Estimating Correlations and Reliabilities of Implicit and Explicit Tests Using a Latent Variable Approach

Ingmar Visser email: i.visser@uva.nl and **Ton Blaazer**
Department of Psychology, University of Amsterdam
Roetersstraat 15
1018 WB Amsterdam
phone: +31 (20) 5256735
fax: +31 (20) 6390279

Abstract

In many fields in psychology implicit tests are used to measure some construct of interest, such as priming measures to test implicit memory and the implicit association test to measure implicit attitudes. In sequence learning, reaction times are regarded as an implicit test of sequence knowledge. The validity and interpretation of implicit tests hinges on its relationship with similar explicit tests. Using different explicit measures, both large associations and dissociations have been reported between measures of sequence knowledge. Part of this inconclusiveness of the results may be due to different reliabilities and sensitivities of the measures being used. In this paper, a latent variable model approach is used to assess differences in reliability and sensitivity of reaction times and online prediction performance in a typical sequence learning task, with the aim of arriving at precise estimates of the correlations between implicit and explicit measures of sequence knowledge.

Introduction

In many fields in psychology implicit test procedures are used, for testing self-esteem, for testing attitudes, and in learning and memory research. Many of these procedures rely on reaction times (RTs). That is, when someone has a negative attitude towards caucasians, one would not necessarily expect them to express this in an explicit questionnaire. However, it may very well be possible to detect such attitudes in an implicit association test by detecting differential responding to, say, pictures of black and white people (Fazio & Olson, 2003). Similarly, in memory research, priming effects may occur in the absence of explicit recollection of the presented material. In implicit learning research, a speed-up in RTs is observed in the absence of verbally reportable knowledge (Cleeremans & McClelland, 1991).

The use of implicit measures differs considerably in these fields, and so does the interpretation of associations and dissociations that exist between implicit and explicit measures. However, there is one important similarity: the validity and interpretation of implicit tests depends crucially on precise knowledge about its relationship with corresponding explicit tests. This relationship, be it an association or a dissociation, is greatly influenced by the reliability of the measures involved. As a consequence of differing reliabilities and sensitivities of implicit and explicit measures, the true relationships between these measures may be obscured (Buchner & Wippich, 2000; Meier & Perrig, 2000).

Many debates about the relationships between implicit and explicit measures center on the type of knowledge representations that underly participants' responses to these measures. There are two extreme possibilities for these knowledge representations. The first possibility is that there are two kinds of knowledge, implicit and explicit. One is measured by RTs, as in priming effects or in the implicit association test, and the other is measured by explicit tasks such as questionnaires (in attitude research) or recognition or recall tasks (in memory research). The assumption of two kinds of knowledge implies that it should be possible to dissociate them.

The second possibility is that there is only one kind of knowledge but there are different measures, implicit or explicit (see e.g. work by Shanks & colleagues). Fazio & Olson (2003) state something similar about attitudes: "... it is more appropriate to view the *measure* as implicit or explicit, *not* the attitude" (emphasis added). In taking this position, researchers implicitly take on the challenge of finding plausible explanations of dissociations that are nonetheless found between implicit and explicit measures.

The focus of this paper is on implicit sequence learning, although the proposed method of using factor analysis to arrive at estimates of correlations between implicit and explicit measures of a construct can be applied more generally, and indeed has been recently in the area of attitude research (Blanton, Jaccard, Gonzalez & Christie 2006). In typical sequence learning experiments, participants are presented with a sequence of stimuli that they have to respond to by pressing an appropriate response button. Unbeknownst to participants, the sequence of stimuli contains regularities that make upcoming stimuli predictable by previously seen stimuli. As a result of this, a decrease in RTs is observed relative to a control condition with differently or randomly structured stimuli. To establish that the learning process that underlies the speed-up in RTs is indeed implicit, a test of explicit knowledge is administered after the RT phase of the experiment, and the correlation between implicit and explicit test is computed. The results of such studies however, are mixed, with some studies pointing to limited reportable knowledge (Reber, 1967; Cleeremans & McClelland, 1991) and others pointing to high correlations between implicit and explicit measures (Perruchet & Amorim, 1992; Shanks & Johnstone, 1999). These differences may be partly due to the different reliabilities of implicit and explicit measures that are involved, and in this paper we show a method of estimating those.

