DAT-AR, $\lambda \approx .92$) show large differences in their effect sizes (DAT-SR has the largest effect size $d > .65$ for the three groups, whereas DAT-AR has the lowest effect size, $d \approx$ zero for the three groups). Therefore, it might be suggested that differences in the DAT-SR are due to specific abilities (not to *g*).

Allowing a different mean structure for each group revealed a substantial decrease in fit. Differential practice, thus, does not seem to cause a different pattern of relations between the latent trait and the manifest indicators.

Furthermore, the estimated correlations between the latent factors at pretest and posttest were close to unity across both models and groups. This indicates that the individuals were equally ranked in both measures, and their distances in the latent score remained constant from the pretest to the posttest session. The general implication of this finding is that the latent factor computed for the pretest and posttest sessions can be modeled as a single factor, which explains the variance-covariance structure in both *pre* and *post* measures.

The results reported here, derived from longitudinal multi-group latent-variable analyses, are consistent with previous research concluding that practice effects are not explained by the tapped latent factor. Jensen (1998) summarized and discussed evidence showing that changes in intelligence's scores across testing occasions are unrelated to a general factor of intelligence (*g*). He revised test-retest change in scores, spontaneous changes in scores, the secular increase in intelligence (Flynn effect), and intervention programs aimed at raising intelligence, beginning with the premise that "*an individual's test score is not a measure of the quantity of the latent trait per se possessed by that individual*" (p. 311). The main conclusion was that changes in tests' scores are *hollow* regarding the latent trait of interest, namely, general intelligence (*g*). This same conclusion has been achieved in further studies (Coyle, 2006; te Nijenhuis, Voskuijl, & Schijve, 2001; te Nijenhuis et al., 2007; Wicherts et al., 2004).

Bors and Vigneau (2003) analyzed the effect of practice on the Progressive Matrices Test (RAPM) finding that increased scores across three testing occasions involve learning processes (how to solve reasoning problems). The observed improvements were not explained by modified strategies for completing the test or the memorization of information associated with the specific problems. Importantly, they found that the rank ordering of tests' scores obtained by the participants showed a large stability.

The recent report by Hayes, Petrov, and Sederberg (2015) concluded that score gains across testing occasions do not reflect changes in intelligence. Rather, these gains may reflect strategy refinement (knowledge presumably acquired during testing). Their eye fixation results led to the conclusion that a substantial portion of the variance in scores' gains is derived from improvements in problem-solving strategies. Nevertheless, these researchers analyzed a single test and they acknowledged that intelligence is a latent variable.

As noted at the introduction section, there are now several reports analyzing changes across testing occasions using several intelligence measures (Chooi & Thompson, 2012; Colom et al., 2013; Harrison et al., 2013; Redick et al., 2012; Stephenson & Halpern, 2013; von Bastian & Oberauer, 2013). However, none of these studies was designed to apply the recommended longitudinal multi-group latent variable analyses, mainly because the sample sizes were small. To our knowledge, the study by Schmiedek, Lovden, and Lindenberger (2010) is the only exception. Their study analyzed 204 adults that practiced three tests of working memory, three tests of episodic memory, and six tests of perceptual speed. The effects were analyzed with latent difference score models, which allowed answering the following question: to what degree relationships with the latent variable explain the means and variance in the specific measures? The latent difference scores quantify changes at the latent level. The results showed reliable gains at this latent level, suggesting improvements in cognitive efficiency beyond changes in strategy or mere knowledge.

The results reported in the present study, also derived from latent modeling, do not support this latter conclusion. Quite to the contrary, changes observed at the tests' level were not explained by the latent factor, and this was the case for the three groups (memory span, processing speed, and passive control). Nevertheless, as discussed by Colom et al. (2013), constructs are not homogeneous entities, meaning that their manifest indicators may show differential behaviors across testing occasions. To some extent, this argument is consistent with Flynn (2007): "*imagine that score gains on all of the WISC subtests were three times as great as they are. They would still have the same pattern, that is, they would still flunk the test of factor invariance and not qualify as g gains against that criterion. But could we dismiss the enhancement of performance on so many cognitively complex tasks? (…) the mere fact that the pattern of gains does not correlate with the differential g loadings of the subtests will not make the gains go away*" (p. 61). The ultimate criterion for assessing whether or not changes in tests' scores reflect real improvements in the latent trait might come from the analysis of performance changes in real life settings. We still lack solid evidence regarding this crucial issue. Therefore, further research in this area is strongly required.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.intell.2015.02.004.

