

Ruprecht-Karls-Universität Heidelberg
Fakultät für Verhaltens- und Empirische Kulturwissenschaften

Cognitive ability beyond IQ

Inauguraldissertation
zur Erlangung des akademischen Grades eines
Dr. phil. im Fach Psychologie

eingereicht an der Fakultät für Verhaltens- und Empirische Kulturwissenschaften
der Ruprecht-Karls-Universität Heidelberg

von Dipl.-Psych. Daniel Danner
geboren am 26.05.1983 in Heilbronn

Gutachter:
Prof. Dr. Dirk Hagemann (Betreuer)
Prof. Dr. Joachim Funke

Tag der Disputation:
12. Oktober 2011

Acknowledgments

First and foremost I want to thank my supervisor Prof. Dr. Dirk Hagemann for inspiring, teaching, and supporting me for the last three years. He was a great model in scientific and non-scientific issues. He always encouraged me to follow my own ideas and patiently taught me how to stand on my own academic feet. Without him, I would not be the “Forscher” I am. I also want to thank Prof. Dr. Joachim Funke, for his insights and inputs and for helping to get the project started in the first place.

During the last three years, I spent many hours on this thesis and I particularly want to thank my colleague, “roomie”, and friend Marieke Hager, who made our office feel like home. I am also indebted to my colleagues Dr. Andrea Schankin, Sascha Wüstenberg, Daniel V. Holt, and Markus Nagler for supporting me both inside and outside the office, Marianne Beschorner for her generous help in all administrative matters, and Katharina Weskamp, Andreas Neubauer, and Anna-Lena Schubert for their kind and very reliable help. Likewise, I thank all participants who took part in the studies.

I also want to express my deep gratitude to those who were not primarily involved in this thesis. I want to thank my parents Ursula and Michael Danner for always supporting me in finding and following my way. I deeply thank Sarah for supporting me with her love and her encouragement in everything I do.

Contents

1	Introduction	5
2	Implicit learning	8
3	Dynamic decision making.....	10
4	Some psychometric considerations	12
5	The measurement of psychometric intelligence.....	16
6	The measurement of implicit learning I.....	16
6.1	Measuring individual differences in implicit learning with an artificial grammar learning task (Manuscript 1)	18
6.2	Can artificial grammar learning tasks measure individual differences in implicit learning? (Manuscript 2)	20
6.3	The measurement of implicit learning II.....	21
7	The measurement of dynamic decision making.....	22
7.1	Measuring performance in dynamic decision making: reliability and validity of the Tailorshop simulation (Manuscript 3)	25
8	The measurement of success in real life	25
9	The psychometric properties of implicit learning and dynamic decision making (reported in Manuscript 4).....	26
10	The relation between psychometric intelligence, implicit learning, and dynamic decision making (reported in Manuscript 4).....	29
11	The relation between implicit learning, dynamic decision making, and success in real life (reported in Manuscript 4)	32
12	Summary and Conclusion	33
	References	35

Appendix

Manuscript 1: Danner, D., Hagemann, D., Schankin, A., Hager, M., & Funke, J. (under review). Measuring individual differences in implicit learning with an artificial grammar learning task. <i>Consciousness & Cognition</i>	A1
Manuscript 2: Danner, D., Hagemann, D., Schankin, A., Hager, M., & Funke, J. (under review). Can artificial grammar learning tasks measure individual differences in implicit learning? <i>Journal of Individual Differences</i>	A2
Manuscript 3: Danner, D., Hagemann, D., Holt, D. V., Hager, M., Schankin, A., Wüstenberg, S., & Funke, J. (in press). Measuring performance in dynamic decision making: reliability and validity of the Tailorshop simulation. <i>Journal of Individual Differences</i> . doi: 10.1027/1614-0001/a000055.	A3
Manuscript 4: Danner, D., Hagemann, D., Schankin, A., Hager, M., & Funke, J. (in press). Beyond IQ. A latent state trait analysis of general intelligence, dynamic decision making, and implicit learning. <i>Intelligence</i> . doi: 10.1016/j.intell.2011.06.004.	A4

1 Introduction

If you ask a psychologist to assess a person's cognitive ability, he or she will most probably apply an intelligence test and make a statement based on the person's intelligence quotient (IQ). And there are good reasons for this since IQ is the standard measure for cognitive ability and has a long research tradition in psychological science.

The foundation for the research of individual differences in cognitive abilities lies in the 1880s. Galton (1883) examined participants in his anthropometric laboratory with a test battery containing perceptual discrimination tests, memory tests, and association tests. Cattell and colleagues (Cattell & Farrand, 1896) continued Galton's work and developed a battery of mental tests for the selection of university applicants. This battery contained reaction time tests, perceptual discrimination tests, and memory tests. In a similar fashion, Münsterberg (1891) developed mental tests for the measurement of verbal associations, calculating ability, reading ability, and memory ability. These early approaches to the measurement of mental or cognitive ability consisted of sensory perception tests and rather simple cognitive tasks. The performance in these tests only correlated moderately (Sharp, 1899; Wissler, 1901) which means that a person who showed an above average performance in one task did not necessarily show an above average performance in another task. This started the discussion about the structure of cognitive abilities.

Spearman (1904) was one of the first who formulated a model of the structure of mental or cognitive abilities. He tested school children and reported that their performance in various sensory discrimination tasks was positively correlated. Based on this finding, he concluded that there is a *general mental ability factor* that he called g-factor that determines the performance in all mental tasks. This laid the foundation for characterizing cognitive ability with a single score which has later been called IQ (Stern, 1911). Spearman also suggested that there are (task-)specific ability factors that are independent of g. However, Spearman's test battery contained sensory discrimination tasks only and therefore, the generalizability of his model may be limited. His work nonetheless laid the foundation for further structural models of cognitive ability. Burt (1949) and Vernon (1950) used a broader range of cognitive tasks (e.g., memory tests, association tests, arithmetical tests, and spatial tests) and refined Spearman's model. Like Spearman, they suggested that there is a general cognitive ability factor. However, they further proposed that the specific ability factors do overlap and that this overlap can be explained by more general group factors. For example, Vernon (1950) suggested that the performance in cognitive tasks is determined by task specific ability factors. These tasks specific factors do overlap and therefore can be grouped

into minor group factors like mathematical ability, reading, spelling, or spatial ability. Likewise, the minor group factors can be combined into more general major group factors like a verbal-educational factor or an inductive factor. In turn, the overlap between the major group factors can be explained by a single g-factor that corresponds to Spearman's general mental ability factor.

The idea of a hierarchical structure of cognitive abilities was also proposed by Cattell (1963). He suggested that the performance in cognitive tasks is determined by specific first order factors like figural relations, memory span, or induction. In turn, these first order factors can be grouped into second order factors that he called fluid and crystallized intelligence. According to Cattell, fluid intelligence is the ability to adapt to and solve new problems whereas crystallized intelligence is the product of learning and prior experience. Initially, Cattell suggested that fluid intelligence and crystallized intelligence are two independent factors at the top of his ability model. However, empirical investigations (e.g. Horn & Cattell, 1966) have shown that there is an overlap between these second order factors and hence Cattell suggested that this overlap may be explained by a general, third order factor that may be seen as equivalent to Spearman's g-factor.

Considering the various structural models of cognitive abilities and the various measurement methods involved, Jäger (1984) systematized the available tasks that have been used to measure cognitive ability. He suggested that these tasks can be classified by the cognitive operations that are necessary to solve the tasks and the task's contents. According to Jäger, the cognitive operations are speed of operation, memory, creativity, and processing capacity and the contents can be figural, verbal, or numeric. In Jäger's terms, a participant solves a numerical series task by applying his operational processing to numeric content. Likewise, a participant solves a number-digit test by applying his speed of operation to verbal content. Factor analyses (e.g., Jäger, 1982) revealed that individual performance differences can be explained by the four operational factors and by the three content factors. Jäger's investigations further suggested the existence of a general ability factor.

Maybe the most comprehensive structural intelligence model is Carroll's (1993) three strata theory of intelligence. Based on reanalyses of over 460 factor analytic studies, he suggested a hierarchical model of cognitive ability. On the lowest level (stratum 1) there are 64 different specific ability factors like reading comprehension, memory span, general sound discrimination, numerical facility, or simple reaction time. According to Carroll, these specific abilities are correlated and therefore, may be grouped into eight general ability factors (stratum 2) which are fluid intelligence, crystallized intelligence, general memory and

learning, broad visual perception, broad auditory perception, broad retrieval ability, broad cognitive speediness, and processing speed. On the top of the hierarchy (stratum 3) there is one general ability factor that explains the correlations between the stratum 2 factors.

In summary, the majority of structural models of cognitive ability suggest a hierarchical structure of cognitive abilities with one single ability factor at the top of the hierarchy. This general ability factor or g-factor may be seen as a disposition to be successful in various situations or tasks and was described as the ability to be successful in a culture (Hofstätter, 1957), the ability to act purposeful and to think reasonable (Wechsler, 1975), or the ability to understand complex information, to think deductive, and learn from experience (Neisser et al, 1996). From a statistical point of view, the g-factor may be seen as the proportion of individual differences that is consistent across very different cognitive tasks. Sternberg and Gigorenko (2002) say that g is able to explain about 50% of the performance variance in very different cognitive tasks. Furthermore, g has been shown to be the most powerful predictor of educational attainment and professional success (e.g., Ng, Eby, Sorensen, & Feldman, 2005; Salgado et al, 2003; Schmidt & Hunter, 2004). This underlines the significance and relevance of IQ as a measure of a single general cognitive ability factor.

The different hierarchical models of cognitive abilities have in common that there is one general ability factor at the top of their structure. However, they vary in the number and width of specific ability factors. Early investigations used quite homogeneous tasks to investigate cognitive ability, which led to rather simple structural models (Spearman, 1904). Subsequent investigations used a much wider range of tasks (e.g., Carroll, 1993; Cattell, 1963; Jäger, 1984; Vernon, 1950), which led to more comprehensive and more fine-grained structural models. These hierarchical models may be seen as a framework, in which different ability constructs can be integrated. Such a systematization of abilities or tasks may explain the relation between different models of cognitive ability. In particular, Carroll's three strata theory offers a very comprehensive framework to integrate various ability components. For example, the stratum 2 factor visual perception of Carroll's model may be seen as an equivalent to the grouping factor spatial ability in Vernon's model or the spatial content factor in Jäger's model. Likewise, memory span is a stratum 1 factor in Carroll's model as well as a level 1 ability factor in Cattell's model. Similarly, fluid intelligence is a stratum 2 factor in Carroll's model as well as an element of Cattell's model (as a second order factor) or Vernon's model (as the major group factor induction). In a similar vein, Jäger's Berlin intelligence model allows one to classify cognitive tasks by the contents of the tasks or by the cognitive operations that are used. For example, Cattell's Culture Fair Intelligence Test which

was developed as an indicator of fluid intelligence may be described as a product of figural content and processing capacity in Jäger's model.

However, there are also ability factors that have not been integrated into these hierarchical models so far. These ability factors may characterize cognitive ability beyond IQ. The present thesis investigated two such constructs which were implicit learning and dynamic decision making. The usefulness of these constructs was evaluated in three ways. First, I evaluated the *incremental construct validity* of implicit learning and dynamic decision making. In particular, I evaluated whether implicit learning and dynamic decision making are divergent from measures of psychometric intelligence and I evaluated how they fit into hierarchical models of cognitive ability. Second, I evaluated the *predictive validity* of these constructs. In particular, I investigated whether implicit learning and dynamic decision making can incrementally predict success in real life. Third, I evaluated the *psychometric properties* of the measures of implicit learning and dynamic decision making. In particular, I investigated whether these measures are reliable, stable over time, and consistent across different tasks. The results of these investigations are reported in four manuscripts:

Manuscript 1: Danner, D., Hagemann, D., Schankin, A., Hager, M., & Funke, J. (under review). Measuring individual differences in implicit learning with an artificial grammar learning task. *Consciousness & Cognition*.

Manuscript 2: Danner, D., Hagemann, D., Schankin, A., Hager, M., & Funke, J. (under review). Can artificial grammar learning tasks measure individual differences in implicit learning? *Journal of Individual Differences*.

Manuscript 3: Danner, D., Hagemann, D., Holt, D. V., Hager, M., Schankin, A., Wüstenberg, S., & Funke, J. (in press). Measuring performance in dynamic decision making: reliability and validity of the Tailorshop simulation. *Journal of Individual Differences*. doi: 10.1027/1614-0001/a000055.

Manuscript 4: Danner, D., Hagemann, D., Schankin, A., Hager, M., & Funke, J. (in press). Beyond IQ. A latent state trait analysis of general intelligence, dynamic decision making, and implicit learning. *Intelligence*. doi: 10.1016/j.intell.2011.06.004.

2 Implicit learning

Implicit learning is most often defined as the ability to learn without being aware that something is learned. For example, Shanks and St. John (1994) suggest that "implicit learning occurs without concurrent awareness of what is being learned" (p. 369). Some authors refer to

the cognitive processes that take place. Mackintosh (1998) describes implicit learning as “the product of a basic associative system” (p. 365). Likewise, Mathews et al. (1989) characterize implicit learning as “an alternate mode of learning that is automatic, non-conscious, and more powerful than explicit thinking” (p. 1083). Other authors also refer to the kind of knowledge that is acquired. For example, Reber (1993) describes implicit learning as “largely independent of conscious attempts to learn and largely in the absence of explicit knowledge about what was acquired” (p.5). In essence, most definitions contain two core aspects. First, implicit learning is unintended or even unconscious. Second, the acquired knowledge can not be reported.

Considering implicit learning as a cognitive ability raises the question how implicit learning is related to other ability constructs. In particular, it is a theoretically interesting question, whether implicit learning is an ability which is independent of psychometric intelligence or whether implicit learning can be integrated into a hierarchical model of intelligence. Mackintosh (1998) hypothesizes that implicit learning is independent of psychometric intelligence. According to him, there are two independent learning systems: an implicit, associative learning system and an explicit, hypothesis generating and testing system. He suggests that the explicit learning system is necessary for discovering regularities with intention and awareness (e.g., in a numerical series task). The implicit learning system, on the other hand, detects contingencies without awareness or intention (e.g., judging whether a sentence is grammatically correct without being able to report the respective grammatical rule). Mackintosh criticizes that standard intelligence tests capture individual differences in the explicit system but not individual differences in the implicit learning system. He proposes that implicit learning is independent from psychometric intelligence but nevertheless a determinant of success in real life. There are several findings that support Mackintosh's position. Several studies report low and non-significant correlations between the performance on intelligence tests and the performance on implicit learning tasks (Gebauer & Mackintosh, 2007; Feldman, Kerr, & Sreissguth, 1995; Kaufman et al., 2010; McGeorge, Crawford, & Kelly, 1997; Pretz, Totz, & Kaufman, in press; Reber, Walkenfeld, & Hernstadt, 1991). In addition, there are performance differences in several domains that can not be explained by psychometric intelligence but may be explained by the ability to learn rules implicitly. For example, Ceci and Liker (1986) investigated the performance in horse-racing bets. They reported that individual differences in betting performance could neither be explained by individual differences in reported knowledge nor by individual differences in psychometric intelligence. Comparing successful and unsuccessful bettors, Ceci and Liker found that both

used the same variables to make predictions (e.g., whether a horse has won the last race, the condition of the track, or a horse's lifetime). However, the successful betters used more complex interactions between variables to make predictions (e.g., whether a horse has won the last race on a specific track against a specific rival). Some authors (e.g., Mackintosh, 1998) suggest that these complex interactions may represent implicitly learned rules and the successful betters may be more successful in implicit learning. In a similar vein, Berry and Broadbent (1984) developed a task which they called Process Control. In that task, the participants have to control an outcome variable (e.g., amount of sugar produced in a factory) by manipulating an input variable (e.g., number of workers hired). Typically, the participants are not able to report how the input variable and the outcome variable are connected but there are individual performance differences. The performance differences are independent of psychometric intelligence (e.g., Berry & Broadbent, 1984; Gebauer & Mackintosh, 2007) and several authors suggested that the participants may have learned the connection between the variables implicitly (Berry, Broadbent, 1984; Buchner, Funke, & Berry, 1995; Mackintosh, 1998).

Taken together, these findings suggest that implicit learning may be independent of psychometric intelligence. Furthermore, there are individual performance differences in some cognitive tasks that can not be explained by intelligence but that fit conceptually well with implicit learning. Hence, implicit learning may be an interesting ability construct to describe and understand human cognitive ability beyond IQ.

3 Dynamic decision making

Any cognitive task can be seen as a problem that has to be solved: there is a given state (e.g., an unsolved item) that has to be transferred into a goal state (e.g., a solved item) whereby a barrier has to be overcome (e.g., find a rule). Dörner (1980, 1986) criticizes that standard intelligence tests only measure the speed and accuracy of the ability to solve simple problems (like an analogical reasoning task) but not the ability to solve complex problems in real life (like managing a company). Dörner suggests that real life problems are characterized by complexity, connectivity, non-transparency, dynamics, and polythely. For example, an analogical reasoning task may be seen as a simple task because there is one default solution for a given item and the structure of the task is rather simple (e.g., London is to England as Berlin is to Germany because London is the capital of England and Berlin is the capital of Germany). On the other hand, managing a company may be seen as a much more complex task because it requires considering many variables like the financial situation of the

company, employee satisfaction, the demands of the market, and so on. Such a task may also be seen as connected since several variables are interdependent (like the demands of the market and the financial situation of the company). Furthermore, the task may be seen as non-transparent because not all information which is necessary to solve the task will be available all the time. The task may also be seen as dynamic because the variables (like the demands of the market) will change over time, and the task may be seen as polythelic because a problem solver may have to solve several subgoals (like satisfying the employees, optimizing the production, etc.) to reach the superior goal (manage the company successfully). Dörner's critique laid the foundation for a field of research, which has been called dynamic decision making (Gonzalez, Vanyukov, & Martin, 2005) or complex problem solving (Funke, 2010).

On a conceptual level, the relationship between dynamic decision making and psychometric intelligence is unresolved. On the one hand, dynamic decision making and psychometric intelligence may be seen as different because they are operationalized differently. In particular, Dörner suggested that the demands of items in an intelligence test differ from the demands of complex problems (e.g., in terms of complexity or dynamics). On the other hand, dynamic decision making and psychometric intelligence may be seen as similar because both ability constructs are defined in a similar way. Neisser et al. (1996) described intelligence as the ability to understand complex information, to think deductively, and learn out of experience. This agrees with Dörner's description of complex problem solving. Furthermore, Hofstätter (1957) suggested that intelligence is the ability to be successful in a culture and Dörner (1980) suggested that complex problems are valid representations of real-world problems. Accordingly, there should be a substantial overlap between dynamic decision making performance and psychometric intelligence. In line with that, some authors even describe intelligence as the ability to solve problems (e.g., Berg & Sternberg, 1985).

Beyond similarities in their definitions psychometric intelligence and dynamic decision making have also been described as involving similar cognitive processes. In particular, Dörner (1986) suggested that making dynamic decisions requires gathering information, elaborating goals, planning decisions, and self-management. For example, in order to manage a company, the problem solver has to identify the relevant information (e.g., demands of the market, current production status), set objectives (e.g., increase production), make plans (e.g., hire more workers and buy new machines in order to increase production), and so on. In a similar vein, Funke (2010) suggests that dynamic decision making requires complex cognition, which means actively searching for information with the intention to

make decisions or to solve problems (see also Knauff & Wolf, 2010). Such a description of the problem solving process agrees with Sternberg's (1977) analyses about what cognitive processes are involved in solving items of an intelligence test. In particular, Sternberg suggested that solving inductive reasoning items of an intelligence test requires encoding, inference, mapping, application, justification, and responding. In summary, despite obvious and considerable differences in the tasks used to measure them, dynamic decision making and psychometric intelligence also have a lot in common: Conceptually, both constructs are described in similar terms and some authors suggest that similar cognitive processes are involved when solving complex problems and when solving items of traditional intelligence tests.

However, apart from the constructs' relation on a theoretical level, it may be even more interesting to know how dynamic decision making and psychometric intelligence are related on an empirical level. In particular, it is interesting to know whether dynamic decision making is an ability that is independent of psychometric intelligence and how dynamic decision making fits into a hierarchical model of cognitive abilities. Previous studies found non-significant or only small correlations (for an overview, see Kluwe, Misiak, & Haider, 1991), other studies report significant standardized path coefficients between $\beta = .38$ and $\beta = .54$ from latent intelligence to latent dynamic decision making variables (Kröner, Plass, & Leutner, 2005; Rigas, Carling, & Brehmer, 2002; Wittmann & Hatstrup, 2004). One study even found a correlation between a latent intelligence and a latent dynamic decision making variable of $r = .84$ (Wirth & Klieme, 2003). Given these heterogeneous findings, it is undecided whether dynamic decision making is a facet of intelligence or an independent ability construct. The present thesis will help to clarify this issue.

4 Some psychometric considerations

Several studies reported small and non-significant correlations between implicit learning variables and psychometric intelligence variables. These findings were interpreted as preliminary evidence for the independence of implicit learning and psychometric intelligence. That conclusion may not be warranted. In particular, because these studies treated the performance measures as trait-like variables which are stable over time and consistent across different situations or methods (Stemmler, Hagemann, Amelang, & Bartussek, 2011). This might be inappropriate because the variance of a performance measure may capture additional factors beyond individual differences in a trait which in turn might affect the correlation with other variables.

First, a performance measure may also be influenced by the specific measurement situation even in standardized experiments. For example, one person may be well rested whereas another person may already have worked several hours before testing. One person may be motivated to show maximum performance whereas another person may have gotten a stinging rebuke by his or her supervisor that day and may not be motivated to perform well. This means, that performance in an implicit learning task may not only reflect individual ability differences but also individual situational effects. This may decrease the correlation between an implicit learning variable and an intelligence variable and thus the correlation may not reflect the relation between the ability constructs. Likewise, dynamic decision making variables may reflect occasion specific variance which may affect the correlation with psychometric intelligence variables. In particular, a dynamic decision making variable with a small proportion of occasion specific variance may reveal a substantial relation with an intelligence variable, whereas a dynamic decision making variable with a substantial proportion of occasion specific variance may reveal a small correlation with an intelligence variable.

Second, a performance measure may be influenced by the specific method being used. Hence, there may be individual differences in a performance measurement which are triggered by the method. For example, a verbal intelligence test may capture individual differences in general intelligence as well as individual differences in speech comprehension whereas a figural intelligence test may capture individual differences in general intelligence and visual thinking. Thus, individual differences in speech comprehension or visual thinking are method specific because they can only be assessed with verbal or figural test material. Similarly, a particular implicit learning task may measure performance differences, which are specific to this particular task but not to implicit learning ability in general. Thus, method specificity may be an additional factor that decreases the correlation between psychometric intelligence variables and implicit learning variables. The same applies to dynamic decision making variables. A particular dynamic decision making task may not only reflect individual differences in dynamic decision making ability but also individual knowledge differences (as suggested by Hesse, 1982). A variable with a small proportion of method specific variance may reveal substantial correlations with psychometric intelligence variables whereas a variable with a substantial proportion of method specific variance may reveal small correlations.

Third, a performance measure may be influenced by unsystematic measurement error. For example, instructions may be ambiguous or persons may accidentally make mistakes, which may result in a low reliability of performance measures. These effects may contribute unwanted variance in implicit learning variables and dynamic decision making variables and hence decrease the correlation with other variables. Therefore, it seems worthwhile to investigate the reliability of implicit learning variables in greater detail.

In essence, the occasion specificity, the method specificity, and the reliability of variables may affect the correlation with other variables. Therefore, these factors must be taken into account when investigating the relations between these constructs. The present work investigates these effects which will help to understand the validity of implicit learning and dynamic decision making in greater detail.

The consideration that a variable may reflect trait variance as well as occasion specific, method specific, and unsystematic variance has been formalized in Steyer and colleagues' latent state-trait theory (Steyer, Schmitt, & Eid, 1999). In a nutshell, latent state-trait theory proposes that the measurement i of a variable Y can be decomposed into a trait ξ_i , a state residual ζ_i , a method residual η_i , and an unsystematic error residual ε_i , thus $Y_i = \xi_i + \zeta_i + \eta_i + \varepsilon_i$. Given the independence of these factors (Steyer et al., 1999), the variance of this measurement can be decomposed as $\sigma^2(Y_i) = \sigma^2(\xi_i) + \sigma^2(\zeta_i) + \sigma^2(\eta_i) + \sigma^2(\varepsilon_i)$, and the factor variances may be estimated with a structural equation model as shown in Figure 1. As can be seen in this figure, the latent trait factor is defined as a variable that is consistent across several measurement occasions and methods, whereas the latent state residual and the method factor are specific to the individual measurement occasion or the assessment method. Hence, these models allow separating the different contributions of the trait, the measurement occasion, and the measurement method to the manifest variables.