Implicit sequence learning

The status of knowledge resulting from implicit learning is highly controversial with people claiming that human learning is systematically accompanied by awareness (Shanks & StJohn, 1994) and others claiming that unconscious learning is a fundamental process in human cognition (Reber, 1993). Another exponent of the latter view is Destrebecqz, who has found a strong dissociation between an implicit and an explicit measure of sequence knowledge (Destrebecqz & Cleeremans, 2001). It should be noted though that others have failed replicate their results (Wilkinson & Shanks, 2004).

In much of the implicit learning research, the goal is establish implicit learning effects in the absence of explicit knowledge. As a consequence, the validity of the explicit knowledge test that is used is pivotal in evaluating this research. Since Reber's 1967 paper on implicit learning of artificial grammars, there has been a heated debate about the validity of different explicit knowledge tests. In particular, it has been argued that the verbal report task that was used by Reber to assess explicit knowledge at the end of the experiment, is not sensitive enough to bring out all the explicit knowledge that participants may have in such an experiment (Perruchet & Amorim, 1992; Shanks & StJohn, 1994). As a result, researchers have proposed other measures of explicit knowledge such as the recognition task and the generation task.

In the recognition task, portions of the sequence of stimuli that was presented to participants in the RT task, are presented to them, and they are asked to give a recognition rating. The results from the use of such recognition tasks are similarly inconclusive as the results from the verbal report tasks. Destrebecqz and Cleeremans (2001) found near baseline performance, whereas others found high correlations between priming effects as measured by RTs and a recognition task (Shanks & Perruchet, 2002).

In the generation task, participants are required to reproduce the sequence of stimuli that was presented to them in the RT task. In different versions of the generation task, again different researchers found diverging results, varying from near baseline performance (Cleeremans & McClelland, 1991; Jiménez, Méndez & Cleeremans, 1996) to high correlations between RTs and generation performance on trigrams of generated stimuli (Perruchet & Amorim, 1992.

Some of the objections that were raised against the verbal report task can be raised similarly against different versions of the recognition and generation tasks. In particular, the sensitivity and reliability of these tasks are unknown. This is also true, of course, for the typical RT measures that are used in sequence learning tasks. As a consequence, observed correlations between implicit and explicit measures can not be taken at face value. Buchner & Wippich (2000) have shown that observed correlations between implicit and explicit memory tasks can be greatly reduced by different reliabilities of the measures involved. They propose to correct the observed correlations by split-halves reliability estimates. In this paper, a similar approach is taken to compare observed and true correlations between implicit and explicit measures of sequence knowledge.

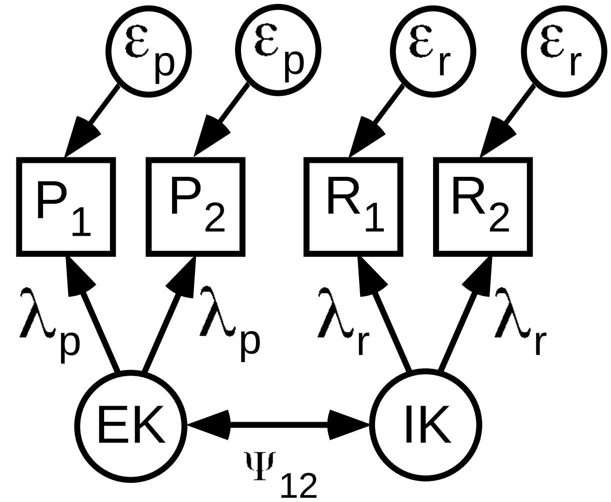

Figure 1: Factor model.

This current paper has three aims. First, to devise a direct measure of sequence knowledge that can be measured repeatedly, and which can be measured concurrently with RTs, which are used as an indirect measure of sequence knowledge. Second, to establish the reliability of each of these measures. Third, and most importantly, to arrive at estimates of the correlations between these measures in such a way as to account for their different reliabilities.