References

Anastasi, A. (1934). Practice and variability. *Psychological Monographs*, *45*, 5. http://dx.doi.org/10.1037/h0093355.

Ashton, M. C., & Lee, K. (2005). Problems with the method of correlated vectors. *Intelligence*, *33*(4), 431–444. http://dx.doi.org/10.1016/j.intell.2004.12.004.

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2014). Improving fluid intelligence with training on working memory: A meta-analysis. *Psychonomic Bulletin & Review*. http://dx.doi.org/10.3758/s13423-014-0699-x.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1990). *Differential Aptitude Test* (5th ed.). Madrid: TEA.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246. http://dx.doi.org/10.1037/0033-2909.107.2.238.

Bors, D. A., & Vigneau, F. (2003). The effect of practice on Raven's Advanced Progressive Matrices. *Learning and Individual Differences*, *13*, 291–312. http://dx.doi.org/10.1016/S1041-6080(03)00015-3.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Buschkuehl, M., Hernandez-Garcia, L., Jaeggi, S. M., Bernard, J. A., & Jonides, J. (2014). Neural effects of short-term training on working memory. *Cognitive, Affective, and Behavioral Neuroscience*. http://dx.doi.org/10.3758/s13415-013-0244-9.

Buschkuehl, M., & Jaeggi, S. M. (2010). Improving intelligence: A literature review. *Swiss Medical Weekly*, *140*(19–20), 266–272.

Chooi, W., & Thompson, L. A. (2012). Working memory training does not improve intelligence in healthy young adults. *Intelligence*, *40*, 531–542. http://dx.doi.org/10.1016/j.intell.2012.07.004.

Colom, R., Abad, F. J., García, L. F., & Juan-Espinosa, M. (2002). Education, Wechsler's Full Scale IQ, and *g*. *Intelligence*, *30*, 449–462. http://dx.doi.org/10.1016/S0160-2896(02)00122-8.

Colom, R., Jung, R., & Haier, J. (2006). Finding the *g*-factor in brain structure using the method of correlated vectors. *Intelligence*, *34*, 561–570. http://dx.doi.org/10.1016/j.intell.2006.03.006.

Colom, R., Quiroga, Ma. A., Shih, P. C., Martínez, K., Burgaleta, M., Martínez-Molina, A., et al. (2010). Improvement in working memory is not related to increased intelligence scores. *Intelligence*, *38*, 497–505. http://dx.doi.org/10.1016/j.intell.2010.06.008.

Colom, R., Román, F. J., Abad, F. J., Shih, P. C., Privado, J., Froufe, M., et al. (2013). Adaptive n-back training does not improve fluid intelligence at the construct level: Gains on individual tests suggest that training may enhance visuospatial processing. *Intelligence*, *41*, 712–727. http://dx.doi.org/10.1016/j.intell.2013.09.002.

Conway, A. R. A., & Getz, S. J. (2010). Cognitive ability: Does working memory training enhance intelligence? *Current Biology*, *20*, 8. http://dx.doi.org/10.1016/j.cub.2010.03.001.

Coyle, T. R. (2006). Test-retest changes on scholastic aptitude tests are not related to *g*. *Intelligence*, *34*, 15–27. http://dx.doi.org/10.1016/j.intell.2005.04.001.

Dolan, C. V., & Hamaker, E. L. (2001). Investigating black–white differences in psychometric IQ: Multi-group confirmatory factor analyses of the WISC-R and K-ABC and a critique of the method of correlated vectors. In F. Columbus (Ed.), *Advances of Psycho- logical Research, vol. 6*. (pp. 31–59). Huntington7: Nova Science Publishers.

Flynn, J. R. (2007). *What is intelligence? Beyond the Flynn effect*. Cambridge: Cambridge University Press.

Haier, R. J. (2014). Increased intelligence is a myth (so far). *Frontiers in Systems Neuroscience*. http://dx.doi.org/10.3389/fnsys.2014.00034.

Harrison, T. L., Shipstead, Z., Hicks, K. L., Hambrick, D. Z., Redick, T. S., & Engle, R. W. (2013). Working Memory Training May Increase Working Memory Capacity but Not Fluid Intelligence. *Psychological Science*. http://dx.doi.org/10.1177/0956797613492984 (Published online 3 October 2013).

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2015). Do we really become smarter when our fluid intelligence test scores improve? *Intelligence*, *48*, 1–14. http://dx.doi.org/10.1016/j.intell.2014.10.005.

Herrnstein, R., & Murray, C. (1994). *The Bell Curve*. New York: Free Press.