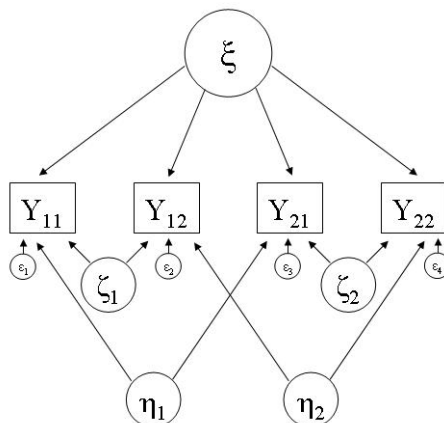


Figure 1. Latent state-trait structural equation model. Y_{11} = variable on measurement occasion 1 with method 1, Y_{12} = variable on measurement occasion 1 with method 2, Y_{21} = variable on measurement occasion 2 with method 1, Y_{22} = variable on measurement occasion 2 with method 2, ξ = trait variable, ζ_1 = state residual 1, ζ_2 = state residual 2, η_1 = method residual 1, η_2 = method residual 2, ε_1 - ε_4 = measurement error.

There have been many applications of latent state-trait models in different domains of personality research, which demonstrated substantial effects of the measurement occasion or the method on behavioral variables (e.g., Eid, Notz, Steyer, & Schwenkmezger, 1994; Schmitt & Steyer, 1993; Steyer, Schwenkmezger, Auer, 1990; Yasuda, Lawrenz, Whitklock, Lubin, & Lei, 2004; Ziegler, Ehrlenspiel, & Brand, 2009) or physiological variables (e.g., Hermes et al., 2009; Hagemann, Hewig, Seifert, Naumann, & Bartussek, 2005). However, there have been no applications of latent state-trait models on performance variables yet, even if some findings suggest that it may be instructive to consider the occasion specificity and method specificity of these variables. For example, in some studies the participants completed the same dynamic decision making task several times (Süß, Kersting, & Oberauer, 1993; Wittmann & Hattrup, 2004) and the performance between subsequent tasks correlated only moderately (between $r = .37$ and $r = .62$). This points either towards a low reliability or towards a substantial occasion specificity of the variables. Moreover, Wirth and Klieme (2003) reported structural equation models, which implied a correlation of $r = .33$ between two dynamic decision making tasks ($r = .47$ when corrected for attenuation) and Gebauer and Mackintosh (2007) reported a correlation of $r = .15$ between two artificial grammar learning tasks ($r = .21$ when corrected for attenuation). This suggests a substantial method specificity of the performance measures.

Taken together, there are some findings which suggest that implicit learning variables and dynamic decision making variables may contain substantial proportions of occasion specific or method specific variance. These unwanted variance proportions may affect correlations with other variables such as psychometric intelligence and thus these correlations may be biased estimates for the relation between constructs. Therefore, one aim of the present thesis was to investigate implicit learning variables and dynamic decision making variables with latent state-trait models. Thus, these unwanted variance proportions can be controlled and the relation between constructs can be estimated without bias.

As can be seen in Figure 1, a construct has to be measured with at least two different methods on at least two different measurement occasions in order to apply latent state-trait models. For the purpose of the present thesis, I therefore ran a *longitudinal study* and measured psychometric intelligence, implicit learning, and dynamic decision making with two different methods on two different measurement occasions. In addition, I measured several indicators of real life performance to investigate whether implicit learning and dynamic decision making are determinants of success in real life as suggested by Mackintosh (1998) and Dörner (1986).

5 The measurement of psychometric intelligence

Carroll (1993) has shown that the Advanced Progressive Matrices (APM; Raven, Court, & Raven, 1994) are an excellent marker for psychometric intelligence. Therefore, I selected the APM as a first indicator for psychometric intelligence. The Berlin Intelligence Structure Test (BIS; Jäger, Süß, & Beauducel, 1997) was used as a second indicator for psychometric intelligence. The BIS was used because Jäger (1973) carefully selected the tasks that he included in the BIS. In particular, he reviewed and systematized 289 different tasks in order to obtain a representative sample of available intelligence tasks. Thus, the performance in the BIS may be seen as a further valid indicator for psychometric intelligence.

6 The measurement of implicit learning I

While there are many investigations on how to measure individual differences in psychometric intelligence, there is a paucity of investigations on how to measure individual differences in implicit learning. However, there are tasks that have been used to investigate implicit learning processes and such tasks may also be suitable to investigate individual differences in implicit learning. In particular, artificial grammar learning tasks have become the standard paradigm to investigate implicit learning (e.g., Altmann, Dienes, & Goode, 1995;

Dulany, Carlson, & Dewey, 1984; Gebauer & Mackintosh, 2007; Knowlton & Squire, 1994, 1996; Meuleman & Van der Linden, 2003; Perruchet & Pacteau, 1990; Pothos & Bailey, 2000; Reber, 1967; Reber et al., 1991; Reber & Perruchet, 2003; Scott & Dienes, 2010; Tunney, 2005). In such a task, the participants are asked to learn a list of arbitrary letter strings (like WNSNXS). Afterwards they are told that these strings were constructed according to a complex rule system or a grammar (see Figure 2 for an example) and they are asked to judge new strings (like WNSNXT) as grammatical or non-grammatical. Typically, the participants' judgment accuracy is above chance, which suggests that they learned something but they are not able to report the grammar rules, which suggests that they learned the rules implicitly. This operationalisation agrees with definitions of the implicit learning process. When the participants are asked to learn the letter strings, they do not know that these letter strings are constructed according to a grammar. Thus, they are not able to learn the grammar intentionally or consciously. In addition, they are not able to report the grammar rules. Accordingly, the judgment accuracy in the testing phase of an artificial grammar learning task may be used as a valid performance indicator for implicit learning.

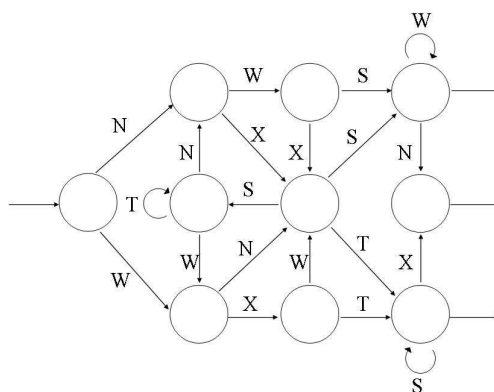


Figure 2. Example of a grammar that is used in artificial grammar learning tasks. Grammatical Stimuli are generated by following any path of arrows (e.g., NWSW).

However, this approach may be limited, in particular, if participants complete an artificial grammar learning task more than once. Participants who complete an artificial grammar learning task for the first time do not know that the letter strings in the learning phase are constructed according to a grammar and thus will not search for grammar rules. However, participants who complete an artificial grammar learning task for the second time will know that the letter strings in the learning phase are constructed according to a grammar

and thus they *may* intentionally search for the grammar rules. This would violate the definition of implicit learning. Therefore, the learning during the second artificial grammar task may not be implicit any more and the performance in the second artificial grammar learning task *may* not be a valid indicator for implicit learning performance. The application of latent state-trait models requires the participants to complete several artificial grammar learning tasks several times. Therefore, I first had to investigate whether artificial grammar learning tasks can be used more than once for measuring individual differences in implicit learning.

6.1 Measuring individual differences in implicit learning with an artificial grammar learning task (Manuscript 1)

In order to use an artificial grammar learning task more than once, Gebauer and Mackintosh (2007) suggested modifying the standard procedure of artificial grammar learning tasks. In particular, they asked their participants to learn a list of grammatical strings in a learning phase. Afterwards, in a testing phase, they did *not* inform the participants that the strings were constructed according to a grammar but they asked their participants to rate the presented letter strings as “old” (already presented in the learning phase) or “new” (not presented before). Indeed all presented strings were new, but half of them were grammatical and the other half was not. Grammatical strings rated as “old” and non-grammatical strings rated as “new” were counted as correct answers. The idea behind this procedure may be that the participants learned something about the grammar, thus felt familiar with a grammatical string and this is why they classified a grammatical string as an “old” one.

In line with this reasoning, there are several authors who suggest that novelty judgments and grammaticality judgments are conceptually similar. For example, Whittlesea and Lobe (2000) demonstrated that several heuristics (fluency, generation, and resemblance) influence the performance in recognition as well as classification tasks. The authors suggest that these heuristics affect the perceived familiarity of stimuli and that familiarity affects novelty judgments as well as grammaticality judgments (see also Kinder, Shanks, Cock, & Tunney, 2003; Scott and Dienes, 2008; Whittlesea, Jacoby, & Girard, 1990). However, as noted by Whittlesea and Lobe (2000) “that does not mean that classification and recognition decisions that are performed heuristically will ordinarily be correlated” (p. 101). Therefore, one aim of this study was to test whether asking subjects for novelty measures the same construct as asking for grammaticality.

Another aim of this study was to investigate whether the performance in an artificial grammar learning task is independent of reportable grammar knowledge when several artificial grammar learning tasks are completed. Therefore, I developed a bi- and trigram knowledge test. The bi- and trigram knowledge test was developed with reference to Perruchet and Pacteau (1990) who suggest that the participants may not use abstract grammar knowledge to make grammaticality decisions but heuristics like bigrams. In a similar fashion, other authors suggest that the participants may use fragments (Dulany, Carlson, & Dewey, 1984) or chunks (Servan-Schreiber & Anderson, 1990). Therefore, I asked the participants to rate whether a bi- or tri-gram occurred more often in grammatical or more often in non-grammatical strings. A zero correlation between n-gram knowledge and accuracy would indicate that the participants did not use bi- or trigrams for their judgments, whereas a positive correlation between n-gram knowledge and accuracy would indicate that the participants may have used bi- or trigrams for their judgments.

I performed a series of experiments, which manipulated whether the participants had to rate the grammaticality of strings in the testing phase (“classical” procedure) or the novelty of strings in the testing phase (modified procedure). There were three central findings of these experiments. First, the reliability estimates of the judgment accuracy variables were rather small (between 0.00 and 0.66). This replicates the findings of Gebauer and Mackintosh (2007) and Reber et al. (1991) who also reported small reliability estimates for the performance in artificial grammar learning tasks. Second, the instruction to rate the novelty of letter strings does not allow one to measure the same construct as the instruction to rate the grammaticality of letter strings. This means that even if both instructions may be seen as similar on a conceptual level (Kinder, Shanks, Cock, & Tunney, 2003; Whittlesea, Jacoby, & Girard, 1989; Whittlesea and Lobe, 2000), they differ substantially on an empirical level. Therefore, novelty judgments are not equivalent to grammaticality judgments. Third, if participants complete a “classical” artificial grammar learning task for the first time, there is a zero correlation between the judgment accuracy and the amount of reportable grammar knowledge. However, if participants complete a “classical” artificial grammar learning task for the second time, there is a substantial correlation between the judgment accuracy and the amount of reportable grammar knowledge. Furthermore, the performance in a first artificial grammar learning task does not significantly correlate with the performance in following artificial grammar learning tasks but the performance in a second artificial grammar learning task correlates significantly with the performance in a third artificial grammar learning task. This suggests that the performance in a first artificial grammar learning task may be seen as an

indicator of implicit learning whereas the performance in subsequent artificial grammar learning tasks may not be seen as implicit any more.

The findings of this first study suggest that artificial grammar learning tasks may only be used once to measure individual differences in implicit learning. However, there is also an alternative interpretation. Our participants completed a knowledge test (containing bi- and trigrams of letter strings) after every artificial grammar learning task. Therefore, it is also possible that the knowledge test and not the grammar awareness changed the participants' strategy and caused the low task consistency as well as the relation with reported knowledge. Investigating this hypothesis was the aim of a second study.

6.2 Can artificial grammar learning tasks measure individual differences in implicit learning? (Manuscript 2)

The initial aim of this study was to investigate whether a knowledge test increases the correlation between two successively completed artificial grammar learning tasks. Therefore, half the participants completed a bigram knowledge test after the first artificial grammar learning task (the bigram group) whereas the other half did not (the control group). A first result was that the correlation between both artificial grammar learning tasks was smaller in the bigram group. Likewise, there was a significant correlation between the performance in the second artificial grammar learning tasks and reported bigram knowledge in the bigram group, but not in the control group. These results suggest that a bigram knowledge test decreases the task consistency of artificial grammar learning tasks and increases the correlation between implicit learning performance and reportable grammar knowledge. This means, artificial grammar learning tasks may only be used once to measure individual differences in implicit learning if the participants complete a bigram knowledge test. However, artificial grammar learning tasks may be used for several times if the participants do not complete a bigram knowledge test. Therefore, participants can complete several artificial grammar learning tasks for several times and latent state-trait models can be used to estimate a latent implicit learning trait variable.

There were some further findings of this study. The reliability estimates of artificial grammar learning performance were rather small (between 0.21 and 0.60). This replicates previous findings (Gebauer & Mackintosh, 2007; Reber et al., 1991) and suggests that the performance in artificial grammar learning tasks is substantially affected by unsystematic measurement error.

In addition, the participants in this study also completed Cattell's Culture Fair Intelligence Test and they were asked to report their final school exams' grade point average. Similar to previous studies (Gebauer & Mackintosh, 2007; Kaufman et al., 2010; McGeorge et al., 1997; Pretz et al., in press; Reber et al., 1991) there was only a moderate correlation between implicit learning performance and psychometric intelligence. Furthermore, the results of this study revealed a significant relation between participants' final school exams and artificial grammar learning performance. This is in line with Kaufman et al. (2010) who also showed a significant association between implicit learning performance and educational success. However, the association, observed in the present study, became non-significant when intelligence was included as a further predictor. This suggests that even though the implicit learning variable and the psychometric intelligence variable only correlated moderately, the relation between artificial grammar learning performance and educational success was due to this overlap.

6.3 The measurement of implicit learning II

Taken together, the central finding of both studies is that artificial grammar learning tasks can be used several times to measure individual differences in implicit learning. The present findings suggest that bigram knowledge tests may turn the participants' attention towards bigrams. Thus, the participants may intentionally acquire bigram knowledge in a subsequent artificial grammar learning task and learning may not be implicit any more. However, if no bigram knowledge test is completed, several artificial grammar learning tasks can be used to measure individual differences in implicit learning. Therefore, I used two different artificial grammar learning tasks (without bigram knowledge tests) to measure individual differences in implicit learning.

Furthermore, both studies revealed small reliability estimates of the implicit learning performance variable. This replicates the findings of Gebauer and Mackintosh (2007) and Reber et al. (1991). As discussed, a small reliability has implications for the investigation of the relation between implicit learning and psychometric intelligence. In particular, a small reliability decreases the correlation between two variables. Hence, to investigate the relation between implicit learning and psychometric intelligence, I used latent state-trait models to control for this lack of reliability. They decompose the variances of the manifest performance variables into a trait proportion, a state residual proportion, a method residual proportion, and a measurement error proportion.

Finally, there is preliminary evidence for a relation between implicit learning and the real life criterion educational success. In the present investigation, this relation became non-significant when psychometric intelligence was included as a predictor. Nevertheless, this finding suggests that it may be worthwhile to investigate the relation between implicit learning and real life criteria in greater detail. Therefore, I used different indicators of real life performance in order to investigate the relation between implicit learning and real life performance in the longitudinal study.

7 The measurement of dynamic decision making

Traditional paper-pencil intelligence tests have been criticized as inadequate methods to measure dynamic decision making (Dörner, 1980, 1986). Therefore, several authors suggest using computer-based simulations to measure dynamic decision performance. Over the years, several dynamic decision making tasks have been developed. For example, the Tailorshop scenario (Dörner, 1979; Funke, 1983) simulates a fictional company where the participants have to control several variables like the number of workers or the costs for advertising in order to maximize their company's value. Other tasks simulate a forestry (Wagener, 2001), a power plant (Wallach, 1998), or a space flight (Wirth & Funke, 2005) where the participants have to control several variables to reach a given goal state. These simulations have in common that they simulate complex, connected, dynamic, non-transparent, and sometimes even polythetic environments.

The *Heidelberg Finite State Automaton* has become a common instrument for measuring individual differences in dynamic decision making, especially since it has been included in the Program for International Student Assessment (PISA; Wirth & Klieme, 2003). Therefore, I chose this simulation as one indicator of dynamic decision making. The scenario simulates a space flight where the participants can control a space ship and a vehicle with a user interface (see Figure 3). During the simulation, the participants are asked to reach several goal states (e.g., landing the space ship on a particular planet) whereby the number of reached goal states is taken as a performance indicator for dynamic decision making (Wirth & Funke, 2005; Wirth & Klieme, 2003).

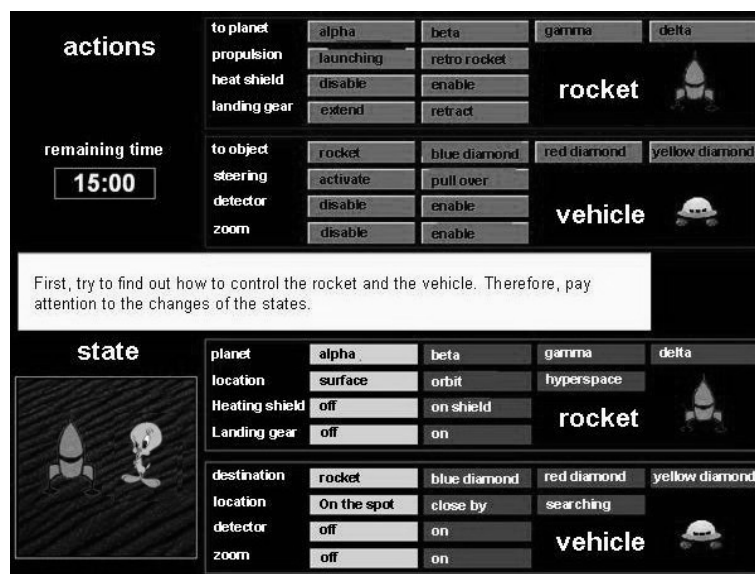


Figure 3. Screenshot of the graphical user interface of the Heidelberg Finite State Automaton (labels translated).

The simulation corresponds to Dörner's and Gonzalez's definition of dynamic decision making. In particular, the simulation may be seen as complex, because it consists of many variables (e.g., the state of the propulsion, the state of the landing gear). The simulation may be seen as connected because the different variables depend on each other (e.g., the ability to fly with the space ship depends on the state of the propulsion, the heat shield, the landing gear, and the state of the vehicle). The simulation may be seen as non-transparent because the participants do not know how the variables in the simulation are connected but have to find out while exploring and controlling. Likewise, the simulations may be seen as dynamic because each intervention in the simulation influences the following state of the simulation. Finally, the simulation may be seen as a polythetic task because it is necessary to achieve different subgoals (e.g., controlling the landing gear, the heating shield, the state of the vehicle) to achieve a greater goal (e.g., landing the space ship on a particular planet).

The *Tailorshop* is another well established dynamic decision making task that has been used for several decades (e.g., Barth & Funke, 2010; Leutner, 1988; Putz-Osterloh, 1981, 1983; Putz-Osterloh, Bott, & Köster, 1990; Putz-Osterloh & Lür, 1981; Süß, Kersting, & Oberauer, 1993; Wittmann & Hatrup, 2004). The scenario simulates a small business that produces and sells shirts. The participants have to manage this business for twelve simulated months by manipulating several variables like the number of workers, the expenses for advertising, etc. (see Figure 4).

Round 1 of 12

Variable	Value	Planning
Account status	165775	
Number of shirts sold	407	
Price of raw material	3.99	
Shirts in stock	81	
Workers 50	8	<input type="text"/>
Workers 100	0	<input type="text"/>
Salary	1080	<input type="text"/>
Price of shirts	52	<input type="text"/>
Shops	1	<input type="text"/>
Worker satisfaction %	57.7	
Loss of production %	0.0	

Variable	Value	Planning
Company value	250685	
Demand	767	
Raw material in stock	16	<input type="text"/>
Machines 50	10	<input type="text"/>
Machines 100	0	<input type="text"/>
Repair & service costs	1200	<input type="text"/>
Social costs per worker	50	<input type="text"/>
Advertising costs	2800	<input type="text"/>
Business location	suburb	<input type="text" value="suburb"/>
Machine damage %	5.9	

Figure 4. Screenshot of the graphical user interface of the Tailorshop (labels translated).

The Tailorshop was initially developed by Dörner (1979) according to his definition of complex problems. In particular, the simulation consists of many variables and connections. Therefore, the Tailorshop may be seen as complex. Furthermore, the variables are highly connected. For example, the ability to produce shirts in the Tailorshop simulation depends on the amount of raw material, the number of machines and workers, the workers' satisfaction, and the state of the machines. In addition, the simulation may be seen as non-transparent because the participants do not know how the variables in the simulation are connected but have to find out while exploring and controlling them. The Tailorshop may also be seen as dynamic because each intervention in the simulation influences the following state of the simulation. Finally, the Tailorshop may be seen as a polythetic task because it is necessary to achieve different subgoals (like buying raw material, hire workers, advertising, etc.) to achieve the greater goal (maximize the company value).

However, even if the simulation has become a standard paradigm to investigate dynamic decision making, there is a discussion which indicator should be used as a performance variable. Some authors suggest using the number of months with a positive trend in the company value to quantify dynamic decision making performance (e.g., Funke, 1983) whereas other authors suggest using the absolute company value at the end of the simulation to quantify the dynamic decision making success (e.g., Barth & Funke, 2010; Süß, Oberauer, & Kersting, 1993). In order to find an appropriate performance indicator for the Tailorshop simulation, I analyzed the data of the first measurement occasion of the longitudinal study.

7.1 Measuring performance in dynamic decision making: reliability and validity of the Tailorshop simulation (Manuscript 3)

In this analysis, I compared two different performance indicators of the Tailorshop simulation: the *change variable* and the *trend variable*. The change variable corresponds to the sum of the changes of the company value between the simulated months (which is equivalent to the final company value after twelve simulated months). The trend variable corresponds to the number of months with a positive trend in the company value. I used structural equation models to test measurement models and estimate the reliability of the performance variables. Furthermore, the validity of the performance variables was evaluated with respect to their correlation with the Heidelberg Finite State Automaton (convergent validity), their correlation with self-rated income and supervisor ratings (predictive validity), and their correlation with the performance in the Advances Progressive Matrices (divergent validity).

The analysis revealed that the measurement models fitted the trend variables well (in particular, the trends between the second month and the twelfth month) but not the change variables. Furthermore, the results revealed good reliability and good overall validity for the trend of the company value. Hence, I decided to use the number of months with a positive trend in the company value (between the second and the twelfth month) as a performance measure in the Tailorshop simulation.

8 The measurement of success in real life

For the purpose of the present study, I focused on a particular aspect of success in real life: professional success. For one thing, the predictive validity of psychometric intelligence has often been evaluated by its association with professional success (e.g., Ng et al., 2005; Salgado et al., 2003; Schmidt & Hunter, 2004). Thus, from a theoretical point of view, professional success is a useful criterion to evaluate the predictive validity of implicit learning and dynamic decision making variables. Second, professional success is an important outcome variable in an economic context. Thus, in an applied context, implicit learning or dynamic decision making may become interesting selection criteria for university or job applications if they are able to predict professional success.

Detle, Abele, and Renner (2004) systematized different indicators of professional success. They suggested that the different indicators may be distinguished by (1) their frame of reference (specific task vs. global career), (2) the type of data (e.g., neutral parameter or comparison with reference), and (3) the data source (document, self-rating, external rating).

For example, a participant's yearly income may be characterized as an indicator with global career as the frame of references. Income can further be seen as a neutral parameter because it can be measured objectively and the data source can be either a document (e.g., payroll) or self-rated.

I measured professional success in order to evaluate the relation with individual differences in implicit learning and dynamic decision making, which are consistent across different methods and stable over time. As different authors noted (e.g., Epstein, 1979; Wittmann, 1988), the relation between construct variables and criterion variables can only be evaluated accurately if the variables are measured on a similar level of abstraction. In order to measure individual differences that are unaffected by method specific effects (such as the type of data or the data source), I used indicators of different data types and data sources. In order to measure individual differences in professional success that are stable over time, I selected the participants' global career as the frame of reference. In particular, I asked the participants to report their yearly income, their highest educational attainment, and I asked the participants to rate their social status. Yearly income and educational attainment may be seen as global career parameters because they refer to a rather long time period. In the same vein, social status may be seen as a global career indicator because it refers to a profession in general and not to the social status of a specific task. These three measures served as indicators of *objective professional success*.