Analyzing reliability

In sequence learning, there are typically many repeated measurements of the implicit measure of knowledge: RTs. This feature of sequence learning experiments can be exploited to arrive at reliability estimates of the RT measure. In the factor model proposed in the current paper, the repeated measurements of RTs are used as indicators of the underlying factor of implicit knowledge. Figure 1 depicts the factor model that is used to analyze the data. In this model, P_1 and P_2 are the repeated (parallel) measures of prediction performance, and R_1 and R_2 are the repeated measures of RTs. These indicators are pairwise regressed on a common underlying factor representing explicit knowledge (EK) and implicit knowledge (IK) respectively. Each of the indicators is also associated with a reliability or measurement error term ε_p, and ε_r for the measurement error of prediction performance and RTs, respectively.

The model for the observed data can hence be expressed as:

$$P_i = \lambda_p EK + \varepsilon_p,\ i = 1, 2, \tag{1}$$

$$R_i = \lambda_r IK + \varepsilon_r,\ i = 1, 2, \tag{2}$$

where ε_p and ε_r are the measurement errors, and λ_p and λ_r express the strength of the relationship between the EK factor and prediction performance, and the strength of the relationship between the IK factor and reaction times respectively. The parameter of most interest in this model is Ψ_{12}, the correlation between the latent factors EK and IK. Estimates of this parameter can be compared with the observed correlations between prediction performance and reaction performance.

To be able to apply this model, repeated measurements of prediction performance must be available as well as repeated measurements of RTs. In the experiment presented below, a task is designed to allow these repeated measurements of prediction performance.

Experiment

The goal of the experiment is to devise a direct measure of sequence knowledge that can be measured repeatedly, and concurrently with RTs. The measure that is used for this purpose is an *online* prediction task in which participants are required to predict the upcoming stimuli at random trials during a serial RT task.

Method

Participants Participants were seven psychology students from the Department of Psychology from the University of Amsterdam who received course credits for their participation in the experiment.

Procedure Participants were seated in front of an Apple Imac computer and given a four-choice serial reaction time (SRT) task. The experiment consisted of 11 blocks of 144 trials each. Each block consisted of 12 repetitions of a second-order conditional sequence of length 12: CDABDCADBACB, except in block 9, in which a transfer sequence was used: CDBCABADCBDA. Such sequences are frequently used in sequence learning experiments (Perruchet & Amorim, 1992; Shanks & Johnstone, 1999). Stimuli were presented in one of four locations that were organized in a 2×2-grid as shown in Figure 2. At RT trials, participants were required to press the appropriate response on the numerical key-pad of an ordinary keyboard. The response were congruently mapped to the screen positions; keys 1, 2, 4, and 5 were used.

Reaction time trials were interspersed with *online* prediction trials at random points in the sequence. Prediction trials were signaled by a display as depicted in Figure 2. Note that the letters were not part of the actual display. Four question marks were placed at each of the possible screen locations, and participants were instructed to predict the location of the next stimulus. In the instructions, participants were made aware that in the RT trials there were no repeating trials. At prediction trials, they were instructed to refrain from typing the same stimulus that was presented at the RT trial immediately before the prediction trial. In each block, there were 24 prediction trials, 2 at each position of the 12-element repeating sequence. A block never started with a prediction trial, and there were at least 3 RT trials in between consecutive prediction trials. Since

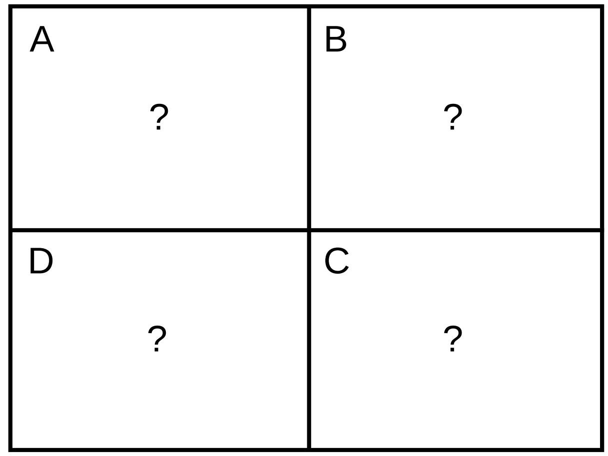

Figure 2: Display for prediction trials.

the sequences that were used in this experiment were second-order conditional, this spacing of prediction trials ensured that correct prediction was always possible based on previously seen stimuli. Earlier research established that online prediction task does not affect the learning process in any significant way (Visser, Raijmakers & Molenaar, 2006).