Hunt, E. B. (2011). *Human intelligence*. Cambridge: Cambridge University Press.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *PNAS*, *105*, 6829–6833. http://dx.doi.org/10.1073/pnas.0801268105.

Jaušovec, N., & Jaušovec, K. (2012). Working memory training: Improving intelligence — Changing brain activity. *Brain and Cognition*, *79*, 96–106. http://dx.doi.org/10.1016/j.bandc.2012.02.007.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Jensen, A. R. (1998). *The g factor*. New York: Praeger.

McArdle, J. J. (1988). Dynamic but Structural Equation Modeling of Repeated Measures Data. In J. R. Nesselroade, & R. B. Cattell (Eds.), *Handbook of Multivariate Experimental Psychology* (pp. 561–614) (2nd ed.). US: Springer.

Melby-Lervåg, M., & Hulme, C. (2012). Is working memory training effective? A meta-analytic review. *Developmental Psychology*. http://dx.doi.org/10.1037/a0028228 (Advance online publication).

Meredith, W. M. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*, 525–543. http://dx.doi.org/10.1007/BF02294825.

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In L. M. Collins, & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 203–240). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10409-007.

Moody, D. E. (2009). Can intelligence be increased by training on a task of working memory? *Intelligence*, *37*, 327–328. http://dx.doi.org/10.1016/j.intell.2009.04.005.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus User's Guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., et al. (2012). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology: General*. http://dx.doi.org/10.1037/a0029082 (Advance online publication).

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. http://dx.doi.org/10.1016/j.intell.2005.05.003.

Rudebeck, S. R., Bor, D., Ormond, A., O'Reilly, J. X., & Lee, A. C. (2012). A potential spatial working memory training task to improve both episodic memory and fluid intelligence. *PLoS One*, *7*(11), e50431. http://dx.doi.org/10.1371/journal.pone.0050431.

Schmiedek, F., Lovden, M., & Lindenberger, U. (2010). Hundred days of cognitive training enhance broad cognitive abilities in adulthood: findings from the COGITO study. *Frontiers in Aging Neuroscience*. http://dx.doi.org/10.3389/fnagi.2010.00027.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *The Journal of Educational Research*, *99*, 323–338. http://dx.doi.org/10.3200/JOER. 99.6.323-338.

Schweizer, K. (2010). Some Guidelines Concerning the Modeling of Traits and Abilities in Test Construction. *European Journal of Psychological Assessment*, *26*, 1–2. http://dx.doi.org/10.1027/1015-5759/a000001.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2010). Does working memory training generalize? *Psychological Belgica*, *50*(3–4), 245–276.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin*. http://dx.doi.org/10.1037/a0027473 (Advance online publication).

Stephenson, C. L., & Halpern, D. F. (2013). Improved matrix reasoning is limited to training on tasks with a visuospatial component. *Intelligence*, *41*, 341–357. http://dx.doi.org/10.1016/j.intell.2013.05.006.

te Nijenhuis, J., van Vianen, A. E. M., & van der Flier, H. (2007). Score gains on *g*-loaded tests: No *g*. *Intelligence*, *35*, 283–300. http://dx.doi.org/10.1016/j.intell.2006.07.006.

te Nijenhuis, J., Voskuijl, O. F., & Schijve, N. B. (2001). Practice and coaching on IQ tests: Quite a lot of *g*. *International Journal of Selection and Assessment*, *9*, 302–308. http://dx.doi.org/10.1111/1468-2389.00182.

Tidwell, J. W., Dougherty, M. R., Chrabaszcz, J. R., Thomas, R. P., & Mendoza, J. L. (2013). What accounts as evidence for working memory training? Problems with correlated gains and dichotomization. *Psychological Bulletin & Review*. http://dx.doi.org/10.3758/s13423-013-0560-7.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1–10. http://dx.doi.org/10.1007/BF02291170.

von Bastian, C. C., & Oberauer, K. (2013). Distinct transfer effects of training different facets of working memory capacity. *Journal of Memory and Language*. http://dx.doi.org/10.1016/j.jml.2013.02.002.

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., van Baal, C. M., Boomsma, D. I., et al. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence*, *32*(5), 509–537. http://dx.doi.org/10.1016/j.intell.2004.07.002.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial Invariance Within Longitudinal Structural Equation Models: Measuring the Same Construct Across Time. *Child Development Perspectives*, *4*, 10–18. http://dx.doi.org/10.1111/j.1750-8606.2009.00110.x.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.