In addition, the participants' supervisors rated their overall job performance. Hereto, I developed a supervisor rating scale. Based on a literature review, I selected 18 items from Goodman and Svyantek (1999), Higgins, Peterson, Pihl, and Lee (2007), Tsui and Gutek (1984), and Wayne and Liden (1995). Afterwards, $N = 18$ supervisors from different companies and branches rated the appropriateness of these items and I selected the nine items that were rated as most appropriate. Then, $N = 34$ other supervisors (also from different companies and branches) rated a total of $N = 52$ employees with these items. Finally, the five items with the greatest item-total correlation (all $r_{it} \geq .80$) were selected for the supervisor rating scale.

9 The psychometric properties of implicit learning and dynamic decision making (reported in Manuscript 4)

The longitudinal study consisted of two measurement occasions (five months apart) and the participants completed the Advanced Progressive Matrices, the Berlin Structure Intelligence Tests, the Heidelberg Finite State Automaton, the Tailorshop, and two artificial

grammar learning tasks (without grammar knowledge tests) on both measurement occasions. I used latent state-trait models to decompose the variances of the manifest performance variables (Y) into a trait proportion (ξ), a state residual (ζ), a method residual (η), and a measurement error residual (ε). Then, I evaluated the measures by their trait specificity, their occasion specificity, their method specificity, and their reliability. The *trait specificity* (also referred to as consistency) is the proportion of variance that is stable over time and consistent across different methods [$\sigma^2(\xi) / \sigma^2(Y)$]. The *occasion specificity* is the proportion of individual differences that is specific for a particular measurement situation [$\sigma^2(\zeta) / \sigma^2(Y)$]. The *method specificity* is the proportion of variance that is triggered by a particular method [$\sigma^2(\eta) / \sigma^2(Y)$]. These parameters have a range between zero and one, and a greater value indicates a greater specificity. The *reliability* is the sum of these systematic variance proportions and indicates the proportion of systematic individual differences of the trait, the measurement situation, and the method [$\sigma^2(\xi) + \sigma^2(\zeta) + \sigma^2(\eta) / \sigma^2(Y)$]. These parameters are reported in Table 1.

Table 1

Trait-specificity, occasion-specificity, method-specificity, and reliability of the measures

task	measurement occasion	trait-specificity	occasion-specificity	method-specificity	reliability
APM	1	0.72	0	0.14	0.86
APM	2	0.70	0	0.13	0.83
BIS	1	0.67	0	0.22	0.90
BIS	2	0.71	0	0.24	0.95
AGL1	1	0.29	0	0	0.29
AGL1	2	0.31	0	0	0.31
AGL2	1	0.30	0	0	0.30
AGL2	2	0.25	0	0	0.25
Tailorshop	1	0.36	0	0.16	0.52
Tailorshop	2	0.29	0	0.13	0.42
HFA	1	0.44	0	0.36	0.80
HFA	2	0.44	0	0.36	0.80

Note. APM = Advanced Progressive Matrices, BIS = Berlin Intelligence Structure Test, HFA = Heidelberg Finite State Automaton, AGL1 = artificial grammar learning task with grammar 1, AGL2 = artificial grammar learning task with grammar 2, $N = 173$.

As can be seen, the measures of psychometric intelligence contained great proportions of trait specific variances. This replicates the findings of previous investigations (e.g., Carroll, 1993; Conley, 1984; Larsen, Hartmann, Nyborg, 2008) and demonstrates that individual differences in psychometric intelligence can be measured consistently with different methods and that these differences are stable over time.

The analysis further revealed that the occasion specificity was zero for the implicit learning measures. This indicates that no occasion specific effects influenced the measurements. For example, the measurement of individual differences in implicit learning was not affected by the participants' awareness that there is a grammar defining the strings in the learning phase when they completed an artificial grammar learning task for the second time. This replicates the results of my previous studies and suggests that artificial grammar learning tasks may be used several times in order to measure individual differences in implicit learning. The method specificities were also zero, which indicates that there were no method specific effects such as specific characteristics of the grammars that affected the measurements. Taken together, the latent state-trait analysis of the implicit learning revealed that different artificial grammar learning tasks can be used several times to measure individual differences in implicit learning. This suggests that the small reliabilities that have been reported in previous studies (e.g., Gebauer & Mackintosh, 2007; Reber et al., 1991) were not caused by occasion specific or method specific effects but due to random measurement error. The reliability estimates of the implicit learning variables were rather small (≤ 0.31), which indicates that the manifest variables contain great proportions of unsystematic measurement error. This indicates that the manifest variables are poor indicators of implicit learning ability. These results have two important implications. For one thing, the correlations between the performance in artificial grammar learning tasks and the performance in psychometric intelligence tests that have been reported in previous studies (e.g., Gebauer & Mackintosh, 2007; McGeorge et al., 1997; Pretz et al., in press; Reber et al., 1991) are insufficient for drawing conclusions about the relationship between implicit learning ability and psychometric intelligence. The subsequent structural equation model analyses will separate the implicit learning trait variance from the unsystematic variance proportions and reveal insights into the relation between implicit learning ability and psychometric intelligence. For another thing, the small reliability estimates suggest that the manifest performance variables are not suitable for an individual assessment because a performance score will only yield an inaccurate measurement of a person's implicit learning ability.

The dynamic decision making measures (the Tailorshop and the Heidelberg Finite State Automaton) also revealed trait specificities below 0.50 which indicates that less than half of the variance of the manifest performance variables reflect individual differences in dynamic decision making. The analysis further revealed that both measures contained substantial proportion of method specific variance (between 13% and 36%), which suggests that the Tailorshop and the Heidelberg Finite Automaton capture different aspects of dynamic decision making. In particular, the Tailorshop simulation takes place in an economic context where the participants have to lead a company successfully. The Heidelberg Finite State Automaton, on the other hand, takes place in a rather futuristic setting where the participants have to control a space ship. As Beckmann and Guthke (1995) and Hesse (1982) have shown, the semantic context of a dynamic decision making scenario has impact on the decision making processes that take place. Thus, the method specificity of the simulations may partly be explained by the semantic context in which they take place. This finding has implications for the manifest performance variables. For one thing, their trait specificities are too low to use these dynamic decision making tasks for individual assessments. A participant's performance in a single task is not sufficient for making inferences about this participant's dynamic decision making ability. For another thing, a correlation with a *manifest* dynamic decision making variable is not sufficient for drawing conclusions about the relation to the construct dynamic decision making in general. Structural equation modeling makes it possible to investigate the relation with a *latent* dynamic decision making variable, which is adjusted for these method specific effects.

In sum, the latent state-trait analysis revealed that the manifest implicit learning variables and the manifest dynamic decision making variables are poor indicators for the ability constructs. Therefore, I investigated the relations between the constructs using *latent* ability variables, which were adjusted for method specific effects and measurement error.

10 The relation between psychometric intelligence, implicit learning, and dynamic decision making (reported in Manuscript 4)

The correlations between the latent ability variables allow one to evaluate the construct validity of implicit learning and dynamic decision making. The correlations between the latent variables are shown in Table 2. As can be seen, there was a great correlation between psychometric intelligence and dynamic decision making ($r = .86$), which indicates a poor divergent validity of dynamic decision making. The latent intelligence variable explained about 74% of the variance of the latent dynamic decision making variable. This

suggests that dynamic decision making only offers minor insights into cognitive ability beyond IQ. This result replicates the findings of Wirth and Klieme (2003) who reported a correlation of $r = .84$ between a latent intelligence variable and a latent dynamic decision making variable. The results of the latent state-trait analysis further suggest that the heterogeneous findings of previous studies may be due to the heterogeneous reliabilities or the heterogeneous method specificities of dynamic decision making variables. In sum, these findings suggest that the ability to make dynamic decisions is not much more than psychometric intelligence.

Table 2

Correlation between the latent variables

	psychometric intelligence	dynamic decision making	implicit learning	OPS
dynamic decision making	.86***			
implicit learning	.32**	.26*		
OPS	.78***	.52***	.31*	
SR	.03	.25*	-.02	-.07

Note. OPS = objective professional success, SR = supervisor ratings, *** $p < .001$, ** $p < .010$, * $p < .050$, $N = 173$

On the other hand, the relation between implicit learning and psychometric intelligence was less substantial ($r = .32$), the latent intelligence variable explained only 10% of the variance of the latent implicit learning variable. This replicates the findings of previous studies that reported low correlations between implicit learning and psychometric intelligence (Gebauer & Mackintosh, 2007; Feldman, et al., 1995; Kaufman et al., 2010; McGeorge et al., 1997; Reber et al., 1991). Besides, the present study investigated the relation between latent trait variables which were adjusted for measurement error. Therefore, it can be ruled out that the low correlation is a result of the low reliability of the variables. This in turn suggests good divergent validity of implicit learning. The ability to learn implicitly is only weakly related to psychometric intelligence.

All correlations between psychometric intelligence, implicit learning, and dynamic decision making were positive. Following Spearman (1904), this suggests a hierarchical structure of these abilities. I additionally performed a principal component analysis on the

correlations between the latent ability variables. This analysis revealed eigenvalues of 2.02, 0.84, 0.14, which indicates that a single general ability factor can explain about 67% of the variance of the latent variables. This result is in line with Sternberg and Gigorenko (2002) who suggest that a general ability factor is able to explain about 50% of the variance in various performance tasks. To investigate this issue in a greater detail, I used a structural equation model. The respective model is shown in Figure 4. As can be seen, a hierarchical model with one single ability factor at the top of the hierarchy fitted the data well.

Furthermore, the specific ability factors for psychometric intelligence and dynamic decision making were not significant and thus set to zero. This suggests that the intelligence test as well as the dynamic decision making tasks may be seen as indicators for general cognitive ability whereas the artificial grammar learning tasks capture general cognitive ability as well as an incremental proportion of implicit learning ability.

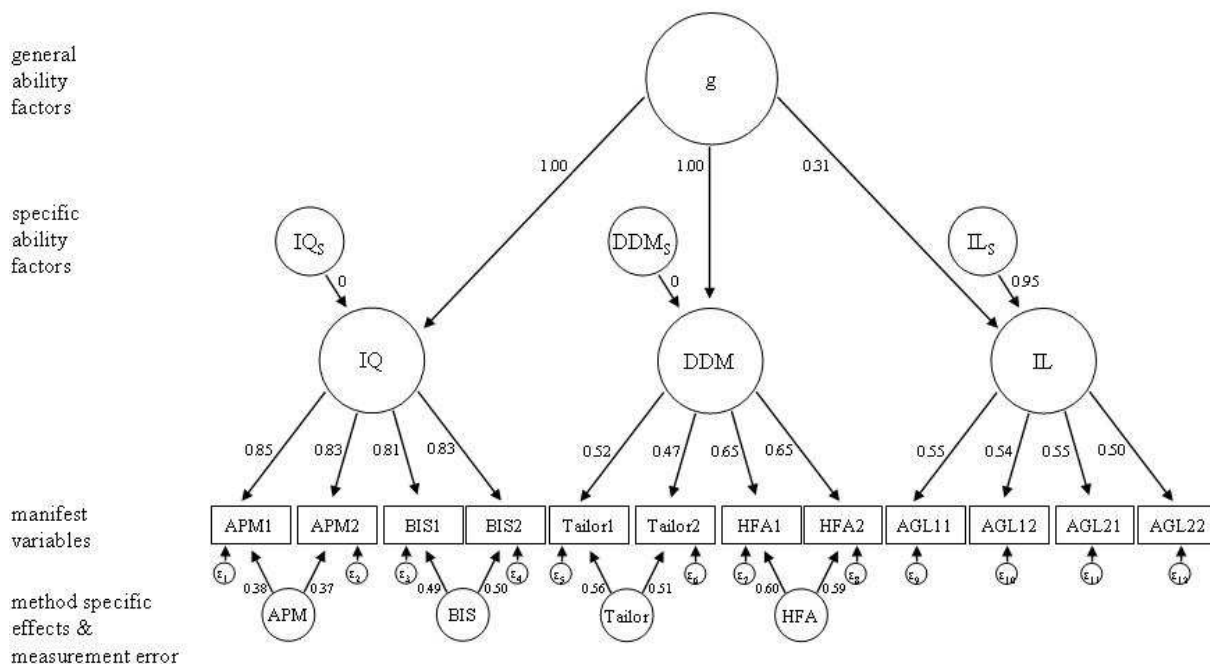


Figure 4. Hierarchical ability model for psychometric intelligence, implicit learning, and dynamic decision making. The standardized path coefficients are reported. *g* = general cognitive ability, *IQ* = psychometric intelligence, *DDM* = dynamic decision making, *IL* = implicit learning, *APM* = Advanced Progressive Matrices, *BIS* = Berlin Intelligence Structure Test, *HFA* = Heidelberg Finite State Automaton, *AGL* = artificial grammar learning task, ϵ_1 - ϵ_{12} = measurement error variables, $\chi^2(58) = 61.60, p = .348, RMSEA = 0.02, CFI = 1.00, N = 173$.

11 The relation between implicit learning, dynamic decision making, and success in real life (reported in Manuscript 4)

The predictive validity of implicit learning and dynamic decision making can be evaluated by their relation to criteria of success in real life. As can be seen in Table 2, there was a significant correlation between implicit learning and objective professional success (indicated by income, social status, and educational attainment). However, a latent regression analysis revealed that this relation decreased and became non-significant when psychometric intelligence was included as a predictor. In addition, the correlation between implicit learning and supervisor ratings was close to zero and not significant. These findings suggest that there is no incremental predictive validity of implicit learning.

There was also a substantial and significant correlation between dynamic decision making and objective professional success. Again, a latent regression analysis revealed that this relation decreased and became non-significant when psychometric intelligence was included as a predictor. In addition, there was a significant correlation between dynamic decision making and supervisor ratings. More importantly, this association remained significant when adjusted for psychometric intelligence. This indicates the incremental predictive validity of dynamic decision making. Thus, even if dynamic decision making is not *much* more than psychometric intelligence, this *little* more offers insights into aspects of success in real life that can not be explained by psychometric intelligence.

The results reported so far refer to latent variables that were adjusted for method specific effects or measurement error. However, in an applied context it may be worthwhile to know, how *manifest* measures can predict manifest criteria. For example, a company which is conducting an assessment center may wish to know how the performance scores of a particular task are related to supervisor ratings. Therefore, I additionally investigated the correlations between the manifest variables. In sum, the greatest correlations were between the measures psychometric intelligence and the indicators of objective professional success (between $r = .22$ and $r = .48$). The correlations with the manifest dynamic decision making variables were less substantial (between $r = .05$ and $r = .23$) as were the correlations with the manifest implicit learning variables (between $r = -.04$ and $r = .19$). The supervisor ratings only correlated significantly with the dynamic decision making tasks (between $r = .12$ and $r = .20$) which indicates that even if the dynamic decision making measures offer an incremental predictive value, their explanatory power is limited.

12 Summary and Conclusion

The aim of the present work was to evaluate whether implicit learning and dynamic decision making are useful constructs to describe cognitive ability beyond IQ. Therefore, I investigated the incremental and the predictive validity of the constructs, and the psychometric properties of the performance measures. There are six core findings of my investigations. First, implicit learning is only weakly related to psychometric intelligence, even after adjusting for measurement error. This indicates that implicit learning captures individual performance differences beyond IQ and suggests a good divergent validity of implicit learning. Second, there is a great association between dynamic decision making and psychometric intelligence. This speaks against the divergent validity of dynamic decision making. Third, implicit learning as well as dynamic decision making can be integrated into hierarchical models of cognitive ability. The present findings revealed that both constructs load substantially on a general ability factor. In addition, implicit learning reveals a specific ability component whereas dynamic decision making captures no incremental variance. Fourth, there is no evidence for the incremental predictive validity of implicit learning. In the present study, there was only a weak association between implicit learning and professional success. Furthermore, this association vanished when adjusted for psychometric intelligence. Fifth, dynamic decision making can incrementally predict supervisor ratings, even though there is a great overlap between psychometric intelligence and dynamic decision making. This was true for the latent ability variables as well as for the manifest performance indicators. Hence, even if there are only minor individual differences in dynamic decision making beyond IQ, these individual differences can explain success in real life in greater detail. Sixth, the trait specificities of the manifest measures of implicit learning and dynamic decision making were too small to use these measures for individual assessments. Investigating the measurement of implicit learning and dynamic decision making in greater detail will make these constructs valuable supplements not only in research contexts but also in applied contexts.

Taken together, these findings show that implicit learning as well as dynamic decision making are useful constructs for investigating individual differences in cognitive ability. Implicit learning is largely independent of psychometric intelligence and offers insights in cognitive ability beyond IQ. Even though there are only minor individual differences in dynamic decision making beyond psychometric intelligence, these ability differences play a significant role for achieving success in real life. However, in order to use implicit learning

tasks and dynamic decision making tasks for an individual assessment, the psychometric properties of the performance measures have to be improved.

References

- Altmann, G. T. M., Dienes, Z., & Goode, A. (1995). Modality independence of implicitly learned grammatical knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 899-912. doi: 10.1037/0278-7393.21.4.899
- Barth, C. M. & Funke, J. (2010). Negative affective environments improve complex solving performance. *Cognition and Emotion*, *24*, 1259-1268. doi: 10.1080/02699930903223766
- Beckmann, J. F. & Guthke, J. (1995). Complex problem solving, intelligence, and learning ability. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: the European perspective* (pp. 177-200). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Berg, C. A. & Sternberg, R. J. (1985). A triarchic theory of intellectual development during adulthood. *Developmental Review*, *5*(4), 334-370. doi: 10.1016/0273-2297(85)90017-6
- Berry, D. C. & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *36A*(2), 209-231. doi: 10.1080/14640748408402156
- Buchner, A., Funke, J., & Berry, D. C. (1995). Negative correlations between control performance and verbalizable knowledge: Indicators for implicit learning in process control tasks? *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *48A*(1), 166-187. doi: 10.1080/14640749508401383
- Burt, C. (1949). The structure of the mind; a review of the results of factor analysis. *British Journal of Educational Psychology*, *19*, 176-199. doi: 10.1111/j.2044-8279.1949.tb01612.x
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.
- Cattell, J. M. & Farrand, L. (1896). Physical and mental measurements of the students of Columbia University. *Psychological Review*, *3*(6), 618-648. doi: 10.1037/h0070786
- Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1-22. doi: 10.1037/h0046743
- Ceci, S. J. & Liker, J. K. (1986). A day at the races: A study of IQ, expertise, and cognitive complexity. *Journal of Experimental Psychology: General*, *115*(3), 255-266. doi: 10.1037/0096-3445.115.3.255

- Conley, J. J. (1984). The hierarchy of consistency: A review and model of longitudinal findings on adult individual differences in intelligence, personality and self-opinion. *Personality and Individual Differences*, 5(1), 11-25. doi: 10.1016/0191-8869(84)90133-8
- Detle, D. E., Abele, A. E., & Renner, O. (2004). Zur Definition und Messung von Berufserfolg: Theoretische Überlegungen und metaanalytische Befunde zum Zusammenhang von externen und internen Laufbahnerfolgsmäßen. *Zeitschrift für Personalpsychologie*, 3(4), 170-183. doi: 10.1026/1617-6391.3.4.170
- Dörner, D. (1979). *Programm TAILORSHOP in der Version für TI-59 mit Drucker PC-100. Modifizierte und kommentierte Fassung von Norbert Streitz*. Aachen.
- Dörner, D. (1980). On the difficulty people have in dealing with complexity. *Simulation & Games*, 11(1), 87-106.
- Dörner, D. (1986). Diagnostik der operativen Intelligenz. *Diagnostica*, 32(4), 290-308.
- Dulany, D. E., Carlson, R. A., & Dewey, G. I. (1984). A case of syntactical learning and judgment: How conscious and how abstract? *Journal of Experimental Psychology: General*, 113(4), 541-555. doi: 10.1037/0096-3445.113.4.541
- Eid, M., Notz, P., Steyer, R., & Schwenkmezger, P. (1994). Validating scales for the assessment of mood level and variability by latent state-trait analyses. *Personality and Individual Differences*, 16, 63-76. doi: 10.1016/0191-8869(94)90111-2
- Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology*, 37(7), 1097-1126. doi: 10.1037/0022-3514.37.7.1097
- Feldman, J., Kerr, B., & Streissguth, A. P. (1995). Correlational analyses of procedural and declarative learning performance. *Intelligence*, 20(1), 87-114. doi: 10.1016/0160-2896(95)90007-1
- Funke, J. (1983). Einige Bemerkungen zu Problemen der Problemlöseforschung oder: Ist Testintelligenz doch ein Prädiktor? *Diagnostica*, 29, 283-302.
- Funke, J. (2010). Complex problem solving: A case for complex cognition? *Cognitive Processing*, 11(2), 133-142. doi: 10.1007/s10339-009-0345-0
- Galton, F. (1883). *Inquiries into human faculty and its development*. London: MacMillan.
- Gebauer, G. F. & Mackintosh, N. J. (2007). Psychometric intelligence dissociates implicit and explicit learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(1), 34-54. doi: 10.1037/0278-7393.33.1.34

- Gonzalez, C., Vanyukov, P., & Martin, M. K. (2005). The use of microworlds to study dynamic decision making. *Computers in Human Behavior, 21*(2), 273-286. doi: 10.1016/j.chb.2004.02.014
- Goodman, S. A. & Svyantek, D. J. (1999). Person–organization fit and contextual performance: Do shared values matter. *Journal of Vocational Behavior, 55*(2), 254-275. doi: 10.1006/jvbe.1998.1682
- Hagemann, D., Hewig, J., Seifert, J., Naumann, E., & Bartussek, D. (2005). The latent state-trait structure of resting EEG asymmetry: Replication and extension. *Psychophysiology, 42*, 740-752. doi: 10.1111/j.1469-8986.2005.00367.x
- Hermes, M., Hagemann, D., Britz, P., Lieser, S., Bertsch, K., Naumann, E. (2009). Latent state–trait structure of cerebral blood flow in a resting state. *Biological Psychology, 80*, 196-202. doi: 10.1016/j.biopsycho.2008.09.003
- Hesse, F. W. (1982). Effekte des semantischen Kontexts auf die Bearbeitung komplexer Probleme. *Zeitschrift für Experimentelle und Angewandte Psychologie, 29*(1), 62-91.
- Higgins, D. M., Peterson, J. B., Pihl, R. O., & Lee, A. G. M. (2007). Prefrontal cognitive ability, intelligence, Big Five personality, and the prediction of advanced academic and workplace performance. *Journal of Personality and Social Psychology, 93*(2), 298-319. doi: 10.1037/0022-3514.93.2.298
- Hofstätter, P. R. (1957). *Psychologie*. Oxford: Fischer Buecherei.
- Horn, J. L. & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57*(5), 253-270. doi: 10.1037/h0023816
- Jäger, A. O. (1973). *Dimensionen der Intelligenz*. Göttingen: Hogrefe.
- Jäger, A. O. (1982). Mehrmodale Klassifikation von Intelligenzleistungen: Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells. *Diagnostica, 28*(3), 195-225.
- Jäger, A. O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven. *Psychologische Rundschau, 35*(1), 21-35.
- Jäger, A. O., Süß, H.-M., & Beauducel, A. (1997). *Berliner Intelligenzstruktur - Test. Form 4*. Göttingen: Hogrefe.
- Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition, 116*(3), 321-340. doi: 10.1016/j.cognition.2010.05.011

- Kinder, A., Shanks, D. R., Cock, J., & Tunney, R. J. (2003). Recollection, fluency, and the explicit/implicit distinction in artificial grammar learning. *Journal of Experimental Psychology: General*, *132*(4), 551-565. doi: 10.1037/0096-3445.132.4.551
- Kluwe, R. H., Misiak, C., & Haider, H. (1991). The control of complex systems and performance in intelligence tests. In H. A. H. Rowe (Ed.), *Intelligence: Reconceptualization and measurement*. (pp. 227-244). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Knauff, M. & Wolf, A. G. (2010). Complex cognition: The science of human reasoning, problem-solving, and decision-making. *Cognitive Processing*, *11*(2), 99-102. doi: 10.1007/s10339-010-0362-z
- Knowlton, B. J. & Squire, L. R. (1994). The information acquired during artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(1), 79-91. doi: 10.1037/0278-7393.20.1.79
- Knowlton, B. J. & Squire, L. R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 169-181. doi: 10.1037/0278-7393.22.1.169
- Kröner, S., Plass, J. L., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence*, *33*(4), 347-368. doi: 10.1016/j.intell.2005.03.002
- Larsen, L., Hartmann, P., & Nyborg, H. (2008). The stability of general intelligence from early adulthood to middle-age. *Intelligence*, *36*(1), 29-34. doi: 10.1016/j.intell.2007.01.001
- Leutner, D. (1988). Computersimulierte dynamische Systeme: Wissenserwerb unter verschiedenen Lehrmethoden und Sozialformen des Unterrichts. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *20*(4), 338-355.
- Mackintosh, N. J. (1998). *IQ and human intelligence*. New York: Oxford University Press.
- Mathews, R. C., Buss, R. R., Stanley, W. B., Blanchard-Fields, F., Cho, J. R., & Druhan, B. (1989). Role of implicit and explicit processes in learning from examples: A synergistic effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(6), 1083-1100. doi: 10.1037/0278-7393.15.6.1083
- McGeorge, P., Crawford, J. R., & Kelly, S. W. (1997). The relationships between psychometric intelligence and learning in an explicit and an implicit task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(1), 239-245. doi: 10.1037/0278-7393.23.1.239

- Meulemans, T. & Van der Linden, M. (2003). Implicit learning of complex information in amnesia. *Brain and Cognition*, 52(2), 250-257. doi: 10.1016/s0278-2626(03)00081-2
- Münsterberg, H. (1891). Zur Individualpsychologie. *Centralblatt für Nervenheilkunde und Psychatrie*, 14, 196-198.
- Neisser, U., Boodoo, G., Bouchard, T. J. Jr., Boykin, A. W., Brody, N., Ceci, S. J. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, 51(2), 77-101. doi: 10.1037/0003-066x.51.2.77
- Ng, T. W. H., Eby, L. T., Sorensen, K. L., & Feldman, D. C. (2005). Predictors of objective and subjective career success. A meta-analysis. *Personnel Psychology*, 58(2), 367-408. doi: 10.1111/j.1744-6570.2005.00515.x
- Perruchet, P. & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General*, 119(3), 264-275. doi: 10.1037/0096-3445.119.3.264
- Pothos, E. M. & Bailey, T. M. (2000). The role of similarity in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(4), 847-862. doi: 10.1037/0278-7393.26.4.847
- Pretz, J. E., Totz, K. S., & Kaufman, S. B. (in press). The effects of mood, cognitive style, and cognitive ability on implicit learning. *Learning and Individual Differences*. doi: 10.1016/j.lindif.2009.12.003
- Putz-Osterloh, W. (1981). Über die Beziehung zwischen Testintelligenz und Problemlöseerfolg. *Zeitschrift für Psychologie mit Zeitschrift für angewandte Psychologie*, 189(1), 79-100.
- Putz-Osterloh, W. (1983). Über Determinanten komplexer Problemlöseleistungen und Möglichkeiten zu ihrer Erfassung. *Sprache & Kognition*, 2(2), 100-116.
- Putz-Osterloh, W., Bott, B., & Köster, K. (1990). Modes of learning in problem solving: Are they transferable to tutorial systems? *Computers in Human Behavior*, 6(1), 83-96. doi: 10.1016/0747-5632(90)90032-c
- Putz-Osterloh, W. & Lüer, G. (1981). The predictability of complex problem solving by performance on an intelligence test. *Zeitschrift für Experimentelle und Angewandte Psychologie*, 28(2), 309-334.
- Raven, J. C., Court, J. H., & Raven, J. (1994). *Manual for Raven's Progressive Matrices and Mill Hill Vocabulary Scales. Advanced Progressive Matrices*. Oxford: Oxford Psychologists Press.

- Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning & Verbal Behavior*, 6, 855-863. doi: 10.1016/s0022-5371(67)80149-x
- Reber, A. S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*. New York: Oxford University Press.
- Reber, A. S., Walkenfeld, F. F., & Hernstadt, R. (1991). Implicit and explicit learning: Individual differences and IQ. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(5), 888-896. doi: 10.1037/0278-7393.17.5.888
- Reber, R. & Perruchet, P. (2003). The use of control groups in artificial grammar learning. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 56A(1), 97-115. doi: 10.1080/02724980244000297
- Rigas, G., Carling, E., & Brehmer, B. (2002). Reliability and validity of performance measures in microworlds. *Intelligence*, 30(5), 463-480. doi: 10.1016/s0160-2896(02)00121-6
- Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European Community. *Journal of Applied Psychology*, 88(6), 1068-1081. doi: 10.1037/0021-9010.88.6.1068
- Schmidt, F. L. & Hunter, J. (2004). General Mental Ability in the World of Work: Occupational Attainment and Job Performance. *Journal of Personality and Social Psychology*, 86(1), 162-173. doi: 10.1037/0022-3514.86.1.162
- Schmitt, M. J. & Steyer, R. (1993). A latent state-trait model (not only) for social desirability. *Personality and Individual Differences*, 14, 519-529. doi: 10.1016/0191-8869(93)90144-r
- Scott, R. B. & Dienes, Z. (2008). The conscious, the unconscious, and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(5), 1264-1288. doi: 10.1037/a0012943
- Scott, R. B. & Dienes, Z. (2010). Fluency does not express implicit knowledge of artificial grammars. *Cognition*, 114(3), 372-388. doi: 10.1016/j.cognition.2009.10.010
- Servan-Schreiber, E. & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(4), 592-608. doi: 10.1037/0278-7393.16.4.592
- Shanks, D. R. & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences*, 17(3), 367-447. doi: 10.1017/S0140525X00035032

- Sharp, S. E. (1899). Individual psychology: A study in psychological method. *The American Journal of Psychology*, *10*(3), 329-391. doi: 10.2307/1412140
- Spearman, C. (1904). 'General intelligence', objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201-293. doi: 10.2307/1412107
- Stemmler, G., Hagemann, D., Amelang, M., & Bartussek, D. (2011). *Differentielle Psychologie und Persönlichkeitsforschung*. Stuttgart: Kohlhammer.
- Stern, W. (1911). *Intelligenzproblem und Schule*. Leipzig: Teubner.
- Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. Oxford: Lawrence Erlbaum.
- Sternberg, R. J. & Grigorenko, E. L. (2002). The theory of successful intelligence as a basis for gifted education. *Gifted Child Quarterly*, *46*(4), 265-277. doi: 10.1177/001698620204600403
- Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality*, *13*(5), 389-408. doi: 10.1002/(sici)1099-0984(199909/10)13:5<389::aid-per361>3.0.co;2-a
- Steyer, R., Schwenkmezger, P., & Auer, A. (1990). The emotional and cognitive components of trait anxiety: A latent state-trait model. *Personality and Individual Differences*, *11*, 125-134. doi: 10.1016/0191-8869(90)90004-b
- Stinson, B. (2008). *The Bro Code*. New York: Touchstone.
- Süß, H.-M., Kersting, M., & Oberauer, K. (1993). Zur Vorhersage von Steuerungsleistungen an computersimulierten Systemen durch Wissen und Intelligenz. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *14*(3), 189-203.
- Tsui, A. S. & Gutek, B. A. (1984). A role set analysis of gender differences in performance, affective relationships, and career success of industrial middle managers. *Academy of Management Journal*, *27*(3), 619-635. doi: 10.2307/256049
- Tunney, R. J. (2005). Sources of confidence judgments in implicit cognition. *Psychonomic Bulletin & Review*, *12*(2), 367-373.
- Vernon, P. E. (1950). *The structure of human abilities*. Oxford: Wiley.
- Wagener, D. (2001). *Psychologische Diagnostik mit komplexen Szenarios. Taxonomie, Entwicklung, Evaluation*. Lengerich: Pabst Science Publishers.
- Wallach, D. (1998). *Komplexe Regelungsprozesse. Eine kognitionswissenschaftliche Analyse*. Wiesbaden: Deutscher Universitäts-Verlag.
- Wayne, S. J. & Liden, R. C. (1995). Effects of impression management on performance ratings: A longitudinal study. *Academy of Management Journal*, *38*(1), 232-260. doi:

10.2307/256734

- Wechsler, D. (1975). Intelligence defined and undefined: A relativistic appraisal. *American Psychologist*, 30(2), 135-139. doi: 10.1037/h0076868
- Whittlesea, B. W., Jacoby, L. L., & Girard, K. (1990). Illusions of immediate memory: Evidence of an attributional basis for feelings of familiarity and perceptual quality. *Journal of Memory and Language*, 29(6), 716-732. doi: 10.1016/0749-596x(90)90045-2
- Whittlesea, B. W. A. & Leboe, J. P. (2000). The heuristic basis of remembering and classification: Fluency, generation, and resemblance. *Journal of Experimental Psychology: General*, 129(1), 84-106. doi: 10.1037/0096-3445.129.1.84
- Wirth, J. & Funke, J. (2005). Dynamisches Problemlösen: Entwicklung und Evaluation eines neuen Messverfahrens zum Steuern komplexer Systeme. In E. Klieme, D. Leutner & J. Wirth (Eds.), *Problemlösekompetenz von Schülerinnen und Schülern*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Wirth, J. & Klieme, E. (2003). Computer-based assessment of problem solving competence. *Assessment in Education: Principles, Policy & Practice*, 10(3), 329-345. doi: 10.1080/0969594032000148172
- Wissler, C. (1901). The correlation of mental and physical tests. *Psychological Monographs*, 3(6), 1-62.
- Wittmann, W. (1988). Multivariate reliability theory: Principles of symmetry and successful validation strategies. In J. R. Cattell (Ed.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 505-560). New York: Plenum Press.
- Wittmann, W. & Hatrup, K. (2004). The relationship between performance in dynamic systems and intelligence. *Systems Research and Behavioral Science*, 21, 393-409. doi: 10.1002/sres.653
- Yasuda, T., Lawrenz, C., Whitlock, R. V., Lubin, B., & Lei, P.-W. (2004). Assessment of intraindividual variability in positive and negative affect using latent state-trait model analyses. *Educational and Psychological Measurement*, 64, 514-530. doi: 10.1177/0013164403258445
- Ziegler, M., Ehrlenspiel, F., & Brand, R. (2009). Latent state-trait theory: An application in sport psychology. *Psychology of Sport and Exercise*, 10, 344-349. doi: 10.1016/j.psychsport.2008.12.004

Appendix A1

Manuscript 1: Measuring individual differences in implicit learning with an artificial grammar learning task

Measuring individual differences in implicit learning
with an artificial grammar learning task

Daniel Danner, Dirk Hagemann, Andrea Schankin, Marieke Hager, & Joachim Funke
Institute of Psychology, University of Heidelberg, Germany

Author Note

This research was funded by German Research Foundation Grant DFG, Ha 3044/7-1. We gratefully thank Andreas Neubauer, Anna-Lena Schubert, and Katharina Weskamp for conducting the experiments and two anonymous reviewers for their helpful comments on an earlier draft of this manuscript.

Correspondence concerning this paper should be addressed to Daniel Danner, University of Heidelberg, Institute of Psychology, Hauptstrasse 47-51, D-69117 Heidelberg, Germany, Phone: +49 (0) 6221-547354, Fax: +49 (0) 6221-547325, Email: daniel.danner@psychologie.uni-heidelberg.de

Abstract

The present study investigates whether an artificial grammar learning task may be used to measure individual differences in implicit learning. In three experiments, the participants had to rate either the grammaticality or the novelty of letter strings. The results indicate that only a task with the instruction to rate the grammaticality but not a task with the instruction to rate the novelty measures reliable and consistent individual differences in implicit learning.

Furthermore, it is shown that when the participants are asked to rate the grammaticality of letter strings, the task can only be used once to measure implicit learning. Subsequently, the role of strategy use and implications for further and past research are discussed.

Keywords: implicit learning, artificial grammar learning, individual differences, reliability

Measuring individual differences in implicit learning
with an artificial grammar learning task

Implicit learning is a process of acquiring complex information without awareness of what has been learned (Frensch & Rüniger, 2003; Frensch, 1998; Seger, 1993). Like in any learning process there may be substantive individual differences, but not much is known about their magnitude and meaning for a successful completion of cognitive laboratory tasks or mastering the challenges of everyday life. Whereas some authors reason that implicit learning is executed by evolutionary old systems, is essential for survival and therefore shows only minor individual differences (Reber, 1992), others have postulated that individual differences in implicit learning may be a powerful determinant of success in educational and work achievement and thus may have the same preponderance as general intelligence (Mackintosh, 1998). Recently, Kaufman, DeYoung, Gray, Jiménez, Brown, and Mackintosh (2010) and Pretz, Trotz, and Kaufman (2009) reported significant associations between implicit learning performance and academic achievement, which supports Mackintosh's hypothesis. However, beyond these investigations, there is only weak empirical evidence to support such claims. One of the major reasons for this may be the lack of a reliable task for the measurement of individual differences in implicit learning. The present paper reports on a series of experiments that aim to fill this gap by investigating the reliability and the task consistency of artificial grammar learning tasks (Reber, 1967), which is a standard procedure in implicit learning research. Experiment 1 and 2 will reveal that not every type of artificial grammar learning tasks is suitable for the measurement of individual differences. However, experiment 3 will demonstrate how the artificial grammar learning task can be used to measure reliable individual differences in implicit learning.

The artificial grammar learning task

An artificial grammar learning task consists of a learning phase and a testing phase. In the learning phase, the participants are asked to learn a list of apparently arbitrary letter strings (like WNSNXS). Afterwards in the testing phase, they are told that these strings were constructed according to a complex rule system (a grammar) and they are asked to judge newly presented strings (like NWSWWN) as either grammatical or non-grammatical. Typically, the participants show above chance performance, which suggests that they learned something but they are not able to report the grammar rules, which suggests that they learned the rules implicitly. Therefore, implicit learning may be assessed based on two criteria: the judgment accuracy in the testing phase and the amount of reportable grammar knowledge.

Judgment accuracy. The most popular indicator for implicit learning success within an artificial grammar learning paradigm is the judgment accuracy, which is commonly quantified as the percentage of correct judgments within the testing phase. In particular, a mean percentage of correct judgments that is significantly above chance suggests that implicit learning took place.

Grammar knowledge. Artificial grammar learning tasks are labeled implicit learning tasks because there appears to be no relation between participants' judgment accuracy and their amount of knowledge about the grammar. However, there is a lively discussion what kind of knowledge may be relevant for an artificial grammar learning task and how it should be assessed. When we ask the participants to reproduce the underlying grammar, we presume that people reach an above-chance accuracy because they learned something about the underlying *grammar*. However, this does not have to be true. Several authors suggested that the participants may not learn the grammatical rules implicitly but instead may use *heuristics* like bigrams (Perruchet & Pacteau, 1990), fragments (Dulany, Carlson, & Dewey, 1984), or chunks (Servan-Schreiber & Anderson, 1990). In particular, Perruchet and Pacteau

(1990) have conducted a series of experiments and showed that learning bigrams in the learning phase is as effective as learning grammatical letter strings and that the classification of bigrams corresponds to the classification of letter strings. In the same vein, Dulany et al. (1984) have shown that the participants in an artificial grammar learning task can report which fragments they use to make grammaticality judgments. Furthermore, they have shown that the reported knowledge (containing bi- and trigrams) can predict grammaticality judgments. In sum, these findings suggest that knowledge of n-grams (bi- and trigrams) may be relevant for the performance in artificial grammar learning tasks. In addition to that, an n-gram knowledge test may also be seen as an indirect form of other forms of grammar knowledge. In particular, using knowledge of n-grams is just one possibility to succeed in an n-gram knowledge test. Another strategy may be to use more abstract knowledge and deduce the answers for the knowledge. Thus, an n-gram knowledge test may measure different forms of grammar knowledge. Therefore, we asked the participants to rate whether an n-gram occurred more often in grammatical or more often in non-grammatical strings. A zero correlation between n-gram knowledge and accuracy would indicate that the participants did not use n-grams for their judgments, whereas a positive correlation between n-gram knowledge and judgment accuracy would indicate that the participants may have used n-grams for their judgments.

Individual differences and reliability

The reliability of an implicit learning measure is important for several reasons. First, some studies showed that there is no relation between measures of artificial grammar learning and measures of knowledge about the underlying grammar (e.g., Reber, & Allen, 1978) or general intelligence (e.g., Gebauer & Mackintosh, 2007; Kaufman et al., 2010; McGeorge, Crawford, & Kelly, 1997; Reber, Walkenfeld, and Hernstadt, 1991). This is often taken as an evidence for the divergent validity of the measurement, i.e. the proposition that individual

differences in implicit learning may constitute an independent ability. However, this argument only holds true if both measures are reliable. If not, a low correlation between the measures may also be explained by the low reliability of the measurement.

Second, Mackintosh (1998) suggested that implicit learning is a powerful predictor of educational or professional success (see also Kaufman et al., 2010; Pretz et al., 2010). If this holds true, implicit learning may be a very useful construct to describe human mental ability or predict success in later life. However, such a construct would only be useful if it can be measured reliably. In particular, if an artificial grammar learning task may be used to measure a single person's implicit learning ability (e.g., as part of an assessment center) then this measurement is only useful if it is reliable because otherwise it will yield incorrect decisions.

Third, when implicit learning is considered as a mental ability or a trait it is also an important issue whether this ability can be measured with more than one method (Campbell & Fiske, 1959). In principle, a great correlation between several procedures that are designed to measure the same construct indicates a good convergent validity of the measures. However, such a correlation is only informative if the measurements are reliable.

Given the importance of reliability considerations, it is surprising that there are only a few publications that report reliability estimates for measures of artificial grammar learning. Reber et al. (1991) examined $N=20$ students and reported a Cronbach's alpha of $\alpha=.51$ for 100 grammaticality judgments. This result shows that it is possible to measure individual differences in implicit learning although this measurement is not very consistent. However, one limitation of this study is that only a single grammar was used. Gebauer and Mackintosh (2007) assessed $N=605$ pupils. They used two different grammars and presented 80 letter strings in the testing phase. Based on 80 grammaticality decisions they reported a split-half correlation of $r=.70$. Although this sample was much larger, there may be another difficulty in their study. Gebauer and Mackintosh conducted two artificial grammar learning tasks and

reported the split-half correlation pooled over both tasks. However, the mean accuracy in task one was 67.08%, whereas the mean accuracy in task two was 61.36%. Because both tasks apparently varied in their difficulty, it may be possible that the reported correlation was increased by the pooling and thus overestimated the true reliability of the measurement.

Task consistency

There is also an obstacle for any study on the task consistency of the measurement. The task consistency may be important in a research context. For example, to test whether the artificial grammar learning performance measures a trait-like ability that is stable over time. In an applied context, it may be important for an individual assessment (e.g., if an applicant is tested more than one time). Estimating the task consistency would require the same participants to complete at least two artificial grammar learning tasks with two different underlying grammars. There lies one difficulty in this approach. When participants complete the learning phase for the first time, they do not know that there is a grammar behind the letter strings. In the testing phase they are told that there is a grammar and that they should rate the grammaticality of newly presented strings. Thus, when participants complete the learning phase for the *second* time, they already know about the grammar. The participants will also know that there will be a testing phase and that they will be asked to judge new strings as grammatical or non-grammatical. This may cause them not to memorize the strings but to search for the grammar or simple heuristics that may help them later in the testing phase. For that reason, Gebauer and Mackintosh (2007) modified the standard paradigm and asked their participants not to rate the grammaticality but the novelty of the strings in the testing phase. However, none of the strings were previously presented and if the participants (inadvertently) classified a newly presented letter string as an “old” one, this was scored as a correct decision. The idea behind this procedure may be that the participants learn something about the grammar, thus they feel familiar with the grammatical strings and therefore they classify a

grammatical string as an “old” one. In Gebauer and Mackintosh’s (2007) study, there was a significant correlation of $r=.15$ between the judgment accuracy of the two different artificial grammar learning tasks, which suggests a low task consistency. However, it remains unclear if rating the novelty of the strings measures the same construct than rating their grammaticality.

From a conceptual point of view, novelty judgments and grammaticality judgments may be seen as similar. For example, Whittlesea and Leboe (2000) demonstrated that several heuristics (fluency, generation, and resemblance) influence the performance in recognition tasks as well as in classification tasks. The authors suggest that these heuristics affect the perceived familiarity of stimuli and that the familiarity affects novelty judgments as well as grammaticality judgments. In line with this suggestion, Scott and Dienes (2008) demonstrated that grammaticality ratings can be predicted by the perceived familiarity of strings. Furthermore, there are findings, which suggest that the fluency of the processing of the stimuli affects novelty ratings (e.g., Whittlesea, Jacoby, & Girard, 1990) as well as grammaticality ratings (Kinder, Shanks, Cock, & Tunney, 2003). This further points towards the conceptual similarity of both measures. However, from an empirical point of view, it is unclear whether asking participants to rate the novelty of letter strings measures the same construct than asking participants to rate the grammaticality of letter strings. Therefore it is not known at present if this result indicates a low consistency of artificial grammar learning in general or just in case the participants are asked to rate the novelty instead of the grammaticality of the strings.

The present study

Taken together, there is only weak support for the measurement of reliable individual differences in artificial grammar learning. The task consistency of those measures is also unclear. Therefore, the general aim of the present study was to examine an artificial grammar

learning task as a measure of reliable and consistent individual differences in implicit learning with three experiments.

Experiment 1 was designed to test whether asking the participants to rate the grammaticality of a newly presented letter string in the testing phase quantifies the same construct as asking the participants to judge the novelty of the letter strings. The major aim of experiment 2 was to test the reliability and the task consistency of an artificial grammar learning task when the participants were asked to judge the novelty of the letter strings. Experiment 3 aimed at the reliability and the task consistency of an artificial grammar learning task when the participants were asked to judge the grammaticality of the letter strings. Finally, a conjoint analysis of experiment 1, 2, and 3 was conducted in order to test whether individual differences may be quantified with the instruction to rate the grammaticality of strings as well as with the instruction to rate its novelty.

Experiment 1

To estimate the task consistency of the performance in an artificial grammar learning task, it is necessary that the same participants complete more than one task. Because this procedure may cause a validity problem, Gebauer and Mackintosh (2007) asked their participants to rate the novelty of the letter strings instead of their grammaticality.

The idea behind this procedure may be that the participants learn something about the grammar, thus feel familiar with the grammatical strings and therefore classify a grammatical string as an “old” one. Although this idea is theoretically sound, there is no empirical evidence for the presumed similarity of grammaticality and novelty ratings. Hence the aim of experiment 1 was to test if asking participants for novelty measures the same construct as asking for grammaticality. Therefore two artificial grammar learning tasks were presented along with these two instructions. The correlation between the two tasks indicates the extent to which both judgments measure the same construct.

Method

Participants. The participants were $N=21$ students from the University of Heidelberg who were recruited from the campus and were paid €5 for their participation. This sample size was chosen because it allows a detection of a population correlation of $r=.50$ between accuracy of novelty and grammaticality rating with a type-one-error probability of 0.05 (one-tailed) and a power of 0.80 (Faul, Erdfelder, Lang, & Buchner, 2007).

Stimulus material. The letter strings were the same as used by Gebauer and Mackintosh (2007). There were two grammars. For each grammar, there were 30 grammatical strings in the learning phase and 40 grammatical and 40 non-grammatical strings in the testing phase (see Appendix, Table A1 and Table A2). The grammatical strings were constructed according to Figure 1 and Figure 2. The non-grammatical strings contained one violation of the grammar at random positions of the strings. The length of the strings varied between three and eight letters.

Please insert Figure 1 and 2 about here

To test the reportable grammar knowledge of the participants, 12 n-grams were selected for each grammar. There were 6 n-grams which occurred in the learning phase and which also occurred in the testing phase more frequently in grammatical than non-grammatical strings (NX, XS, SN, NXS, WNS, NWS for grammar 1 and MM, LM, RH, LMM, MMM, RHP for grammar 2, respectively). These n-grams were chosen because they may help to identify grammatical strings as grammatical. In addition, there were 6 n-grams which did not occur in the learning phase but which did occur in the testing phase more frequently in non-grammatical strings than in grammatical ones (NN, XN, XX, WSS, NWW, SSW for grammar 1 and MP, RM, LH, HHP, HPL, LMH for grammar 2, respectively). Those

strings were chosen because they may help to identify non-grammatical strings. The strings were presented on a 17" screen of a personal computer with a standard German keyboard.

Procedure. Each participant completed two artificial grammar learning tasks. The *first artificial grammar learning task* was run with grammar 1. In the *learning phase* 30 letter strings were presented and the participants were instructed to memorize them (e.g., WNSNXS). Each string was presented individually for 4 s on a 17" screen of a personal computer. The participants were asked to repeat the strings correctly by pressing the respective letters on the keyboard. When a string was repeated correctly, the next string occurred. When a string was repeated incorrectly, the string was displayed again until repeated correctly. After a participant repeated ten strings correctly, these ten strings were simultaneously displayed for 90 s on the screen and the participant was asked to repeat them silently. After a participant repeated all 30 string correctly the learning phase was finished. In the *testing phase* 80 new strings were presented. Even though all strings were new (have not been presented in the learning phase), the participants were instructed to rate the strings as "old" (presented in the learning phase) or "new" (not presented in the learning phase). To judge a string as "old", the participants had to press the A-key of the keyboard, to judge a string as "new" they had to press the L-key. The strings were presented in a new random order for each participant. Immediately after the testing phase, the participants completed the knowledge tests. In the *n-gram knowledge test*, the participants were instructed to judge whether an n-gram occurred more often in "old" strings or whether an n-gram occurred more often in "new" strings. To judge an n-gram as occurring more often in "old" strings, the participants had to press the A-key of the keyboard, to judge an n-gram as occurring more often in "new" string, they had to press the L-key. The n-grams were presented in a new random order for each participant.