After the reaction time and online prediction phase of the experiment were completed, a free generation task was administered. In this task, participants are presented an initial stimulus, and are then required to generate a sequence of trials that mimics the sequence of trials they were exposed to in the reaction time task. In none of the tasks, feedback was provided to participants. It has been argued that especially in the generation task, providing feedback may lead to undesirable side-effects (Perruchet & Amorim, 1992; Shanks & Johnstone, 1999). Hence, in order to have maximal congruence between the RT task, the prediction task and the generation task, no feedback was provided at any of them.

Results

To establish the well-known effects that are found in sequence learning, mean RTs were computed for each block and for each participant separately. The means of these means are plotted in Figure 3. The results unsurprisingly replicate standard findings in sequence learning: RTs decreased as a function of exposure to the repeating sequence, and increased when the transfer sequence was presented in block 9. The main effect of block number on RTs is $F(10, 60) = 5.21, p < 0.001$. More importantly, there is a significant effect of transferring participants to previously unseen stimuli in block 9 when compared with the reaction times in block 10, $F(1, 6) = 16.24$, $p < 0.01$. Because of possible deviations from normality, this difference was also tested using a Wilcoxon test providing essentially the same result $Z = -2.37$, $p < 0.05$.

In Figure 4, the percentages correct predictions in each block of trials are plotted. The pattern of prediction performance mirrors that of the RT performance.

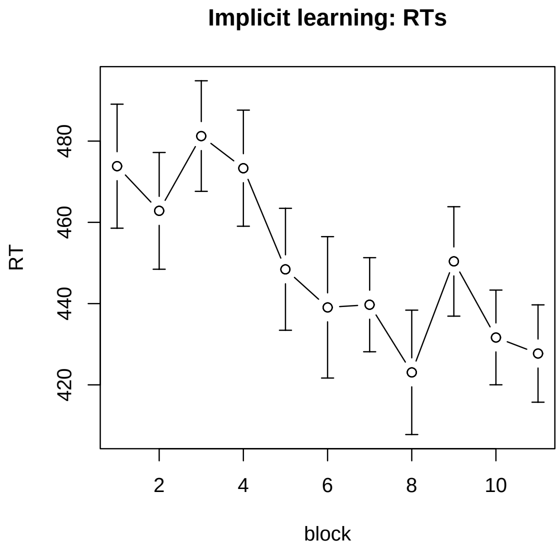

Figure 3: Mean reaction times.

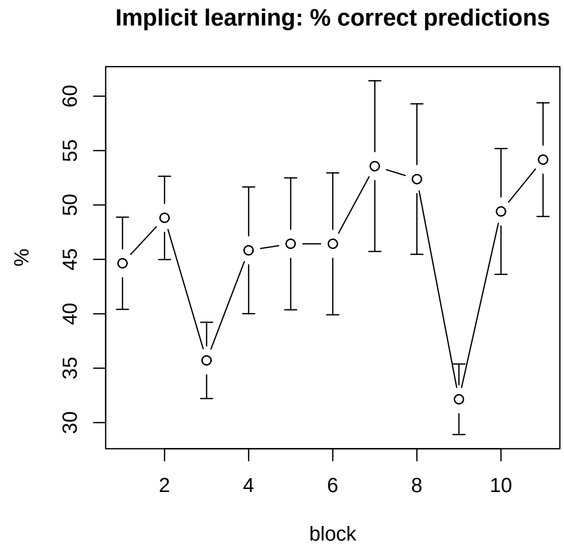

Figure 4: Mean percentage correct predictions.