The *second artificial grammar learning task* was run with grammar 2. The procedure of the *learning phase* was the same as in the first artificial grammar learning task. However, after the learning phase was finished, the participants were informed that all strings in the learning phase were constructed according to a complex rule system. In the *testing phase* 80 new strings were presented (see Appendix, Table A2). The participants were instructed to rate the strings as grammatical or non-grammatical. To judge a string as grammatical, the participants had to press the A-key of the keyboard, to judge a string as non-grammatical, they had to press the L-key. The strings were presented in a new random order for each participant. In the *n-gram knowledge test*, the participants were instructed to judge whether an n-gram occurred more often in grammatical strings or whether an n-gram occurred more often in non-grammatical strings. To judge an n-gram as occurring more often in grammatical strings, the participants had to press the A-key of the keyboard, to judge an n-gram as occurring more often in non-grammatical strings, they had to press the L-key. The n-grams were presented in a new random order for each participant.

Measures. Judgment accuracy. The judgment accuracy was quantified as the percentage of correct classifications of the 80 strings in the testing phase. As suggested by Gebauer and Mackintosh (2007), grammatical strings which were rated as “old” strings and non-grammatical strings which were rated as “new” strings were counted as correct classifications.

N-gram knowledge. The amount of n-gram knowledge was quantified as the percentage of correct classifications of n-grams in the knowledge test. Analog to the testing phase, grammatical bi- and trigrams which were rated as “old” and non-grammatical bi- and trigrams which were rated as “new” were counted as correct classifications.

Statistical analysis. The psychometric properties of the judgment accuracy in the testing phase were quantified with Cronbach’s alpha and the split-half correlation (odd-even-

split, Spearman-Brown corrected). A *t*-test was used to evaluate the null hypothesis that there was no above-chance accuracy of grammaticality judgments in the testing phase.

Results

Judgment Accuracy. In task 1, the judgment accuracy was as expected above chance, $M=61.78\%$, $t(20)=10.18$, $p<.001$, $d=2.40$, and the same was true in task 2, $M=57.80\%$, $t(20)=4.31$, $p<.001$, $d=0.95$. In task 1, Cronbach's alpha of the 80 judgments was $\alpha=.12$ and the split-half correlation was $r=.29$. In task 2, Cronbach's alpha was $\alpha=.58$ and the split-half correlation was $r=.27$. The correlation between judgment accuracy in task 1 (grammar 1, instruction to judge old vs. new) and task 2 (grammar 2, instruction to judge grammatical vs. non-grammatical) was $r=.23$, $p=.300$. A visual inspection of the frequency distributions revealed that the judgment accuracy variables were approximately normally distributed.

N-gram knowledge. In task 1, the performance in the n-gram knowledge test was $M=48.41\%$, $SD=7.74\%$. Cronbach's alpha of the twelve items of the knowledge test was $\alpha=.216$ and the split-half correlation was $r=-.31$. In this task, the correlation between judgment accuracy in the testing phase and n-gram knowledge was $r=.59$, $p=.005$. In task 2, the performance in the n-gram knowledge test was $M=66.27\%$, $SD=18.16\%$. Cronbach's alpha of the knowledge test was $\alpha=.53$ and the split-half correlation was $r=.54$. In this task, the correlation between judgment accuracy in the testing phase and n-gram knowledge was $r=-.22$, $p=.329$.

Discussion

From the perspective of cognitive psychology, this experiment successfully demonstrates an instance of implicit learning because the above-chance accuracy in the testing phase has been replicated. From an individual differences perspective, however, there are several critical points that need attention.

Reliability of judgment accuracy. The internal consistency of the judgment accuracy in task 1 was surprisingly low, which may indicate an unreliable measurement. However, according to classical test theory, Cronbach's alpha is only a point estimation of the reliability if all items are homogenous (or technically spoken have the same true score), elsewhere it is just a lower border of reliability (Lord & Novick, 1968). The same principle applies for the split half correlation, i.e. only if both test halves are homogeneous (have the same true score) the split half correlation is a point estimate of the reliability. Of course, this does not have to be true in empirical applications of the classical test theory. With respect to the present experiment, we do not know that much about the decision processes that take place, and very different judgment patterns may result in equally successful response patterns. In particular, when the participants were instructed to rate the novelty of the letter strings, some "correct" judgments (grammatical strings which were rated to be "old") were actually false alarms because none of the strings of the testing phase were previously presented in the learning phase. This may have shrunk the consistency of judgment patterns even more. To avoid this problem in the following experiment, we used a reliability estimation that is not biased by heterogeneous items or test halves, which is the retest correlation. Hence, in the following studies, 20 out of the 80 strings in the testing phase were presented repeatedly so that the retest correlation could be computed for these 20 strings.

In addition, there is another factor that may also shrink the reliability of the measurement, which is the order of presentation of the strings. This order was different for each participant and thus may have caused different effects of order for each participant, which in turn may have increased the error variance. To control this potential nuisance variable, the order of presentation of strings was fixed across participants in the following experiments.

Task consistency. The low and insignificant correlation between the two tasks replicates the results of Gebauer and Mackintosh (2007) who reported a correlation of $r=.15$ between two artificial grammar learning tasks in which the participants were asked to rate the novelty of letter strings. In the present study, the low correlation may be due to several reasons. First, the reliability of the measurements may be low and therefore the correlation between the tasks was low. Second, the two tasks did not measure the same construct because they used different artificial grammars. Third, the instruction to judge strings for novelty may measure something different than to judge for grammaticality. Clearly this low correlation cannot be interpreted just in the light of the results of experiment 1. Experiment 2 and 3 will help to clarify this point.

The relation with n-gram knowledge. There was a substantial and significant correlation ($r=.59$) between the magnitude of n-gram knowledge and the judgment accuracy in task 1, which suggests that about 35% of the variance of the novelty ratings may be explained by n-gram knowledge. On the other hand, there was no significant correlation between n-gram knowledge and the judgment accuracy in task 2, which indicates that the grammaticality ratings could not be explained by n-gram knowledge. This may be seen as preliminary evidence against the similarity of both measures.¹

Taken together, the aim of experiment 1 was to test whether asking the participants for grammaticality or novelty measures the same construct. This question could not be answered properly. It remains unclear whether the low correlation between the judgment accuracy in the two tasks was due to a low reliability of the measurements, due to the different artificial grammars, or due to the different instructions to judge either for novelty or grammaticality. Thus, two further experiments were conducted to clarify these issues.

Experiment 2

Experiment 2 was designed to follow up several questions. The first issue was to test whether the retest correlation offers greater reliability estimates than Cronbach's alpha or the split-half correlation for the performance in the testing phase. The second issue was to test whether the performance in the testing phase may be consistent across two different grammars when the participants are instructed to rate the novelty of the strings in both instances. The third issue was to test whether the performance in a third task, during which the participants are asked to rate the grammaticality of the strings, measures the same construct as the performance during the first and second task.

Method

Participants. A total of $N=21$ students from the University of Heidelberg who did not participate in experiment 1 were recruited from the campus and were paid €7 for their participation.

Stimulus material. There were three grammars. The strings of grammar 1 and grammar 2 were the same as in experiment 1. The grammatical strings for grammar 3 were constructed according to Figure 3. There were also 30 grammatical strings in the learning phase and 40 grammatical and 40 non-grammatical strings in the testing phase. The non-grammatical strings contained one violation of the grammar at random positions of the strings. The length of these strings also varied between three and eight letters (see Appendix, Table A3).

Please insert Figure 3 about here

To test the grammar knowledge of the participants, 24 n-grams were selected for each grammar. There were 12 n-grams which occurred in the learning phase and which also

occurred in the testing phase more frequently in grammatical than non-grammatical strings. These n-grams were chosen since they may help to identify grammatical strings as grammatical. In addition, there were 12 n-grams which did not occur in the learning phase but which did occur in the testing phase more frequently in non-grammatical strings than in grammatical ones. Those strings were chosen because they may help to identify non-grammatical strings. The n-grams are shown in Table 1. The strings were presented on a 17" screen of a personal computer with a standard German keyboard.

Please insert Table 1 about here

Procedure. Each participant completed three artificial grammar learning tasks. The *first artificial grammar learning task* was run with grammar 1. In the *learning phase* 30 letter strings were presented and the participants were instructed to memorize them (e.g., WNSNXS). The strings were presented in a counterbalanced order across participants. String one was presented to participant one first, string two was presented to participant two first, string three to participant three and so on. Each string was presented individually for 3 s on a 17" screen of a personal computer. The participants were asked to repeat the strings correctly by pressing the respective letters on the keyboard. When a string was repeated correctly, the feedback "correct" was given and the next string occurred. When a string was repeated incorrectly, the feedback "false" was given and the string was displayed again until repeated correctly. The feedback was given to increase the participants' motivation to memorize the strings properly. After a participant repeated ten strings correctly, these ten strings were simultaneously displayed for 90 s on the screen and the participant was asked to repeat them silently. After a participant repeated all 30 string correctly the learning phase was finished. In the *testing phase* 80 new strings were presented (see Appendix, Table A1). Ten grammatical

and ten non-grammatical strings were presented twice. These strings were randomly selected out of the original 80 strings. Thus there were a total of 100 strings in the testing phase and the retest correlation of the 20 strings could be computed. Even though all strings were new (have not been presented in the learning phase), the participants were instructed to rate the strings as “old” (presented in the learning phase) or “new” (not presented in the learning phase). To judge a string as “old”, the participants had to press the A-key of the keyboard, to judge a string as new, they had to press the L-key. The order of presentation of the strings was fixed across participants in a random order. This was done to ensure that possible effects of order would affect all participants in the same way. Immediately after the testing phase, the participants completed the n-gram knowledge test. In the *n-gram knowledge test*, the participants were instructed to judge whether an n-gram occurred more often in “old” strings or whether an n-gram occurred more often in “new” strings. To judge an n-gram as occurring more often in “old” strings, the participants had to press the A-key of the keyboard, to judge an n-gram as occurring more often in “new” string, they had to press the L-key. The order of presentation of the n-grams was fixed across participants in a random order. All n-grams were presented twice so that the retest correlation could be computed.

The *second artificial grammar learning task* was run with grammar 2. The procedures of the learning phase, the testing phase and the knowledge test were the same as in the first artificial grammar learning task.

The *third artificial grammar learning task* was run with grammar 3. The procedure of the *learning phase* was the same as in the first and the second artificial grammar learning task. After the learning phase was finished, the participants were informed that all strings in the learning phase were constructed according to a complex rule system. In the *testing phase* 80 new strings were presented. Ten grammatical and ten non-grammatical strings were presented twice. These strings were randomly selected out of the original 80 strings. Thus there were a

total of 100 strings in the testing phase. The participants were instructed to rate the strings as grammatical or non-grammatical. To judge a string as grammatical, the participants had to press the A-key of the keyboard, to judge a string as non-grammatical, they had to press the L-key. The order of presentation of the strings was fixed across participants in a random order. In the *n-gram knowledge test*, the participants were instructed to judge whether an n-gram occurred more often in grammatical strings or whether an n-gram occurred more often in non-grammatical strings. To judge an n-gram as occurring more often in grammatical strings, the participants had to press the A-key of the keyboard, to judge an n-gram as occurring more often in non-grammatical strings, they had to press the L-key. The order of presentation of the n-grams was fixed across participants in a random order. All n-grams were presented twice so that the retest correlation could be computed.

Measures. As in experiment 1, the judgment accuracy and the amount of n-gram knowledge were recorded.

Results

Judgment Accuracy. As expected, the judgment accuracy was above chance in task 1, $M=64.00\%$, $t(20)=15.79$, $p<.001$, $d=3.45$, in task 2, $M=63.62\%$ $t(20)=9.96$, $p<.001$, $d=2.18$, and in task 3, $M=57.29\%$ $t(20)=5.96$, $p<.001$, $d=1.30$. In task 1, Cronbach's alpha was $\alpha=.16$, the split-half correlation was $r=-.23$, and the retest correlation was $r=.18$. In task 2, Cronbach's alpha was $\alpha=.46$, the split-half correlation was $r=.29$, and the retest correlation was $r=.58$. In task 3, Cronbach's alpha was $\alpha=.30$, the split-half correlation was $r=-.10$, and the retest correlation was $r=.07$. The correlation between judgment accuracy in task 1 (grammar 1, instruction to judge "old" vs. "new") and task 2 (grammar 2, instruction to judge "old" vs. "new") was $r=-.18$, $p=.443$. The respective correlation between task 2 and task 3 (grammar 3, instruction to judge grammatical vs. non-grammatical) was $r=-.08$, $p=.728$. The correlation between judgment accuracy in task 1 and task 3 was $r=.58$, $p=.006$. A visual

inspection of the frequency distributions revealed that the judgment accuracy variables were approximately normally distributed.

N-gram knowledge. In task 1, the performance in the n-gram knowledge test was $M=72.94\%$, $SD=11.18\%$. Cronbach's alpha of the measured knowledge was $\alpha=.45$, the split half correlation was $r=.24$, and the retest correlation was $r=.50$. In this task, the correlation between judgment accuracy in the testing phase and n-gram knowledge was $r=.27$, $p=.245$. In task 2, the performance in the n-gram knowledge test was $M=62.70\%$, $SD=8.17\%$. Cronbach's alpha was $\alpha=-.29$, the split half correlation was $r=-.07$, and the retest correlation was $r=.49$. In this task, the correlation between judgment accuracy in the testing phase and n-gram knowledge was $r=-.10$, $p=.653$. In task 3, the performance in the n-gram knowledge test was $M=71.63\%$, $SD=11.53\%$. Cronbach's alpha was $\alpha=.37$, the split half correlation was $r=.31$, and the retest correlation was $r=.64$. In this last task, the correlation between judgment accuracy in the testing phase and n-gram knowledge was $r=.16$, $p=.477$.

Discussion

Reliability of judgment accuracy. One aim of experiment 2 was to examine whether the retest correlation provides a greater reliability estimate for the judgment accuracy than Cronbach's alpha or the split-half correlation. However, this was not the case since all reliability estimates of the judgment accuracy were rather small. Two factors may have worked against a reliable measurement. First, the instruction in task 1 and task 2 was to rate the novelty, not the grammaticality of the strings in the testing phase. Therefore it may be possible that specifically the judgment accuracy of novelty ratings is not a reliable measure. Second, the reliability estimates in the third task were also in a low range, but at that time, the participants already completed two artificial grammar learning tasks during which they got the instructions to rate the novelty of the letter strings. Although the instruction of task 3 explicitly states to rate the grammaticality of the strings, it may be possible that some

participants did not realize the change of the instruction properly whereas others did. This possibility is supported by the observation that some participants reported that it was boring to complete the same task three times after the experiment was over. To clarify this point, experiment 3 was conducted, in which the participants completed three artificial grammar learning tasks with the instruction to rate the grammaticality of the strings.

Task consistency. Another aim of experiment 2 was to check whether the performance in the testing phase may converge across two tasks when the instruction is to judge the novelty of the strings. The low correlation between task 1 and task 2 suggests that the task consistency of the measurements was low and two realizations of the same paradigm do not appear to measure the same construct. On the one hand, the estimated reliability was low and therefore the small correlation should not be overstated. On the other hand, this result is consistent with Gebauer and Mackintosh's (2007) work because they also report a low correlation between two artificial grammar learning tasks in which the participants had to judge the novelty of letter strings. Taken together, we conclude that the judgment accuracy in an artificial grammar learning task is not a consistent measure when the participants are asked to rate the novelty of the strings.

Effects of the instruction. The third aim was to test whether the performance in the testing phase quantifies the same construct regardless whether the participants are asked to judge the novelty or the grammaticality of letter strings. This was checked by the correlation between task 1 and task 3, and the correlation between task 2 and task 3. Since the reliability estimates of these measurements were low, one may not expect a high correlation between the tasks. Not surprisingly, there was no significant correlation between task 2 and task 3. However, there was a significant and unexpected high correlation between judgment accuracy of task 1 and task 3, which is not easy to explain. Sometimes a correlation between two variables may be a cue for their reliability even if other reliability estimates are low. However,

this does not seem to be plausible here because Cronbach's alpha, the split-half correlation as well as the retest correlation consistently indicated a low reliability of the measurement.²

The relation with n-gram knowledge. Similar to experiment 1, Cronbach's alpha of the knowledge measure was quite small. As discussed above, this may be due to a heterogeneous knowledge structure because a participant who learned a specific n-gram did not have to learn another n-gram necessarily. To account for that circumstance, the retest correlation was additionally computed. Since the retest correlations of the knowledge tests were in a more acceptable range (between $r=.49$ and $r=.64$), this finding suggests that the acquired knowledge was measured reliably. Moreover, the correlation between the judgment accuracy and the amount of n-gram knowledge was insignificant and rather small in all three tasks. This result is in line with the suggestion that the acquired knowledge, which affects the above-chance accuracy in the testing phase, is implicit. However, the estimated reliability for the judgment accuracy was low and may explain these small correlations as well.

Once again, experiment 2 showed rather low reliability estimates for the judgment accuracy, regardless whether Cronbach's alpha, the split-half correlation or the retest correlation was considered. The estimated consistency across tasks was also low. This speaks against the idea that artificial grammar learning tasks may be used to measure individual differences in implicit learning. However, in task 1 and task 2 the participants were asked to rate the novelty but not the grammaticality of the strings. This is a renunciation of Reber's original paradigm. Therefore experiment 3 was conducted in which the participants were asked to rate the grammaticality of strings during three tasks.

There is one additional circumstance that may have influenced the measures. In experiment 1 as well as in experiment 2, grammar 1 was first presented and grammar 2 afterwards. Therefore we cannot exclude the possibility that there were effects of grammar order that may have influenced the participants' judgment. That would be the case if the

participants still think about letter strings of grammar 1 while completing the second task. Since the judgment accuracy was significantly above chance in all tasks, this concern does not appear to be striking. However, to counteract this possible problem, the order of presentation of grammar 1 and grammar 2 was added as a between participant variable in experiment 3.

Experiment 3

Experiment 3 was conducted to test whether the judgment accuracy in the testing phase may be assessed reliably and consistently across different tasks when the participants are asked to rate the grammaticality of strings. As outlined above, there lies one difficulty in this approach. When the participants complete a second artificial grammar learning task, they already know that there is a grammar constituting the strings during the learning phase and they have to rate the grammaticality of strings in the testing phase. Hence, it may be possible that they do not only memorize the strings but also try to discover the grammar explicitly. Therefore three artificial grammar learning tasks were conducted. A change of the participants' strategy after the first task and using the same strategy for task 2 and 3 may result in a low correlation between task 1 and task 2 (as well as between task 1 and task 3) and a great correlation between task 2 and 3. In addition, to examine possible effects of order, we added the order of presentation of grammar 1 and grammar 2 as a between participant variable.

Method

Participants. The participants were $N=42$ students from the University of Heidelberg who were recruited from the campus and were paid €7 for their participation. The order of presentation was added as a between participant variable and therefore the sample size was doubled so that the power within both order conditions was the same as in experiment 2 and 3. One participant already had participated in experiment 2 and therefore was excluded from the analysis.

Stimulus material. The stimuli were the same as used in experiment 2.

Procedure. All participants completed three artificial grammar learning tasks. Half of the participants completed task 1 with grammar 1, task 2 with grammar 2, and task 3 with grammar 3 (order 1). The other half of the participants completed task 1 with grammar 2, task 2 with grammar 1, and task 3 with grammar 3 (order 2). The order of the presentation of grammar 3 was not included as a between participant variable since that would have required a larger sample size. The procedures of the learning phase, the testing phase, and the knowledge test were the same for all three artificial grammar learning tasks.

In the *learning phase* 30 letter strings were presented and the participants were instructed to memorize them. The strings were presented in a counterbalanced order across participants. String one was presented to participant one first, string two was presented to participant two first, string three to participant three and so on. Each string was presented individually for 3 s on a 17" screen of a personal computer. The participants were asked to repeat the strings correctly by pressing the respective letters on the keyboard. When a string was repeated correctly, the feedback "correct" was given and the next string occurred. When a string was repeated incorrectly, the feedback "false" was given and the string was displayed again until repeated correctly. After a participant repeated ten strings correctly, these ten strings were simultaneously displayed for 90 s on the screen and the participant was asked to repeat them silently. After a participant repeated all 30 string correctly the learning phase was finished. After the learning phase was finished, the participants were informed that all strings in the learning phase were constructed according to a complex rule system.

In the *testing phase* 80 new strings were presented. Ten grammatical and ten non-grammatical strings were presented twice. These strings were randomly selected out of the original 80 strings. Thus there were a total of 100 strings in the testing phase. The participants were instructed to rate the strings as grammatical or non-grammatical. To judge a string as

grammatical, the participants had to press the A-key of the keyboard, to judge a string as non-grammatical, they had to press the L-key. The order of presentation of the strings was fixed across participants in a random order.

In the *n-gram knowledge test*, the participants were instructed to judge whether an n-gram occurred more often in grammatical strings or whether an n-gram occurred more often in non-grammatical strings. To judge an n-gram as occurring more often in grammatical strings, the participants had to press the A-key of the keyboard, to judge an n-gram as occurring more often in non-grammatical strings, they had to press the L-key. The order of presentation of the n-grams was fixed across participants in a random order. Because of a software problem, only 18 out of the 24 n-grams were presented and all bi- and trigrams were only presented once instead of twice.

Measures. As in experiment 1 and 2, the judgment accuracy, and the amount of n-gram knowledge were recorded.

Results

Judgment accuracy. Table 2 shows the means, *t*- and *p*-values, and the effect sizes (Cohen's *d*) for the judgment accuracy. As expected, the judgment accuracy was above chance in all tasks.

Please insert Table 2 about here

Cronbach's alpha, the split-half correlation, and the retest correlation of the judgment accuracy are shown in Table 3. All coefficients are positive and considerably greater than in experiment 2.

Please insert Table 3 about here

Table 4 reports the correlations between tasks separated by order of presentation and additionally pooled over both orders of presentation. As can be seen, there was a substantial correlation between task 2 and 3, $r=.38$, $p=.014$, but not between task 1 and 2, $r=.05$, $p=.751$, or task 1 and 3, $r=.08$, $p=.631$. A visual inspection of the frequency distributions revealed that the judgment accuracy variables were approximately normally distributed.

Please insert Table 4 about here

N-gram knowledge. The performance in the first n-gram knowledge test was $M=66.12\%$, $SD=8.94\%$, the performance in the second n-gram knowledge test was $M=67.21\%$, $SD=11.90\%$, the performance in the third n-gram knowledge test was $M=64.90\%$, $SD=11.53\%$. Table 5 shows Cronbach's alpha, the split-half correlation of the measured knowledge, and the correlation between n-gram knowledge and the judgment accuracy. It is obvious from this table that there were substantial correlations between n-gram knowledge and judgment accuracy in task 2 and 3 but not in task 1.

Please insert Table 5 about here

Discussion

Reliability of judgment accuracy. The results of the present experiment suggest that individual differences may be measured reliably if the participants were asked to rate the grammaticality of strings. Most reliability estimates were in a range between 0.40 and 0.60,

which is better than the reliability estimates in experiment 2. The major difference to experiment 2 was that the participants in experiment 3 were asked to rate the grammaticality, whereas the participants in experiment 2 were asked to rate the novelty of strings. These findings suggest that individual differences may only be quantified reliably when the participants are asked to rate the grammaticality of strings. To test this hypothesis statistically, we conducted a conjoint analysis of experiment 1, 2, and 3 (see below).

Task consistency and the relation with n-gram knowledge. The results of experiment 3 suggest that the learning performance in the first artificial grammar learning task may be implicit because the judgment accuracy was significantly above chance and there was no significant relation with n-gram knowledge. The performance of the second and third task, on the other hand, may not be called implicit due to two reasons.

First, the correlation between task 1 and task 2 (or task 3) was low and insignificant, but there was a substantial and significant correlation between task 2 and task 3. This result showed up for both task orders and was even more distinct if the judgment accuracy was computed after pooling over both orders. This finding indicates that a first realization of an artificial grammar learning task seems to measure something different than a second or third realization, which may be due to the circumstance that the participants already know that there is a grammar in the tasks 2 and 3. Therefore it appears to be impossible to measure individual differences in implicit learning consistently across different tasks if the participants are asked to rate the grammaticality of strings.

Second, the correlation between n-gram knowledge in task 1 was insignificant and low ($r=.06, p=.751$), which indicates that the performance in the testing phase cannot be explained by knowledge about n-grams. However, there was a substantial and marginally significant correlation between judgment accuracy and n-gram knowledge in task 2 ($r=.30, p=.060$) and a substantial and significant correlation in task 3 ($r=.34, p=.023$). Since the judgment accuracies

in task 2 and task 3 were significantly above chance, the participants apparently learned something. However, the substantial correlation with the knowledge test indicates that this learning was not completely implicit. This finding suggests that participants process an artificial grammar learning task differently when they know that there is a grammar constituting the strings in the learning phase.

Effects of the order of the grammars. The results show that there were only minor differences in the judgment accuracies and reliability estimates depending on the order of grammar presentation. The pattern of correlations between the tasks was the same for both orders of presentation and the pattern of correlations between judgment accuracy and n-gram knowledge became even more distinct if the results were pooled over both grammars.

Taken together, experiment 3 showed that an artificial grammar learning task may be used to measure individual differences in implicit learning if the participants are asked to rate the grammaticality of letter strings. However, it was not possible to measure these differences repeatedly across different task. The correlation with the knowledge test in task 2 and task 3 also suggest that the learning that took place in task 2 and task 3 was not implicit.

Conjoint analysis

The reliability estimates of the tasks in experiment 1, 2, and 3 showed a broad variation. However, whereas all reliability estimates for novelty judgments were unacceptably small, most of the reliability estimates for grammaticality judgments were satisfactory. Therefore, we tested the hypothesis that only grammaticality judgments quantify reliable individual differences.

Method

The units of observation were the split-half correlations for each task in experiment 1, 2, and 3. These reliability estimates were employed because they could be computed in all experiments. We used the fixed-effect model for the meta-analysis of correlations of Hedges

and Vevea (1998). In a first step, the correlations were participated to a Fisher's Z -Transformation. In a second step, the transformed correlations of task 1 and 2 of experiment 2 were averaged as well as the correlations of task 1, 2, and 3 of experiment 3, because these correlations resulted from the same sample and therefore were dependent from each other. In a third step, the Z -scores were transformed to averaged effect sizes (M_Z) separately for the tasks in which the participants had to rate the grammaticality vs. the novelty of strings. In a last step, the standard errors of the effect sizes were computed. A z -test was used to test the null-hypothesis that the averaged effect sizes in both conditions did not differ from zero.

Results

The averaged effect size of the grammaticality rating tasks differed significantly from zero, $M_Z=0.32$, $z=2.75$, $p=.006$. On the other hand, the averaged effect size of the novelty rating tasks did not differ significantly from zero, $M_Z=0.09$, $z=0.56$, $p=.575$.

Discussion

The results indicate that individual differences in implicit learning may be measured reliably if the participants are asked to rate the grammaticality of the strings but not if they were asked to rate their novelty. This suggests that the variation in the judgment accuracy that is observed in grammaticality judgments quantifies systematic individual differences, whereas the variation that is observed in novelty ratings quantifies no systematic differences between individuals.

General Discussion

We conducted three experiments to investigate whether an artificial grammar learning task may be used to measure individual differences in implicit learning. The judgment accuracy in the testing phase was taken as an indicator of implicit learning success. The results of these experiments demonstrate that it is possible to measure individual differences in implicit learning when participants are asked to rate the grammaticality of strings in the

testing phase. This conclusion is supported by experiment 3 which showed that the judgment accuracy in a first realization of an artificial grammar learning task is significantly above chance and not systematically related with knowledge of n-grams. However, there are several obstacles when a *repeated* measurement of individual differences in implicit learning ought to be realized.

First, when the participants were asked to rate the grammaticality of strings, an artificial grammar learning task can only be used once. Experiment 3 shows that the performance in a second or third realization is not related with the performance in a first realization, whereas the performance in a second or third realization is related with knowledge about n-grams of letter strings. Thus, a second completion is neither task consistent, nor divergent from n-gram knowledge. Second, the instruction to rate the grammaticality of letter strings in the testing phase cannot be replaced by the instruction to rate the novelty of letter strings. Experiment 1 and 2 showed that the correlation between grammaticality and novelty judgments is small and non-significant. Moreover, reliable individual differences can only be quantified when the participants were asked to rate the grammaticality of letter strings, but not when they are asked to rate the novelty. The conjoint analysis revealed that the reliability estimates of novelty judgments did not differ significantly from zero but reliability estimates of grammaticality judgments did. Third, the reliability estimates of the grammaticality judgments are too low to make inferences about the abilities of individuals. In order to use an artificial grammar learning task as an assessment tool, its reliability needs to be enhanced.

The role of strategy use. The reliability estimates of the judgment accuracy variables were rather small. One explanation for this may be that different strategies are used to make grammaticality judgments. In particular, the participants may use implicit as well as explicit strategies to solve implicit learning tasks (as suggested by Dienes & Berry, 1997;

Norman, Price, & Duff, 2006). This may affect the reliability estimates as well as the correlation between artificial grammar learning tasks in different ways.

First, some *items* may be solved with a greater extent of implicit strategies whereas other items may be solved with a greater extent of explicit strategies. Accordingly, some items may reflect individual performance differences in implicit strategy use whereas other items may reflect individual performance differences in explicit strategy use. Technically spoken, the items may not be τ -equivalent (Lord & Novick, 1968). This may affect the reliability estimates. For example, Cronbach's alpha is a point estimate of the reliability only, if the items are τ -equivalent. Elsewise, it offers just a lower bound of the reliability. Likewise, the split-half correlation is a point estimate of the reliability only if the test-halves are τ -equivalent. Elsewise, it underestimates the reliability. It further may affect the correlation between two different artificial grammar learning tasks, since the items of one grammar may measure implicit strategies in a greater extent than the items of another grammar.

Second, *persons* may differ in the extent in which they use implicit and explicit decision strategies (e.g., Buchner, Funke, & Berry, 1995). Accordingly, the judgment accuracy of one person may indicate the success of an implicit strategy whereas the judgments accuracy of another person may indicate the success of an explicit strategy. This means, the judgment accuracy may not only capture individual differences in implicit learning performance but also individual differences in strategy use. This may additionally shrink the consistency of judgments and the correlation between artificial grammar learning tasks.

Third, the use of implicit and explicit strategies may *change* over time (as suggested by Mathews, Buss, Stanley, Blanchards-Fields, Cho, & Druhan, 1989). For example, one participant may use an implicit strategy first and then switch to a more explicit strategy later. Another participant may use an implicit strategy all the time and another participant may use an explicit strategy all the time. Accordingly, the judgment accuracy may also capture

individual differences in strategy change, which may additionally shrink the correlation between artificial grammar learning tasks.

Forth, the *instruction* to rate the novelty of the strings may induce other judgment strategies than the instruction to rate the grammaticality of strings. From a theoretical point of view, grammaticality judgments and novelty judgments may be seen as conceptually similar (Gebauer & Mackintosh, 2007; Scott & Dienes, 2008; Whittlesea et al, 1990, Whittlesea & Lobe, 2000). However, from an empirical point of view, the present results suggest that the instructions measure different constructs. Likewise, the conjoint analysis has shown that the split-half correlations are significantly above chance for the grammaticality judgments, but not for the novelty judgments. A possible explanation may be that the novelty instruction induces several independent judgment strategies, which may lead to a more heterogeneous performance variable. However, the results of the present study do not offer insights in the different judgment processes that may have been used. Identifying these processes may be a worthwhile goal for future research.

Taken together, the use of different strategies can affect the reliability estimates and the correlation between two artificial grammar learning tasks in several ways. The items may measure implicit learning success to different degrees, the participants may use implicit and explicit strategies in different extents, and the use of strategies may change over time. Furthermore, the instruction to rate the novelty of strings may induce other processes than the instruction to rate the grammaticality of strings. However, if artificial grammar learning tasks may be used as an assessment tool, then the performance variable has to be measured reliably, regardless of which strategies may be used.

Implications

Individual differences in implicit learning. Reber (1992) and Reber and Allen (2000) suggested that implicit learning is such an evolutionary old system that there are only weak

differences between individuals. However, the present study shows that individual differences in implicit learning can be measured reliably. This conforms to Mackintosh (1998) who claims that implicit learning is an ability that varies between individuals and replicates the findings of Reber et al. (1991) who also reported reliable individual differences in the performance of an artificial grammar learning task.

The present research differs from other approaches that investigated individual differences in implicit learning. In particular, the present work investigated the reliability and the task consistency of artificial grammar learning tasks, whereas previous studies investigated the relation between the implicit learning and other performance variables. Those studies have shown that the performance in artificial grammar learning tasks is rather unrelated with general intelligence (Gebauer & Mackintosh, 2007; McGeorge, Crawford, & Kelly, 1997; Pretz et al., 2010; Reber et al., 1991) or the performance in explicit learning tasks (McGeorge et al., 1997; Reber et al., 1991). These results suggest that implicit learning may measure an ability that is independent from traditional performance variables such as IQ.

However, the reliability of implicit learning tasks, such as artificial grammar learning tasks, have only sparsely been investigated, which makes these findings difficult to interpret. For example, Reber et al. (1991) suggested that an insignificant correlation between the performance in an artificial grammar learning task and an intelligence test may be taken as an indicator for the divergent validity of an implicit learning ability. In this vein, they interpreted an insignificant correlation of $r=.25$ between the performance in an artificial grammar learning task and a general intelligence test (four subscales of the WAIS-R). However, the estimated reliability of their performance measurement was only 0.51 and therefore we would not expect a large correlation of this variable with any measure of intelligence even if their true-scores have a correlation close to unity. A more realistic size of the correlation between these two measures may be in a magnitude of $r=.30$, which qualifies as a medium effect size

according to Cohen (1988). A post-hoc power analysis reveals that the power to detect a medium effect was only 0.39 in the sample of Reber et al. (1991), which had a total sample size of 20 participants. Thus, the reported insignificance is not a compelling evidence for the divergent validity of the measurement. In the same vein, McGeorge et al. (1997) and Pretz et al. (2010) reported non-significant correlations between the performance in an artificial grammar learning task and the performance in cognitive ability tests. However, the authors did not report reliability estimates for the performance in the artificial grammar learning task. The findings of the present study may suggest that the non-significant correlation may be a result of an unreliable measurement. Gebauer and Mackintosh (2007) also reported an insignificant correlation between several measures of intelligence and the performance in an artificial grammar learning task. However, Gebauer and Mackintosh asked their participants to rate the novelty of letter strings and our results show that novelty ratings measure not the same construct as grammaticality ratings. Therefore, the question how implicit learning and general intelligence is related is yet not answered properly. A fertile approach for future research may be to investigate the relation between implicit learning and other ability constructs with structural equation models. This would allow to separate systematic individual differences from unsystematic measurement error.

Individual assessment and reliability. To make inferences about the abilities of individuals, the reliability of the measurement procedure has to be improved beyond the level that has been achieved in the present study. Gebauer and Mackintosh (2007) used an item analysis to select the letter strings with the greatest item-total correlation. On the one hand, this procedure may be useful to get homogeneous items. On the other hand, the validity of the measurement may shrink because the remaining items may not be a representative sample of the underlying grammar anymore. Another approach would be to repeat the letter strings in

the testing phase and enhance the reliability this way. However, whether lengthening the test really increases the reliability or just causes fatigue or memory effects is an open issue.

The use of alternative instructions. When the participants are asked to rate the grammaticality of letter strings, an artificial grammar learning task may only be used once. The results of the conjoint analysis suggests that the performance is not reliable when the participants are asked to rate the “novelty” of letter strings. Furthermore, experiment 1 and experiment 2 revealed that individual differences in novelty ratings are unrelated with individual differences in grammaticality ratings. Thus, alternative instructions for an artificial grammar learning task may be considered. For example, Manza and Bornstein (1995; see also Helman & Berry, 2003; Zizak & Reber, 2004), suggested to use liking instead of grammaticality ratings in the testing phase because liking ratings would be a more implicit measure. This procedure would also avoid telling the participants that there is a grammar constituting the letter strings in the testing phase. However, there are no data available which would support the notion that liking ratings are a reliable and valid measurement of implicit learning.

Limitations

Measurement models of classical test theory. We used Cronbach’s alpha, the split-half correlation, and the retest correlation as reliability estimates in the present study. These estimates are based on measurement models of classical test theory which make particular assumptions (Lord & Novick, 1968). For example, Cronbach’s alpha is a point estimate of reliability only if items are τ -equivalent. Elsewise, it offers just a lower bound of reliability. Since the reliability estimates were rather low in the present samples, it would have been a worthwhile goal to test these assumptions with structural equation models. However, the use of structural equation models would have required larger sample sizes and therefore could not be realized in the present study.

Operationalisation of implicit learning. We measured implicit learning success by the judgment accuracy in the testing phase of artificial grammar learning tasks. The generalizability of the present findings rests on this particular operationalisation. However, the judgment accuracy seems to be an appropriate measure for several reasons. First, it is the standard performance measure in artificial grammar learning studies (e.g., Altmann, Dienes, & Goode, 1995; Dulany et al., 1984; Gebauer & Mackintosh, 2007; Knowlton & Squire, 1994, 1996; Meulemann & Van der Linden, 2003; Perruchet & Pacteau, 1990; Pothos & Bailey, 2000; Reber, 1967; Reber et al., 1991; Reber & Perruchet, 2003; Scott & Dienes, 2010; Tunney, 2005). Second, there were also great correlations between the overall judgment accuracy and the signal detection parameter d' in all tasks of the present study (all $r_s > .98$, $p_s < .001$). Third, there is empirical evidence for the validity of judgment accuracy as a performance measure. Dulany et al. (1984) have shown that a control group without a learning phase showed a significant worse judgment accuracy than experimental groups with a learning phase. In the same vein, Reber and Perruchet (2003) have shown that a control group which learned randomly generated stimuli performed worse in the testing phase than an experimental group which learned grammatical stimuli. Taken together, the judgment accuracy appears to be a valid indicator for implicit learning success.

The measurement of n-gram knowledge. We used an n-gram knowledge test in order to measure the amount of reportable grammar knowledge. The development of the knowledge test was inspired by the work of Perruchet and Pacteau (1990) and Dulany et al. (1984) who suggested that the participants acquire explicit knowledge of n-grams and therefore show above chance performance in the testing phase. However, implicitly learned knowledge may also help the participants to pass the n-gram knowledge test and therefore, the performance in the n-gram test may reflect explicit as well as implicit knowledge. This goes in line with several authors (e.g., Norman, Price, Duff, & Mentzoni, 2007; Seger, 1994;

Tunney & Shanks, 2003) who suggested that the participants in an artificial grammar learning tasks acquire implicit as well as explicit knowledge. Therefore, it might have been worthwhile measuring the participants' knowledge with an additional method. For example, Dienes and Scott (2005; Scott & Dienes, 2008) distinguish between structural knowledge (e.g., n-gram knowledge that indicates *why* a strings is grammatical) and judgment knowledge (the knowledge *that* a string is grammatical). To measure judgment knowledge, Dienes and colleagues (Dienes, 2008; Dienes, Altman, & Kwan, 1995; Dienes & Seth, 2010; Tunney, 2005) have suggested to use confidence ratings. In particular, they suggested that decisions that are based on unconscious, implicit knowledge should be made with low confidence (guessing criterion) and accordingly there should be no correlation between confidence ratings and accuracy (zero correlation criterion). Therefore, asking the participants to rate the confidence of their judgments would have offered further insights in participants' knowledge.

Effect of the knowledge test. In experiment 3, there were low and non-significant correlations of performance between task 1 and task 2, and between task 1 and task 3, but there was a substantial and significant correlation between task 2 and task 3. There was also a low correlation between judgment accuracy and n-gram knowledge in task 1, but substantial correlations between judgment accuracy and knowledge in task 2 and task 3. We interpreted this result as an effect of grammar awareness. During the first task, the participants do not know that there is a grammar constituting the letter strings, but they do so during a second and third task. However, this finding could also be interpreted as an effect of the knowledge test. After completing a knowledge test, the participants may draw their attention towards n-grams and this may affect their judgments in subsequent tasks. However, if this would have been the case, the same pattern of results should have been found in experiment 1 and experiment 2, which was not the case. Nonetheless, it might be a worthwhile goal for future research to investigate possible effects of the knowledge test in greater detail.

Conclusion

We demonstrated that an artificial grammar learning task can be used to measure individual differences in implicit learning. Future research may investigate whether lengthening the test may substantially increase reliability, whether the use of a liking instruction may allow to perform several realizations of the task, and how individual differences in implicit learning are related to intelligence, educational attainment or even professional success in later life. The present study provides the empirical basis for pursuing these questions.

References

- Altmann, G. T. M., Dienes, Z. n., & Goode, A. (1995). Modality independence of implicitly learned grammatical knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 899-912.
- Buchner, A., Funke, J., & Berry, D. C. (1995). Negative correlations between control performance and verbalizable knowledge: Indicators for implicit learning in process control tasks? *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *48*, 166-187.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J.: Erlbaum.
- Dienes, Z. (2008) Subjective measures of unconscious knowledge. *Progress in Brain Research*, *168*, 49 - 64.
- Dienes, Z., Altmann, G. T. M., Kwan, L., & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 1322-1338. doi: 10.1037/0278-7393.21.5.1322
- Dienes, Z., & Berry, D. (1997). Implicit learning: Below the subjective threshold. *Psychonomic Bulletin & Review*, *4*, 3-23.
- Dienes, Z., & Scott, R. (2005). Measuring unconscious knowledge: Distinguishing structural knowledge and judgment knowledge. *Psychological Research*, *69*, 338-351. doi: 10.1007/s00426-004-0208-3

- Dienes, Z., & Seth, A. (2010). Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness and Cognition: An International Journal*, *19*, 674-681. doi: 10.1016/j.concog.2009.09.009
- Dulany, D. E., Carlson, R. A., & Dewey, G. I. (1984). A case of syntactical learning and judgment: How conscious and how abstract? *Journal of Experimental Psychology: General*, *113*, 541-555. doi: 10.1037/0096-3445.113.4.541
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.
- Frensch, P. A. (1998). One concept, multiple meanings: On how to define the concept of implicit learning. In M. A. Stadler & P. A. Frensch (Eds.), *Handbook of implicit learning*. (pp. 47-104). Thousand Oaks, CA: Sage Publications.
- Frensch, P. A., & Rüniger, D. (2003). Implicit learning. *Current Directions in Psychological Science*, *12*, 13-18. doi: 10.1111/1467-8721.01213
- Gebauer, G. F., & Mackintosh, N. J. (2007). Psychometric intelligence dissociates implicit and explicit learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 34-54. doi: 10.1037/0278-7393.33.1.34
- Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486-504.
- Helman, S., & Berry, D. C. (2003). Effects of divided attention and speeded responding on implicit and explicit retrieval of artificial grammar knowledge. *Memory & Cognition*, *31*, 703-714.

- Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition, 116*, 321-340. doi: 10.1016/j.cognition.2010.05.011
- Kinder, A., Shanks, D. R., Cock, J., & Tunney, R. J. (2003). Recollection, Fluency, and the Explicit/Implicit Distinction in Artificial Grammar Learning. *Journal of Experimental Psychology: General, 132*, 551-565. doi: 10.1037/0096-3445.132.4.551
- Knowlton, B. J., & Squire, L. R. (1994). The information acquired during artificial grammar learning. *Journal of experimental Psychology: Learning, Memory, and Cognition, 20*, 79-91.
- Knowlton, B. J., & Squire, L. R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of experimental Psychology: Learning, Memory, and Cognition, 22*, 169-181.
- Lord, F. M. & Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Oxford: Addison-Wesley.
- Mackintosh, N. J. (1998). *IQ and human intelligence*. Oxford Oxford University Press.
- Manza, L., & Bornstein, R. F. (1995). Affective discrimination and the implicit learning process. *Consciousness and Cognition: An International Journal, 4*, 399-409.
- Mathews, R. C., Buss, R. R., Stanley, W. B., Blanchard-Fields, F., Cho, J. R., & Druhan, B. (1989). Role of implicit and explicit processes in learning from examples: A synergistic effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 1083-1100. doi: 10.1037/0278-7393.15.6.1083
- McGeorge, P., Crawford, J. R., & Kelly, S. W. (1997). The relationships between psychometric intelligence and learning in an explicit and an implicit task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 239-245. doi: 10.1037/0278-7393.23.1.239

- Meulemans, T., & Van der Linden, M. (2003). Implicit learning of complex information in amnesia. *Brain and Cognition, 52*, 250-257.
- Norman, E., Price, M. C., & Duff, S. C. (2006). Fringe consciousness in sequence learning: The influence of individual differences. *Consciousness and Cognition: An International Journal, 15*, 723-760. doi: 10.1016/j.concog.2005.06.003
- Norman, E., Price, M. C., Duff, S. C., & Mentzoni, R. A. (2007). Gradations of awareness in a modified sequence learning task. *Consciousness and Cognition: An International Journal, 16*, 809-837. doi: 10.1016/j.concog.2007.02.004
- Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General, 119*, 264-275. doi: 10.1037/0096-3445.119.3.264
- Pothos, E. M., & Bailey, T. M. (2000). The role of similarity in artificial grammar learning. *Journal of experimental Psychology: Learning, Memory, and Cognition, 26*, 847-862.
- Pretz, J. E., Totz, K. S., & Kaufman, S. B. (2009). The effects of mood, cognitive style, and cognitive ability on implicit learning. *Learning and Individual Differences*. doi: 10.1016/j.lindif.2009.12.003
- Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning & Verbal Behavior, 6*, 855-863. doi: 10.1016/s0022-5371(67)80149-x
- Reber, A. S. (1992). The cognitive unconscious: An evolutionary perspective. *Consciousness and Cognition: An International Journal, 1*, 93-133.
- Reber, A. S., & Allen, R. (1978). Analogic and abstraction strategies in synthetic grammar learning: A functionalist interpretation. *Cognition, 6*, 189-221. doi: 10.1016/0010-0277(78)90013-6

- Reber, A. S., & Allen, R. (2000). Individual differences in implicit learning: Implications for the evolution of consciousness. In R. G. Kunzendorf & B. Wallace (Eds.), *Individual differences in conscious experience*. (pp. 227-247). Amsterdam: John Benjamins Publishing Company.
- Reber, A. S., Walkenfeld, F. F., & Hernstadt, R. (1991). Implicit and explicit learning: Individual differences and IQ. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 888-896. doi: 10.1037/0278-7393.17.5.888
- Reber, R., & Perruchet, P. (2003). The use of control groups in artificial grammar learning. *The Quarterly Journal of experimental Psychology A: Human experimental Psychology*, *56*, 97-115.
- Scott, R. B., & Dienes, Z. (2008). The conscious, the unconscious, and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1264-1288. doi: 10.1037/a0012943
- Scott, R. B., & Dienes, Z. (2010). Fluency does not express implicit knowledge of artificial grammars. *Cognition*, *114*, 372-388.
- Seger, C. A. (1994). Implicit learning. *Psychological Bulletin*, *115*, 163-196. doi: 10.1037/0033-2909.115.2.163
- Servan-Schreiber, E. & Anderson, J. R. (1990). Chunking as a mechanism of implicit learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 592-608.
- Tunney, R. J. (2005). Sources of confidence judgments in implicit cognition. *Psychonomic Bulletin & Review*, *12*, 367-373.
- Tunney, R. J., & Shanks, D. R. (2003). Subjective measures of awareness and implicit cognition. *Memory & Cognition*, *31*, 1060-1071.

Whittlesea, B. W., Jacoby, L. L., & Girard, K. (1990). Illusions of immediate memory:

Evidence of an attributional basis for feelings of familiarity and perceptual quality.

Journal of Memory and Language, 29, 716-732. doi: 10.1016/0749-596x(90)90045-2

Whittlesea, B. W. A., & Leboe, J. P. (2000). The heuristic basis of remembering and

classification: Fluency, generation, and resemblance. *Journal of Experimental*

Psychology: General, 129, 84-106. doi: 10.1037/0096-3445.129.1.84

Zizak, D. M., & Reber, A. S. (2004). Implicit preferences: The role(s) of familiarity in the

structural mere exposure effect. *Consciousness and Cognition: An International*

Journal, 13, 336-362.

Footnotes

¹ The performance in the second n-gram knowledge test was significantly above chance ($t(20)=4.11, p=.001, d=0.90$). This result suggests that the participants acquired n-gram knowledge. However, it does not indicate that the participants used this knowledge for making grammaticality judgments. Only a positive correlation between the performance in the testing phase and the performance in the n-gram knowledge test would suggest that the participants used their n-gram knowledge for making grammaticality judgments.

² One possible explanation for this result may be a greater similarity between grammar 1 and grammar 3. In particular, a detailed inspection of the grammatical strings revealed that the strings of grammar 1 and grammar 3 may be more similar to each other than the strings of grammar 1 and grammar 2 or grammar 2 and grammar 3.