There is an overall increase in prediction performance as confirmed by a main effect of block on percentage correct $F(10, 60) = 3.32, p < 0.01$. Similarly, there is a large drop in prediction performance when participants are transferred to a different sequence; the difference between prediction performance in block 9 and block 10 is significant, $F(1, 6) = 7.76, p < 0.05$. Similar to the RT data, a Wilcoxon test provides the same result $Z = -1.997$, $p < 0.05$.

The result of decreasing RTs mimics standard findings in sequence learning. The increase of prediction performance is also unsurprising, although only a few studies have used a concurrent direct measure of sequence knowledge, but see Shanks & Perruchet (2002). Only Visser, Raijmakers & Molenaar (2006) have used concurrent direct and indirect measurements of sequence knowledge throughout the learning phase of a sequence learning experiment. In the next section, the relationship between this measure and the reaction times is explored in more detail.

Correlations between reaction times and prediction performance The observed correlation between mean RTs and mean prediction performance over the 11 blocks of the experiment is $r = -0.60, t = -2.2128, df = 9, p = 0.0542$, where r has a 95 percent confidence interval from -0.880 to 0.0097. When leaving out block 9, the correlation changes to $r = -.80$ with $p < 0.05$.

In order to compare prediction performance with reaction times in more detail, the factor model from Figure 1 was fitted on the reaction times and percentages correct predictions of blocks 1 through 8 and blocks 10 and 11. Block 9 was left out of these analyses because in that block the transfer sequence was presented to participants. Model fitting was done in the following way. First, mean RTs to the last trial of each trigram occurring in the repeating sequence were computed. This was done separately for each block. For example, the mean reaction time to a D stimulus occurring after AB was 438.3 ms in block 1, and 413.5 ms in block 11. Second, and similarly, the percentages correct predictions were computed for each trigram, and for each block separately. For example, the percentage correct predictions for predicting a D after AB was 21.4% in block 1 and 67.1% in block 11. Factor models were estimated on the basis of the correlation matrices between two measures of RTs of subsequent blocks and two measures of prediction performance of subsequent blocks. Such observed correlations between RTs and an explicit measure of performance are typically reported in research on the relationship between implicit and explicit sequence knowledge (see e.g. Perruchet & Amorim, 1992; Jiménez et al., 1996).

First, five separate factor models were fitted on data from five pairs of consecutive blocks of the experiment; the resulting model parameters, along with the goodness-of-fit measures, and the observed correlations, are reported in Table 1. Reaction times from block 1 served as indicator R_1 in the factor model, and RTs from block 2 served as indicator R_2 in the model. Similarly, prediction performance in block 1 served as indicator P_1 in the model, and performance in block 2 as indicator P_2; and similarly for blocks 5 through 8 and blocks 10 and 11. To identify the factor model, the measurement error parameters ε_{p_1} and ε_{p_2} were set to be equal, and so were the measurement error parameters ε_{r_1} and ε_{r_2}. Similarly, the factor loadings related to the EK factor λ_{p_1} and λ_{p_2} were set to be equal, and so were the factor loadings related to the IK factor λ_{r_1} and λ_{r_2}. Models were fitted using Lisrel (Jöreskog & Sörbom, 1999).

Table 1 reports three goodness-fit-measures, the χ^2, along with the corresponding df and p-value, the CFI

Table 1: Observed (obs) correlations and latent (lat) correlations.

model	obs	low/up	lat	low/up	$\varepsilon_p/\varepsilon_r$	χ^2	*df*	p	CFI	SRMR
bl 1 & 2	-.67	-.84/-.36	-.94	-1.00/-.72	.09/.35	2.46	5	.78	1.0	.061
bl 3 & 4	-.75	-.88/-.49	-.75	-.94/-.50	.13/.16	5.56	5	.33	1.0	.076
bl 5 & 6	-.64	-.83/-.32	-.62	-.96/-.42	.13/.31	1.30	5	.93	.98	.021
bl 7 & 8	-.50	-.75/-.12	-.96	-1.00/-.58	.22/.50	4.27	5	.51	1.0	.10
bl 10 & 11	-.42	-.70/-.02	-1.0	-1.00/-.62	.39/.45	1.76	5	.88	1.0	.062
combined	-.60	-.70/-.47	-.84	-.98/-.70	.18/.36	30.3	45	.95	1.0	.090