Table 1

N-grams used in the knowledge test

grammar	grammatical n-grams	non-grammatical n-grams
1	NS NWS NWX NX NXS NXT SN SSS ST WNS XS XT	NN NNW NNX NWN NWW SSW WNW WWN WWS XN XWX XX
2	HHH HL HPH LM LMM LRH ML MM MMM PR RH RHP	HHP HPL HPM LH LMH LPH LRM MHM MP PHP PLR RM
3	BG BGK BK BKD DG GD GDF GFD GK GKD KD KFD	BF DD DFF DK FDF FFD FGG GFF GGF KDD KFF KK

Note. Grammatical n-grams are n-grams that occurred in the learning phase and which occurred in the testing phase more often in grammatical than in non-grammatical strings. Non-grammatical n-grams are n-grams that did not occur in the learning phase and which occurred in the testing phase more often in non-grammatical than in grammatical strings.

Table 2

Judgment accuracy in experiment 3

task	grammar	<i>M</i>	<i>t</i>	<i>df</i>	<i>p</i>	<i>d</i>
1	1	59.67%	5.46	20	.000	1.19
1	2	62.85%	10.26	19	.000	2.29
1	1+2	61.22%	10.12	40	.000	1.58
	(pooled)					
2	1	61.75%	7.40	19	.000	1.65
2	2	61.76%	7.61	20	.000	1.66
2	1+2	61.76%	10.75	40	.000	1.68
	(pooled)					
3	3	59.15%	8.88	40	.000	1.39

Table 3

Cronbach's alpha, split-half, and retest correlation of the judgment accuracy in experiment 3

task	grammar	α	r_{sh}	r_{tt}
1	1	.66	.43	.58
1	2	.56	.54	.41
1	1+2 (pooled)	.55	.38	.43
2	1	.56	.54	.46
2	2	.32	.22	.46
2	1+2 (pooled)	.54	.60	.44
3	3	.49	.45	.18

Note. α = Cronbach's alpha, r_{sh} = split-half correlation, r_{tt} = retest correlation.

Table 4

Correlation between tasks separated by order of presentation in experiment 3.

	order 1		order 2		pooled	
	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>
task 1 – task 2	-.07	.759	.24	.314	.05	.715
task 1 – task 3	-.02	.908	.14	.560	.08	.631
task 2 – task 3	.41	.067	.39	.089	.38	.014

Note. In order 1 the participants completed grammar 1 first and then grammar 2 and grammar 3. In order 2 the participants completed grammar 2 first and then grammar 1 and grammar 3.

Table 5

Cronbach's alpha of n-gram knowledge and correlation with the judgment accuracy in experiment 3

task	grammar	α	r_{sh}	r
1	1	-.61	-.15	.24 (.303)
1	2	-.08	-.08	-.09 (.704)
1	1+2 (pooled)	-.42	-.11	.06 (.715)
2	1	.49	.40	.41 (.076)
2	2	-.12	.05	.16 (.493)
2	1+2 (pooled)	.24	.25	.30 (.060)
3	3	.20	.09	.34 (.023)

Note. α = Cronbach's alpha of measured knowledge, r_{sh} = split-half correlation, r =

Pearson correlation between n-gram knowledge and judgment accuracy (p -values in brackets).

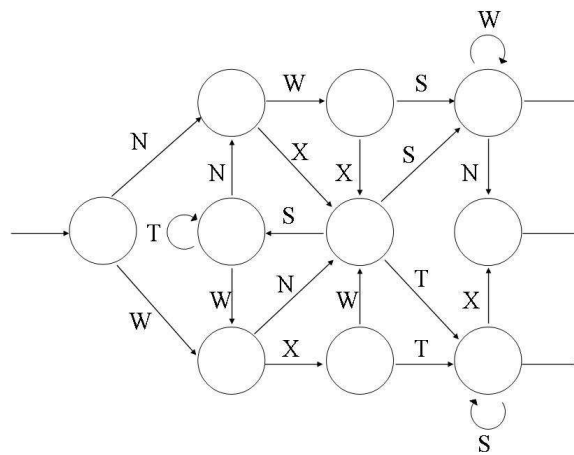


Figure 1: Grammar 1

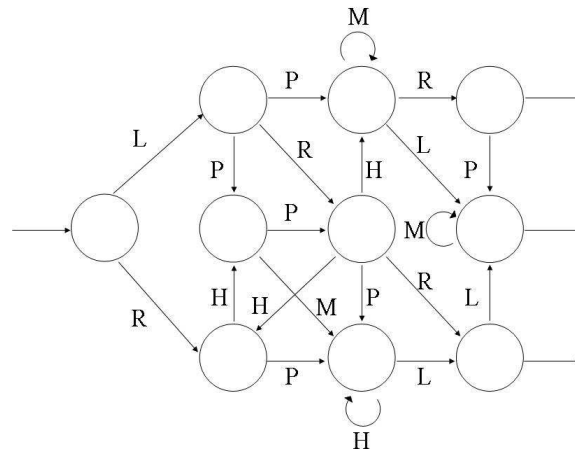


Figure 2: Grammar 2

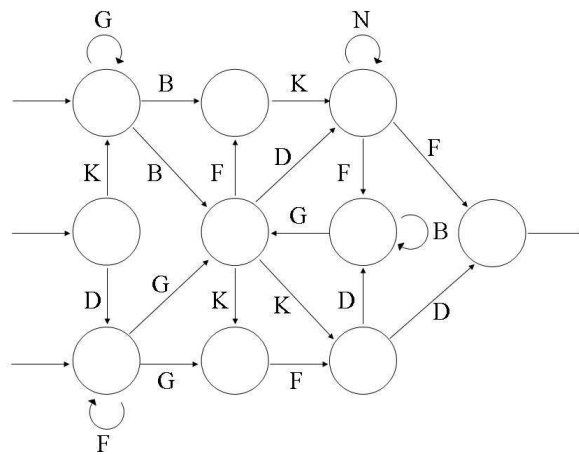


Figure 3: Grammar 3

Appendix

Table A1

Letter strings for grammar 1 sorted for different parts of the experiment

Phase	Strings
Learning phase	<p>WNSNXS NXSTWXT WNSTTWXT WXWSNXT NWXTS NXSNWXS</p> <p>WNTSSS NXSWXTX WXWSNWSN WNSNWXS NXSWXT NXSWWWW</p> <p>NWSWN WNSNWXTX NXSWNTX NXS WNSWWWW WXWSN</p> <p>NXSTNWS WNSTWNSW WNSWXTX WNTSSX WNSNWSW WNTX</p> <p>NXSWNSW WNSNXTS NXTSSS NXSNWSN WNSNWSN NXSWNTSS</p>
Testing phase (correct items)	<p>WNSN NWSW NWSN NXSWW NWXSX NXTSX WNSWNS NXSWNT</p> <p>NWXTSS WNSWWW WNSWNT NXTSSX WXWTSX WNSNXT</p> <p>NWSWWN NXSNWS NXSNWXT NXTSSSX WNSTWNS NXSTNXS</p> <p>WNTSSSX WNSWXWT NXSNXTX WXWSWXT NWXSWS</p> <p>NWXSNS NXSTXWNT WNSTNWXT WXWSNWXS NXSTWNTX</p> <p>WXWSTWXT WXWSNWSW WXWSWXTX NXSTNWSN NWXSXWXS</p> <p>WXWTSSSS WNSTNXTX WXWSWNSW NWXSXWXTX NXSNXSWN</p>
Testing phase (incorrect items)	<p>TXSWNT TWXTSX NTSWWN WWSWNS WNWWNT NWSXWN</p> <p>NWXSSW WXWTST TXWTSSX SWXSWNS WSSWWW NWSWXTS</p> <p>NWWSWXT NXNTNXS WNTTSSX NWXWSSX NWXSNTS WNSNXXX</p> <p>WXWSWST WNSWXWN WXWSWNW XNSTWNTS TWXSXWXS</p> <p>TWSWWWWN NNXSWXWT WSSTTNXT WNNWNSWW WNNNWXS</p> <p>NWXWWXTX WXWXWXTX WXWSXWXT WXWSNWSW</p> <p>NXSWWXWN WXWSWNWW WXWSNWS</p>

Table A2

Letter strings for grammar 2 sorted for different parts of the experiment

Phase	Strings
Learning phase	LRHMMLM LRPHELLM RHPHR RHPHMMLM LRHL LPMHLLMM LPPHLM RHPRLMMM LRHMRP RHPHMMP LPPPL RPHHLLM RPHL LPPRLMMM LPR LRHRPMMM LPPRL LPMMRPMM RHPHRP LPMHLLM LPMMRP RHMHLLMM LPLM RPHHHLL LRR LRRLMMM RHMHLL LPPRLMM RPLMMM RPHLMM
Testing phase (correct items)	LRPHHL RHMHHLL LRHMLMMM LPRP LPRPMM LPRPMMM LPLMMMM RHPHMML LPMR LPMRPM RPHHLL LPPHMLM LPPHMMP RPHL LRHLMM RHPHMML RPHLMMM LPMLLMMM LPLMM LRHMML RPHLLMM LPMHHLL LPPHMMM RPHLM LPPHMML RHMLLMM RPLMMMM LPPHLLM LPMML LPMLLMM RHMHLLM RPHMLMM LPRPMMM LPLMM RPHMLM RHPRLMM LRPHHLL RHPHLLM
Testing phase (incorrect items)	RPRL LLRPMM RHPHHL RHMHHPL LRPHHLL LPLR LPPMRP LPHMMR RHPRLMH LPMHHPL HHMLL LPLRMM RPPLMM PPLMMMM LRHMHPM LPHHL LPMMHM LPPLRPM PHPHMML LPPHMHM LPPL RPHHPL RPHHLL MPPHMMP LPPHLRM LPLMP LPMMP LPPHML LPHMMMP RHMHHLL HRHLMM MHPPHL LPPHRP LPPMRPM LPMLLMP LLMHLL RMPPLM LPPMHM LPPLHHLL RHPHLLL

Appendix A2

Manuscript 2: Can artificial grammar learning tasks measure individual differences in implicit learning?

Can artificial grammar learning tasks measure
individual differences in implicit learning?

Daniel Danner, Dirk Hagemann, Andrea Schankin, Marieke Hager, & Joachim Funke

Institute of Psychology, University of Heidelberg, Germany

Author Note

This research was funded by German Research Foundation Grant DFG, Ha 3044/7-1.

We gratefully thank Andreas Neubauer, Anna-Lena Schubert, and Katharina Weskamp for conducting the experiment.

Correspondence concerning this paper should be addressed to Daniel Danner,
University of Heidelberg, Institute of Psychology, Hauptstrasse 47-51, D-69117 Heidelberg,
Germany, Phone: +49 (0) 6221-547354, Fax: +49 (0) 6221-547325, Email:
daniel.danner@psychologie.uni-heidelberg.de

Abstract

The present study investigates whether artificial grammar learning tasks can measure individual differences in implicit learning. In particular, we investigated (1) the reliability and the task consistency of implicit learning performance, (2) the association between implicit learning performance, reportable grammar knowledge, and general intelligence, and (3) whether implicit learning performance can predict educational attainment. $N=106$ participants completed two artificial grammar learning tasks and the Culture Fair Intelligence Test. The results indicate that the reliability of the performance measure is only moderate and the task consistency is adequate as long as no bigram knowledge test is performed. Artificial grammar learning performance is independent from reportable grammar knowledge and independent from general intelligence. Furthermore, there is a predictive but not an incremental predictive value on educational attainment.

Keywords: implicit learning, artificial grammar learning, individual differences, reliability, validity

Can artificial grammar learning tasks measure individual differences in implicit learning?

Sometimes we make correct decisions based on our gut feeling but can not explain them. Regarding this, several authors suggested that we can learn implicitly, which means without intention and awareness (Frensch & Rüniger, 2003; Reber, 1967, 1992; Seger, 1994). For example, sometimes we are able to classify a sentence to be grammatical correct or incorrect but we are not able to report the determining grammatical rule. Relating to this, Reber (1967) suggested that we may learn complex rules implicitly. He further suggested that implicit learning is an evolutionary mechanism that is independent from explicit learning (Reber, 1992; Reber & Allen, 2000). In the same vein, Mackintosh (2006) proposes an implicit associative learning system, and an explicit hypothesis generating and testing system. In particular, the implicit learning system may detect contingencies *without* awareness or intention whereas the explicit learning system is necessary for discovering regularities *with* intention and awareness. Mackintosh hypothesized that individual differences in implicit learning are independent from general intelligence but powerful predictors of educational success. In order to test this hypothesis, it is necessary to measure individual differences in implicit learning. For one thing, implicit learning may help to characterize cognitive ability in greater detail. For another thing, implicit learning measures may be used as selection criteria for university or job applications.

To measure implicit learning, Reber and Mackintosh suggested to use artificial grammar learning tasks. In such a task, the participants are asked to learn a list of arbitrary letter strings (like KTQHXTJ). Afterwards they are told that these strings were constructed according to a complex rule system (grammar) and they are asked to judge new strings as grammatical or non-grammatical. The percentage of correct judgments is taken as an indicator for implicit learning success. Typically, the participants show above chance performance

which suggests that they learned something but they are not able to report the grammar rules, which suggests that they learned the rules implicitly.

In order to use an artificial grammar learning task for individual assessment or the investigation of individual differences, the performance measures must meet several psychometric criteria. (1) The reliability of performance measures should be acceptable. (2) Measures should be independent from reportable knowledge to attest that the learning performance is implicit. (3) Implicit learning measures should be divergent from general intelligence to attest their divergent validity. (4) Implicit learning performance should be related with real life performance to reveal its predictive validity. (5) Performance should be task consistent, meaning measureable with more than one artificial grammar learning task in order to establish sufficient generalizability. There have been only sparse attempts to investigate the psychometric properties of artificial grammar learning measures. Therefore, the purpose of the present work was to evaluate these five issues.

(1) Reliability. There are only few studies that investigated the reliability of artificial grammar learning measures. Reber, Walkenfeld, and Hernstadt (1991) examined $N=20$ students and reported a Cronbach's alpha of $\alpha=.51$ for 100 grammaticality judgements. Gebauer and Mackintosh (2007) assessed $N=605$ pupils and reported a split-half correlation of $r=.70$ for two artificial grammar learning tasks with 80 grammaticality judgements each. In addition, Danner, Hagemann, Schankin, Bechtold, and Funke (submitted) conducted a series of experiments with a total of $N=83$ students and reported Cronbach's alphas between $\alpha=.32$ and $\alpha=.66$ for grammaticality judgments in different artificial grammar learning tasks. These findings suggest that the performance scores of individuals should be interpreted carefully and the reliability of implicit learning performance variables should be taken into account when interpreting correlations with other variables. However, the previous findings may also be a

specific feature of the grammars that have been used. For the purpose of the present study, we developed two new grammars and investigated the reliability of the performance measures.

(2) Relation with reportable knowledge. Reber (1967) suggested that the participants in an artificial grammar learning task learn the grammar rules implicitly because they are not able to report their grammar knowledge. However, to test whether grammaticality judgments are independent from reportable knowledge, it is necessary to define what kind of knowledge is relevant for the performance in artificial grammar learning tasks. Over the years, there have been controversial and fertile discussions about this topic. For example, Reber and Allen (1978) found that their participants were not able to report any knowledge about grammar rules and therefore suggested that they learned the grammar rules implicitly. Dulany, Carlson, and Dewey (1984) criticized that asking participants to report the grammar rules is too difficult and therefore the participants might not have been able to report their knowledge. To avoid this problem, Dulany et al. (1984) asked their participants to report letter string features on which they based their grammaticality judgments. They showed that the reported knowledge was sufficient to explain the above chance accuracy of grammaticality judgments and concluded that the acquired knowledge was not implicit at all. In a similar vein, Perruchet and Pacteau (1990) showed that knowledge of bigrams was sufficient to explain the above chance accuracy of grammaticality judgments. Other authors (e.g., Knowlton & Squire, 1996) suggested that the participants make grammaticality judgments based on the similarity of letter strings with previously learned strings. Having these different explanation attempts in mind, it seems difficult to find an appropriate measurement for the relevant knowledge. Shanks and St. John (1994) concluded that it is only possible to measure the relevant knowledge for implicit learning tasks, when the information criterion and the sensitivity criterion are met. Meeting the information criterion means to find an operationalisation that captures all kind of relevant knowledge. Meeting the sensitivity

criterion means to make the knowledge test as sensitive as the implicit learning task itself. Thus, to investigate the relation between implicit learning performance and reportable knowledge in the present study, we used a knowledge test that was designed to meet the information as well as the sensitivity criterion.

(3) Relation with general intelligence. Reber et al. (1991) reported a correlation of $r=.25$ between the performance in an artificial grammar learning task and IQ. Gebauer and Mackintosh (2007) reported correlations between $r=-.03$ and $r=.17$ depending on the task and the instruction. Hence, there is preliminary evidence pointing towards the divergent validity of implicit learning measures. A further aim of the present study was to replicate these findings.

(4) Predictive value. From a practical point of view, the most important characteristic of a measure may be its predictive value. Mackintosh (2006) hypothesizes that performance in artificial grammar learning may be a powerful predictor of educational attainment. However, there are no investigations of this hypothesis yet. Therefore, the present study will test whether the performance in an artificial grammar learning task can predict educational success.

(5) Task consistency. A further purpose of the present work was to evaluate the task consistency of performance measures. This is of particular importance within the framework of artificial grammar learning tasks. During an artificial grammar learning task, the participants are asked to memorize a series of arbitrary letter strings. Only after this learning phase, they will be informed that there was a grammar constituting the strings and their task will be to classify new letter strings as grammatical or non-grammatical. During a subsequent artificial grammar learning task, the participants will already know that there is a grammar constituting the strings in the learning phase and that his or her job will be to rate the grammaticality of letter strings afterwards. Hence, it may be that the participants do not only

memorize the strings but also try to discover the grammar explicitly. To investigate this hypothesis, Danner et al. (submitted) performed an experiment where the participants completed three artificial grammar learning tasks with a subsequent knowledge test after each grammar learning task. They reported that there was no correlation between the performance in the first and the second artificial grammar learning task, which indicates a low task consistency of artificial grammar learning task measures. Likewise, the participants' performance in a first task was unrelated with the reported grammar knowledge whereas the performance in subsequent tasks correlated with the reported grammar knowledge. This finding suggests that the learning process in the second and third artificial grammar learning task was not implicit anymore. However, there is also an alternative interpretation of the results of Danner et al. (submitted). Their participants completed a knowledge test (containing bi- and trigrams of letter strings) after every artificial grammar learning task. Therefore, it is also possible that the knowledge test and not the grammar awareness changed the participants' strategy and caused the low task consistency as well as the relation with reported knowledge. A further aim of the present study was to test the hypothesis that a knowledge test decreases the task consistency between two artificial grammar learning tasks and causes a substantial correlation between performance and reported grammar knowledge. Thus, in two separate conditions the participants completed either a grammar knowledge test or a dummy knowledge test.

Aim of the present study. The aim of the present study was to evaluate the reliability and the validity of artificial grammar learning measures. Therefore, we investigated the reliability and the task consistency of performance measures as well as the relation with reportable knowledge, general intelligence and educational attainment.

Method

Participants

There were $N=106$ students of the University of Heidelberg participating in the present study. The participants were randomly assigned to either the bigram group ($N=53$) or the control group ($N=53$).

Procedure

All participants completed first an artificial grammar learning task, second a knowledge test, third the Culture Fair Intelligence Test (CFT3), and fourth an additional artificial grammar learning task and a further knowledge test.

The first artificial grammar learning task. The stimuli for the first artificial grammar learning task were constructed according to Figure 1. The task consisted of a learning phase and a testing phase.

Please insert Figure 1 about here

In the *learning phase* 39 letter strings were presented and the participants were instructed to memorize them. Each string was presented individually for 3 s on a 17" screen of a personal computer (e.g. KTQHXTJ). The participants were asked to repeat the strings correctly by pressing the respective letters on the keyboard. When a string was repeated correctly, the feedback "correct" was given and the next string occurred. When a string was repeated incorrectly, the feedback "false" was given and the string was displayed again until repeated correctly. After a participant repeated ten strings correctly, these ten strings were simultaneously displayed for 90s on the screen and the participant was asked to repeat them silently. After a participant repeated all 39 string correctly the learning phase was finished and

the participant was informed that all strings in the learning phase were constructed according to a complex rule system.

In the *testing phase* 78 new strings were presented (see Appendix, Table A1). There were 39 grammatical strings that were constructed according to the same rule system as the strings in the learning phase (e.g. KXTJTTH). In addition, there were 39 non-grammatical strings that contained one letter at a position that violated the rule system (e.g. KXTXJK). All strings were presented twice so that there was a total of 156 items in the testing phase. The participants were instructed to judge the letter strings as grammatical or non-grammatical. To judge a string as grammatical, the participants had to press the A-key of the keyboard, to judge a string as non-grammatical, the L-key. The order of presentation of the strings was fixed across the participants in a random order. The percentage of correct judgments in the testing phase was taken as the performance indicator for implicit learning success.

The first knowledge test. Immediately after the testing phase, the participants completed a knowledge test. The bigram group completed a bigram knowledge test and the control group completed a dummy knowledge test

The *bigram knowledge test* assessed participants' knowledge of bigrams. To meet the *information criterion* (Shanks & St. John, 1994), we designed the bigram knowledge test in a manner that the test was sensitive to different forms of knowledge. In particular, the test was a direct test of participants' knowledge of bigrams as well as an indirect test of participants' performance relevant knowledge in general. For example, one participant may have acquired knowledge of bigrams during the learning phase (as suggested by Perruchet & Pacteau, 1990) and therefore achieved above chance accuracy in the testing phase as well as in the bigram knowledge test. Another participant may have used the similarity between previously learned and new strings (as suggested by Knowlton & Squire, 1996) and thus achieved above chance accuracy in the testing phase. However, the knowledge about the similarity of strings would

also help the participant to perform well in the knowledge test. In order to meet the *sensitivity criterion*, we made the instructions and response format analogous to the testing phase (as suggested by Shanks & St. John, 1994).

Thus, all strings of the testing phase were decomposed into bigrams. For example, KXTJTTH was decomposed into KX, XT, TJ, JT, TT, and TH. The participants were instructed to rate a bigram as grammatical (occurring more often in grammatical strings) or non-grammatical (occurring more often in non-grammatical strings). To judge a bigram as grammatical, the participants had to press the A-key. To judge a bigram as non-grammatical, the participants had to press the L-key. There were 34 different bigrams for grammar 1 (see Appendix, Table A1). All bigrams were presented twice so that there were a total of 68 items in the bigram knowledge test. The order of presentation of the strings was fixed across the participants in a random order. The percentage of correct judgments in the bigram knowledge test was taken as an indicator for the amount of reportable knowledge.

In order to make the procedure for the bigram and the control group parallel, the control group completed a *dummy knowledge test* which was unrelated with the letter strings. The dummy knowledge test consisted of statements like “Alberto Fujimori was president of Japan from 1990 to 2000” (which is right, by the way) and the participants were asked to rate the truth of the statements. To rate a string as true, the participants had to press the A-key of the keyboard, to rate a string as false, the L-key. There were 34 different statements and all statements were presented twice so that there were a total of 68 items in the dummy knowledge test. Participants’ responses in the dummy knowledge test were not analyzed.

The Culture Fair Intelligence Test (CFT3). The CFT3 (Cattell, Krug, & Barton, 1973) was used as an indicator for participants’ general intelligence. The test consists of 48 different figural reasoning items. The speed version of the test was administered, which took

approximately 25 minutes. The number of correctly solved items was taken as the performance indicator for participants' general intelligence.

The second artificial grammar learning task. The stimuli for the second artificial grammar learning task consisted of completely different letters. The stimuli were constructed according to Figure 2. The second artificial grammar learning task also consisted of a learning phase and a testing phase. The procedure was identical to the first artificial grammar learning task.

Please insert Figure 2 about here

The second knowledge test. The procedure for the second knowledge test was identical to the first with the exception that all participants completed a bigram knowledge test after the testing phase and the bigram knowledge test consisted of 34 items only. The stimuli are shown in the Appendix (Table A2).

In addition, the participants were asked to report their final school exams' grade point average (1=very good to 6=failed) and their subject of study (psychology vs. other subject).

Results

Performance in artificial grammar learning tasks

To investigate the effect of the first bigram knowledge test, we analyzed the data separately for the bigram group and the control group. In the *bigram group*, the percentage of correct judgments in the first artificial grammar learning task was significantly above chance, $M=58.09\%$, $t(52)=7.40$, $p<.001$. The split-half correlation (between the first and the second presentation of strings) was $r=.60$. The percentage of correct judgments in the second artificial grammar learning task was also above chance, $M=56.98\%$, $t(52)=11.02$, $p<.001$ and the split-

half correlation was $r=.32$. The correlation between both tasks was $r=.22$, $p=.109$, which points towards a low task consistency.