and the SRMR. The SRMR indicates slight misfit for the model of block 7 & 8 and for the combined data model. According to the χ^2 and CFI criteria all models adequately capture the data (see Hu & Bentler, 1999, for discussion of cut-off values of fit indexes of factor models). Note that the relevant 'sample size' n in these model fits is *not* the number of subjects but rather the number of trigrams that is being analyzed, i.e. $n = 12$.

In each row of the Table, the observed correlation is given along with it's 95% confidence interval, the latent correlation as estimated in the factor model, along with it's 95% confidence interval and the measurement error parameters ε_p and ε_r. The parameter of interest is the correlation between the EK en IK factor that is estimated in the factor model, i.e. the correlation between knowledge measured by prediction performance and the knowledge measured by RTs, as well as the measurement error parameters. As can be seen in Table 1, the correlations between RTs and online prediction are rather high overall. This is even more so for the latent correlations estimated in the fitted factor models, indicating attenuation of the observed correlations. Note that three of the five latent correlations have -1.0 included in their 95% confidence interval, indicating that the data is consistent with a correlation of -1.0 between the EK en IK factors. This observation is confirmed by non-significant χ^2-difference tests for setting the latent correlation to -1 for these three models.

Because the separate models are based on $n = 12$ items only, the confidence intervals of these parameters are quite large. Therefor, we also fitted a combined model on the five data sets together. This model is a multi-group model, in which all parameters were constrained to be equal across the five measurement occasions. The tenability of these constraints together was tested using a χ^2-difference test which was found to be non-significant, $\chi^2 = 14.95$, $df = 20$, $p = .78$, indicating that the constraints did not significantly influence model goodness-of-fit. The parameter estimates and goodness-of-fit indexes of this model are in the last line of Table 1.

In the combined model, the latent correlation between EK and IK is -.84 with a confidence interval of -.70 to -.98, whereas the observed correlation is -.60 with confidence interval endpoints of -.47 and -.70. Hence, the observed correlation between IK and EK is greatly influenced by (different amounts of) measurement error in the implicit and explicit measures. The high latent correlation between the EK and IK factors in the models indicate that a common knowledge base underlies responding to both types of trials. High correlations between direct and indirect measures of sequence learning have been found before, e.g. by Perruchet & Amorim (1992) who found a correlation of -0.8 between RTs and a generation task administered at the end of training.

It should be noted that these correlations have to be interpreted with care, as our sample size is fairly small. Even though sample size is small, the reliability of estimates of correlations need not be threatened. As said above, the appropriate n in these factor models is not the sample size, but the number of trigrams under analysis. Moreover, the data entering into the factor model consists of RTs averaged over many repeated trials and over subjects. As a consequence, those data are much more reliable than the typical situation in factor analysis in which between-subject variability is investigated.

The difference between the observed and latent correlation was tested by fixing the latent correlation to the value of the observed correlation. In the model for the combined data, this results in a χ^2-difference of 6.2 with $df = 1$, $p < .05$, indicating that the latent correlation is significantly larger than the observed correlation.

Measurement error is higher for the RTs (ε_r) than for the prediction performance (ε_p), χ^2-difference of 6.1 with $df = 1$, $p < .05$ for setting those parameters equal, and both are significantly different from zero. Hence, the direct measure of sequence knowledge, *online* prediction, has lower measurement error than does the indirect measure of sequence knowledge, RTs.

Conclusion & discussion

A factor analysis model was presented that allows precise analysis of correlations between implicit and explicit measures of a given construct. This method was applied to a sequence learning task. In order to apply this method, multiple indicators, or multiple measurements of the same type need to be administered to participants. Having multiple measurements provides the possibility of separating measurement error from tests from the underlying constructs they are purported to measure.

An experiment was devised that contains multiple measurements of a direct measure of sequence knowledge in the form of the *online* prediction task. Performance on this task, i.e. the increase in prediction ability, was highly correlated with the (decrease in) RTs. This overall correlation indicates that improvement on both measures proceeds in a similar vein.

The application that was presented clearly indicates that the presence of measurement error attenuates the observed correlations between implicit and explicit measures of underlying constructs. Failing to observe the presence of (different levels of) measurement error may lead to gross underestimates of the correlations of interest. It was shown that in the case of sequence knowledge as measured by reaction times and prediction performance, the data show strong evidence of a single underlying knowledge base for improvement on both measures in a sequence learning experiment. In fact, a number of the estimated correlations were consistent with a correlation of -1.0 between the latent factors EK and IK for explicit and implicit knowledge, and the model for the combined data has a latent correlation of -.84, also indicating a very strong association between implicit and explicit measures of sequence knowledge. These numbers should however be interpreted with care because of the rather small sample size that we tested. Better estimates of the latent correlations may be arrived at by using a multi-level factor model, with random effects for subjects and measurement occasions.

The aim of the present paper was to provide a method for reliably estimating correlations between constructs that are measured with different measurement error. The investigation of measurement error and its influence on correlations between the constructs of interest is essential in the fields of research that use implicit and explicit measures. The factor model that we presented is an excellent tool to investigate these issues, and its application is hence a prerequisite for answering important substantive questions in those fields of research. Moreover, the present methodology can easily be extended to include other measures of sequence knowledge as well, such as the recognition task or a generation task administered at the end of training, as is commonly done in sequence learning research. The only requirement for using this methodology is that multiple indicators are available to measure both implicit and explicit constructs.

Acknowledgments

This research was supported by European Commission grant 51652 (NEST) and by a VENI grant from the Dutch Organization for Scientific Research (NWO).

References

Blanton, H., Jaccard, J., Gonzales, P. M., and Christie, C. (2006). Decoding the implicit association test: Implications for criterion prediction. *Journal of Experimental Social Psychology*, 42(2) 192–212.

Buchner, A. and Wippich, W. (2000). On the reliability fo implicit and explicit memory measures. *Cognitive Psychology*, 40:227–259.

Cleeremans, A. and McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, 120:235–253.

Destrebecqz, A. and Cleeremans, A. (2001). Can sequence learning be implicit? New evidence with the process dissociation procedure. *Psychonomic Bulletin & Review*, 8(2):343–350.

Fazio, R. H. and Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, 54:297–327.

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling*, 6(1): 1–55.

Jiménez, L., Méndez, C., and Cleeremans, A. (1996). Comparing direct and indirect measures of sequence learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22(4):948–969.

Jöreskog, K. and Sörbom, D. (1999). *LISREL 8 [Computer program].* Scientific Software International., Chicago.

Meier, B. and Perrig, W. J. (2000). Low reliability of perceptual priming: Consequences fot the interpretation of functional dissociations between explicit and implicit memory. *The Quarterly Journal of Experimental Psychology*, 53A(1):211–233.

Perruchet, P. and Amorim, M.-A. (1992). Conscious knowledge and changes in performance in sequence learning: Evidence against dissociation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18(4):785–800.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, 6:317–327.

Reber, A. S. (1993). *Implicit learning and tacit knowledge : an essay on the cognitive unconscious.* Oxford University Press, New York etc.

Shanks, D. R. and Johnstone, T. (1999). Evaluating the relationship between explicit and implicit knowledge in a sequential reaction time task. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 25(6):1435–1451.

Shanks, D. R. and Perruchet, P. (2002). Dissociation between priming and recognition in the expression of sequential knowledge. *Psychonomic Bulletin & Review*, 9(2):362–367.

Shanks, D. R. and St-John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences*, 17(3):367–447.

Visser, I., Raijmakers, M. E. J., and Molenaar, P. C. M. (2006). Characterizing sequence learning: Online generation and hidden markov models show associations and dissociations. *Accepted pending minor revisions.*

Wilkinson, L. and Shanks, D. R. (2004). Intentional control and implicit sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2):354–369.