In the *control group*, the performance in the first artificial grammar learning task was also significantly above chance, $M=59.62\%$, $t(52)=10.68$, $p<.001$, and the split-half correlation was $r=.39$. The percentage of correct judgments in the second artificial grammar learning task was $M=57.28\%$, $t(52)=12.26$, $p<.001$, and the split-half correlation was $r=.21$. The correlation between both tasks was $r=.39$, $p=.004$, which points towards an adequate task consistency.

To investigate the effect of the knowledge test on the task consistency in greater detail, we tested whether the correlation between the first and the second artificial grammar learning task differed between the bigram group and the control group. As expected, the task consistency in the control group was greater, $r=.39$, $Z=.39$, than in the bigram group, $r=.22$, $Z=.22$. However, the difference between groups was not significant, $z=0.94$, $p=.347$.

Reportable knowledge

In the *bigram group*, the percentage of correct judgments in the first bigram knowledge test was significantly above chance, $M=55.45\%$, $t(52)=5.01$, $p<.001$. The split-half correlation was $r=.55$. The correlation between the performance in the first testing phase and the first knowledge test was $r=.01$, $p=.942$. The percentage of correct judgments in the second bigram knowledge test was also significantly above chance, $M=55.03\%$, $t(52)=5.50$, $p<.001$. The split-half correlation was $r=.32$. The correlation between the performance in the second testing phase and the second knowledge test was $r=.30$, $p=.029$.

In the *control group*, there was no bigram knowledge test after the first artificial grammar learning task. The percentage of correct judgments in the bigram knowledge test after the second artificial grammar learning task was significantly above chance, $M=56.40\%$,

$t(52)=6.79, p<.001$. The split-half correlation was $r=.33$. The correlation between the performance in the testing phase and the knowledge test was $r=.02, p=.884$.

Relation with general intelligence

To investigate the relation between implicit learning and general intelligence we took the performance in the first artificial grammar learning task as an indicator for implicit learning performance. The procedure for the bigram group and the control group was identical until the completion of the first artificial grammar learning task and therefore the data of both groups were analyzed together. The number of solved items in the CFT3 served as a measure of participants' general intelligence.

The mean number of solved items in the CFT3 was $M=28.69 (SD=4.76)$. Cronbach's alpha for the 48 items was $\alpha=.73$. The correlation between performance in the first artificial grammar learning task and the CFT3 was $r=.16, p=.112$. Taking the reliability estimates of the variables into account, this reveals a correlation corrected for attenuation of $r=.25$.

Prediction of educational success

The participants' final school exams' grade point averages (GPA) ranged between 1.0 and 3.1 with a mean of $M=1.81 (SD=0.66)$. We performed a series of linear regression analyses with GPA as the criterion and subject of study, the performance in the first artificial grammar learning task, and the performance in the CFT3 as predictors. The subject of study was included as a confounder because there is a severe restriction on admission for psychology in Germany and we expected psychology students to have a better GPA than students of other subjects. There were $N=47$ psychology students and $N=59$ students of other subjects.

Table 1 shows the results of four regression analyses. As can be seen, the performance in the first artificial grammar learning task (analysis 2) as well as the performance in the CFT3 (analysis 3) is significantly related with educational success. However, if both

predictors are considered simultaneously, then only general intelligence remains significant (analysis 4).

Please insert Table 1 about here

Discussion

The present study demonstrates that it is possible to measure individual differences in implicit learning with an artificial grammar learning task. In particular, there are some findings that need attention.

The reliability of performance measures is only moderate. The reliability estimates in the present study range between 0.21 and 0.60, which suggests that performance measures of artificial grammar learning are not suitable for individual assessment. This replicates the findings of Reber et al. (1991), Gebauer and Mackintosh (2007), and Danner et al. (submitted). Therefore, the moderate reliability seems to be a general property of artificial grammar learning task measures and not a specific feature of the grammar used or the sample investigated.

Implicit learning performance is divergent from general intelligence. There was a small correlation between the performance in an artificial grammar learning task and the performance in the CFT3. Even if the reliabilities of the variables were taken into account, the correlation corrected for attenuation was only moderate. This result replicates the findings of Reber et al. (1991) and Gebauer and Mackintosh (2007) and points toward the divergent validity of artificial grammar learning measures.

There is a predictive value but not an incremental predictive value of artificial grammar learning measures. The results of the regression analysis demonstrate that artificial grammar learning performance is related with the participants' graduation grade.

However, the regression coefficient becomes non-significant when participants' general intelligence is included as a predictor. This finding suggests that even though both variables overlap only moderately, the relation between artificial grammar learning performance and educational attainment is due to this overlap. Therefore, the present findings speak against Mackintosh's (2006) hypothesis that implicit learning is independent from general intelligence and relevant for success in real life.

A grammar knowledge test decreases the task consistency and increases the correlation between performance and reportable knowledge. As expected, there was a substantial and significant correlation between both artificial grammar learning tasks if the participants did not complete a knowledge test between tasks. If the participants completed a knowledge test between tasks, there was no significant correlation between tasks. This result goes in line with the hypothesis that a knowledge test decreases the task consistency of artificial grammar learning task measures. The correlations between tasks did not differ significantly between the bigram group and the control group, but the sample size of the present study was only sufficient to detect a population difference in the correlation coefficients of $q=0.50$ with a one-tailed type-one-error probability of $\alpha=.05$ and a power of $1-\beta=.80$. To detect a medium effect of $q=0.30$ (Cohen, 1977) a sample size of $N=282$ would have been required and to detect a small effect of $q=0.10$ a sample size of $N=2480$ would have been required. In addition, there was a substantial and significant correlation between the performance in the second artificial grammar learning task and reportable knowledge in the bigram group, but not in the control group. This suggests that the participants changed their strategy after they had completed a bigram knowledge test and this also points towards an effect of a grammar knowledge test on the task consistency. Taken together, this pattern of results suggests that the grammar knowledge test, and not the awareness of a grammar

constituting the letter strings decreases the correlation between subsequent artificial grammar learning tasks.

Conclusion. The present findings demonstrate that artificial grammar learning tasks can be used to measure individual differences in implicit learning, and implicit learning performance is divergent from general intelligence. However, the reliability of performance measures is only moderate and there is no incremental predictive value of implicit learning on educational attainment. Furthermore, a grammar knowledge test decreases the task consistency and increases the correlation between performance and reportable knowledge. Therefore, artificial grammar learning tasks are suitable for investigating individual differences in implicit learning but they are not suitable for individual assessment.

References

- Cattell, R. B., Krug, S. E., & Barton, K. (1973). *Technical Supplement for the Culture Fair Intelligence Tests, Scales 2 and 3*. Champaign, IL: Institute for Personality and Ability Testing.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Danner, D., Hagemann, D., Schankin, A., Bechtold, M., & Funke, J. (submitted). Measuring individual differences in implicit learning with an artificial grammar learning task.
- Dulany, D. E., Carlson, R. A., & Dewey, G. I. (1984). A case of syntactical learning and judgment: How conscious and how abstract? *Journal of Experimental Psychology: General*, *113*, 541-555.
- Frensch, P. A., & Rüniger, D. (2003). Implicit learning. *Current Directions in Psychological Science*, *12*, 13-18.
- Gebauer, G. F., & Mackintosh, N. J. (2007). Psychometric intelligence dissociates implicit and explicit learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 34-54.
- Knowlton, B. J., & Squire, L. R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 169-181.
- Mackintosh, N. J. (2006). *IQ and human intelligence*. Oxford: Oxford University Press.
- Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General*, *119*, 264-275.
- Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning & Verbal Behavior*, *6*, 855-863.

- Reber, A. S. (1992). The cognitive unconscious: An evolutionary perspective. *Consciousness and Cognition: An International Journal*, *1*, 93-133.
- Reber, A. S., & Allen, R. (1978). Analogic and abstraction strategies in synthetic grammar learning: A functionalist interpretation. *Cognition*, *6*, 189-221.
- Reber, A. S., & Allen, R. (2000). Individual differences in implicit learning: Implications for the evolution of consciousness. In R. G. Kunzendorf & B. Wallace (Eds.), *Individual differences in conscious experience*. (pp. 227-247). Amsterdam, Netherlands: John Benjamins Publishing Company.
- Reber, A. S., Walkenfeld, F. F., & Hernstadt, R. (1991). Implicit and explicit learning: Individual differences and IQ. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 888-896.
- Seger, C. A. (1994). Implicit learning. *Psychological Bulletin*, *115*, 163-196.
- Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences*, *17*, 367-447.

Table 1

Regression analyses of GPA

Analysis	Predictor	β	<i>t</i> -value	<i>df</i>	<i>p</i> -value	<i>R</i> ²
1	subject of study	-.56	-6.65	1	<.001	.31
2	subject of study	-.58	-6.95	1	<.001	.34
	AGL	-.17	-2.04	1	.044	
3	subject of study	-.50	-5.99	1	<.001	.36
	CFT3	-.22	-2.67	1	.010	
4	subject of study	-.52	-6.22	1	<.001	.37
	AGL	-.13	-1.62	1	.108	
	CFT3	-.20	-2.35	1	.020	

Note. β = standardized regression coefficient, AGL = performance in the first artificial grammar learning task, CFT3 = performance in the CFT3.

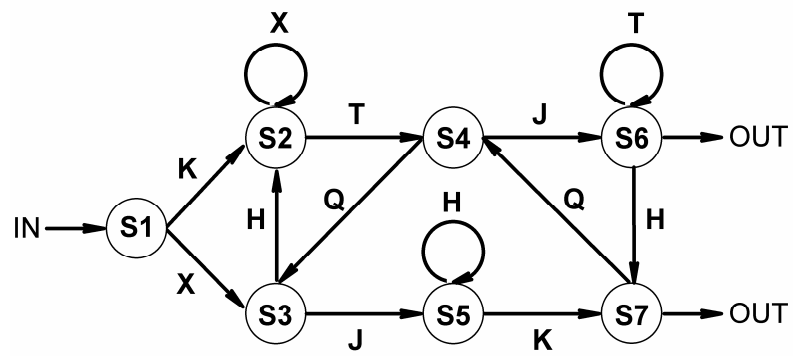


Figure 1: Grammar 1 that was used in the first artificial grammar learning task

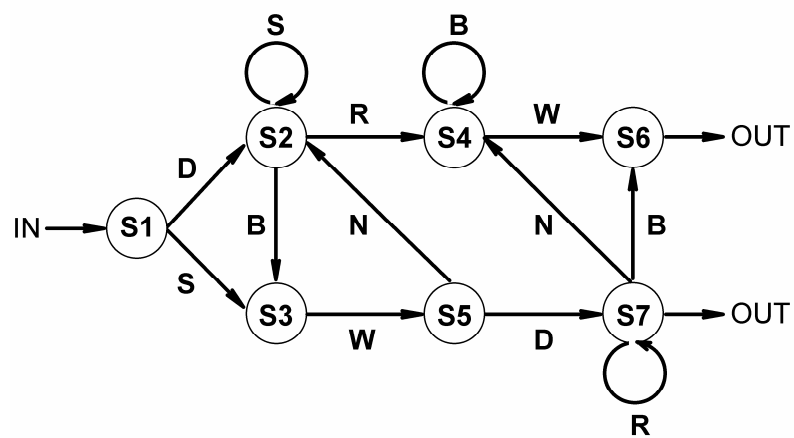


Figure 2: Grammar 2 that was used in the second artificial grammar learning task

Appendix

Table A1

Letter strings for grammar 1 sorted for different parts of the experiment

phase	strings
learning phase	KTJ KTJH KTJHQJ KTJHQJH KTJTH KTJTT KTJTTH KTJTTH KTJTTHH KTQHXTJ KTQJHHK KTQJKQJ KXTJ KXTJHQJ KXTJTTH KXTQHTJ KXXTJ KXXTJTH KXXTQJK KXXXTJ KXXXTJH XHTJ XHTJTT XHTJTTH XHTQHTJ XHTQJK XHXTJT XHXTJTT XHXTQJK XHXXTJ XHXXTJT XJHHHHK XJHHHK XJHHK XJHKQJ XJKQJ XJKQJT XJKQJTT XJKQQJK
testing phase (grammatical strings)	KTJHQJT KTJT KTJTHQJ KTJTTHH KTQHTJ KTQHTJH KTQHTJT KTQJHK KTQJK KXTJH KXTJT KXTJTH KXTJTT KXTJTTH KXTQJHK KXTQJK KXXTJH KXXTJT KXXTJTT KXXXTJT KXXXTJH XHTJH XHTJHQJ XHTJT XHTJTH XHTJTTH XHTJTTH XHXTJ XHXTJH XHXTJTH XHXXTJH XHXXXTJ XJHHKQJ XJHK XJHKQJH XJHKQJT XJK XJKQJH XJKQJTH
testing phase (non- grammatical strings)	KHJT KHQJK KKTJTH KQHTJ KTJQQJT KTJTQT KTQHTH KTQHTTH KTQJJK KXJTHQJ KXJXXTJ KXKTJT KXQJT KXQJTT KXTHH KXTJJK KXTXJK KXXJH KXXJTT KXXTJTT KXXXTKT XHHKQJT XHJTJTH XHTHTTT XHTQJQK XHTTH XHXJJ XHXTXH XHXXQJH XJHHKXJ XJHKQTH XJKK XJKQKTH XJTQJH XKTJT XTK XTTJTH XTXXXTJ XXTJHQJ
knowledge test (bigrams)	HH HJ HK HQ HT HX JH JJ JK JQ JT JX KH KK KQ KT KX QH QJ QK QQ QT TH TJ TK TQ TT TX XH XJ XK XQ XT XX

Table A2

Letter strings for grammar 2 sorted for different parts of the experiment

phase	strings
learning phase	DBWD DBWDNW DBWDNBW DBWDR DBWDRRR DBWNRBW DBWNRW DBWNSRW DRBBBWB DRBBWB DRBW DRBWB DRBWBW DRW DRWB DRWBNBW DRWBRNW DSBWDR DSRBWB DSRWB DSSBWD DSSRBBW DSSRW DSSRWB DSSRWBR DSSSRBW DSSSRW SWDNBW SWDNBWB SWDNW SWDRNBW SWDRNW SWDRRW SWNBWD SWNBWDR SWNRBWB SWNRWB SWNSBWD SWNSRWB
testing phase (grammatical strings)	DBWDNBW DBWDRNW DBWDRR DBWNBWD DBWNRWB DRBBBBW DRBBBB DRBBW DRBBWBR DRBWBR DRWBRR DRWBNWB DRWBR DRWBRRR DSBWD DSRBBW DSRBW DSRW DSRWBR DSSRBW DSSRBWB DSSSBWD DSSSRW DSSSRWB SWD SWDNBBW SWDNWB SWDNWBR SWDR SWDRNBW SWDRR SWDRRR SWDRRRR SWNRBBW SWNRW SWNRWBR SWNSRBW SWNSRW SWNSSRW
testing phase (non- grammatical strings)	DBSDNBW DBSNRRWB DBWDNR DBWDRDW DBWNBSD DRBDBBW DRBNBW DRBWWBR DRSSBWD DRSSRW DRWBNRB DRWDRBR DRWNR DSBRWD DSBW DSBWBR DSBWBRR DSRRBW DSRBWB DSRWSR DSRWW DSSRBBB DSSSNBW DWBBW SBDNWBR SDDR SDNRW SNDRRR SNNSRBW SRD SWBNWB SWDRRSR SWDWNWB SWNNBBW SWNRBWW SWNRR SWNRWRR SWNSRRW SWSSRW
knowledge test (bigrams)	BB BD BN BR BS BW DB DD DN DR DS DW NB ND NN NR NS NW RB RD RN RR RS RW SB SD SN SR SS SW WB WD WN WR WS WW

Appendix A3

Manuscript 3: Measuring performance in dynamic decision making: reliability and validity of the Tailorshop simulation

Measuring performance in dynamic decision making:
reliability and validity of the Tailorshop simulation

Daniel Danner, Dirk Hagemann, Daniel V. Holt, Marieke Hager, Andrea Schankin, Sascha
Wüstenberg, & Joachim Funke

Institute of Psychology, University of Heidelberg, Germany

Author Note

This research was funded by German Research Foundation Grant DFG, Ha 3044/7-1. We gratefully thank Andreas Neubauer, Anna-Lena Schubert, and Katharina Weskamp for conducting the assessment and three anonymous reviewers for helpful comments on an early draft of this manuscript.

Correspondence concerning this paper should be addressed to Daniel Danner, University of Heidelberg, Institute of Psychology, Hauptstrasse 47-51, D-69117 Heidelberg, Germany, Phone: +49 (0) 6221-547354, Fax: +49 (0) 6221-547325, Email: daniel.danner@psychologie.uni-heidelberg.de

Abstract

The Tailorshop simulation is a computer based dynamic decision making task in which participants have to lead a fictional company for twelve simulated months. The present study investigated whether the performance measure in the Tailorshop simulation is reliable and valid. The participants were 158 employees from different companies. Structural equation models were used to test tau-equivalent measurement models. The results indicate that the trends of the company value between the second and the twelfth month are reliable variables. Furthermore, this measure predicted real-life job performance ratings by supervisors and was associated with the performance in another dynamic decision making task. Thus, the trend of the company value provides a reliable and valid performance indicator for the Tailorshop simulation.

Keywords: dynamic decision making, complex problem solving, Tailorshop, reliability, validity

Measuring performance in dynamic decision making:
reliability and validity of the Tailorshop simulation

Real life decisions are complex and sometimes there are no well-defined solutions for problems. A manager has to make decisions even if he or she does not have all relevant information, or an employer has to pursue the interests of his staff as well as the goals of his company, even if both views may be conflicting. Gonzalez, Yanyukov, and Martin (2005) call such decisions *dynamic decisions*. They are characterized by dynamics, complexity, opaqueness, and dynamic complexity. In a similar vein, Dörner (1980) characterizes such problems as *complex problems*, which means that their structure is complex, connected, dynamic, and non-transparent. Recently, dynamic decision making tasks have also been included in the Programme for International Student Assessment (PISA; Wirth & Klieme, 2003). Since the ability to deal with such problems may have impact on important decisions in real life, it is an interesting question whether there are individual differences in dynamic decision making and whether these differences can be measured reliably and validly (e.g., Baker & O'Neil, 2002; Rigas, Carling, & Brehmer, 2002; Süß, 1996, 1999; Strohschneider, 1986; Zaccaro, Mumford, Connelly, Marks, & Gilbert, 2000). Investigating these issues was the aim of the present study.

To investigate dynamic decision making, several authors suggested to study persons' behavior in computer simulations. The *Tailorshop* is such a dynamic decision making task, which has been used for several decades (e.g., Barth & Funke, 2010; Putz-Osterloh, Bott, & Köster, 1990; Süß, Kersting, & Oberauer, 1993; Wittmann & Hatstrup, 2004). The scenario simulates a small business that produces and sells shirts. The participants have to lead this business for twelve simulated months by manipulating several variables like the number of workers, the expenses for advertising, etc. (see Figure 1).

Please insert Figure 1 about here

In total, the Tailorshop consists of 24 variables. Twenty-one variables are visible to the participants and three variables are invisible to the participant. Twelve variables can be manipulated directly (e.g., the costs for advertising) whereas other variables can only be manipulated indirectly (e.g., the demand). The state of a variable in a given month influences the state of the same and other variables in a following month. Figure 2 shows schematically how the variables are connected (see Funke, 1983, for an algebraic definition of all system variables).

Please insert Figure 2 about here

In order to use the performance in the Tailorshop for the investigation of individual differences or for individual assessment, the performance variable should be reliable and valid. The reliability of a performance variable is important in two ways.

In a research context, reliability considerations are important for an understanding of the *validity* of dynamic decision making measures because the reliability of a variable affects its correlation with criterion variables. In an applied context, the Tailorshop may be used to measure a single person's ability to solve complex problems, e.g., as part of an assessment center. This measurement is only useful if it is reliable because otherwise it will yield incorrect decisions.

Reliability estimation.

In classical test theory, the reliability of a variable is defined as the proportion of the true score variance relative to the total variance of a variable (Lord & Novick, 1968). In the Tailorshop scenario, the reliability is defined as the proportion of true individual performance differences relative to the total individual performance differences. The true score τ of a measurement i of a variable Y is defined as the expected value given a particular person P (Lord & Novick, 1968). In the Tailorshop scenario, the true score of a performance variable is defined as the expected performance given a particular person, $\tau_i := E(Y_i|P)$. In addition, the measurement error ε is defined as the deviation of the measured variable from the true score variable, $\varepsilon_i := Y_i - \tau_i$ (Lord & Novick, 1968). To estimate the reliability, multiple, experimentally independent measurements of a variable are necessary.

In addition, two assumptions have to be made which define the τ -equivalent measurement model. The first assumption is that the true score of a measurement i of a particular person is identical with the true score of another measurement j of this person, $\tau_i = \tau_j =: \tau$. The second assumption is that the errors of the measurements are uncorrelated, $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$, for all $i \neq j$. These assumptions may be tested with a structural equation model (Steyer, 1989) as shown in Figure 3. If the assumptions hold, then the variance of the true score may be estimated and the reliability may be computed by reliability $Y_i = \frac{\text{var}(\tau)}{\text{var}(Y_i)}$.

Please insert Figure 3 about here

Validity assessment

According to Dörner (1980) and Gonzalez et al. (2005) dynamic decisions are characterized by complexity, connectivity, non-transparency, and dynamics. Hence, the *content validity* of a performance variable may be evaluated regarding these four criteria. The

convergent validity may be evaluated by the correlation with another dynamic decision making task. Therefore, we expected a substantial correlation with the dynamic decision making task *Heidelberg Finite State Automaton* (Wirth & Funke, 2005), which has also been used in the German PISA assessment in 2000 (Wirth & Klieme, 2003). The *predictive validity* may be evaluated by the correlation with real life performance. Therefore, we expected that the performance in the Tailorshop can predict professional success. Finally, the *divergent validity* may be attested by a low correlation with another ability construct. Hence, we hypothesized that there is a low correlation between the performance in the Tailorshop and the performance in a standard intelligence test.

Performance measurement

At the beginning of the simulation, the participants were instructed to maximize the company value. Thus, the success of dynamic decision making may be measured by the achieved company value. The simplest approach may be to measure the company value after every month. However, the company value of a particular month depends on the company value of the previous month, $\text{company value}_i = \text{company value}_{i-1} + \text{change}_i$. Therefore, the company values are not experimentally independent and the assumption of uncorrelated errors will be violated. On the other hand, there is no such relationship between the *changes of the company values*. Furthermore, the sum of the changes of the company values corresponds to the company value after twelve months because the company value at the beginning of the simulation is identical for all participants, $\text{company value}_{12} = \text{company value}_0 + \sum_{i=1}^{12} \text{change}_i$.

Therefore, the *changes of the company values* after each simulated month may be taken as performance indicators for the Tailorshop simulation.

As an alternative, Funke (1983) suggested to use the *trends of the company value* as performance indicators. The trends of the company value are binary variables. If the company value between two successive months increases, the trend is positive. If the company value

decreases, the trend is zero.¹ This scoring may have several advantages. First, the trend measure is simple to interpret because each point corresponds to a month where the given aim (“maximize company value”) was achieved. Second, the trend measure is robust against outliers, whereas the change value may rise to extreme values (due to the non-linear relationships between the variables). And finally, the measurement model for the trend measure makes fewer assumptions than the measurement model for the change measures on how the company value develops over the months. In particular, the τ -equivalent measurement model for the change measures states that the (true) change of a person is constant over the months, $\tau_i = \tau_j$. On the other hand, the measurement model for the trend measures only states that a person who has a greater probability to make gain in a particular month, also has a greater probability to make gain in another month.

Aim of the present study

The aim of the present study was to investigate the reliability and the validity of (1) the change of the company value and (2) the trend of the company value. The reliabilities of these variables were investigated with τ -equivalent measurement models. Furthermore, the content, convergent, predictive, and divergent validities of these variables were evaluated.

Method

Participants

The participants were $N=158$ employees (111 female, 47 male), who were recruited via newspaper announcement from different branches and different companies around Heidelberg. The participants rated their jobs according to the International Standard Classification of Occupations (ISCO-88 COM). 6% rated themselves as legislators, senior officials, and managers, 25% as professionals, 11% as technical and associate professionals, 14% as clerks, 40% as service workers and shop and market sales workers, 1% as craft and

related trade workers, 1% as plant and machine operators and assemblers, and 1% as elementary occupations. The participants' mean age was $M=43.34$ years ($SD=11.22$).

Measures

Advanced Progressive Matrices. General intelligence was measured using the Advanced Progressive Matrices (Raven, Court, & Raven, 1994). The number of solved items in the second set was taken as a performance indicator. Cronbach's alpha for the 36 items was $\alpha=.85$.

Heidelberg Finite State Automaton. The Heidelberg Finite State Automaton (Wirth & Funke, 2005) was used as a second indicator for dynamic decision making. The scenario is computer based and simulates a space flight where the participants control a space ship and a ground vehicle with a graphical user interface (see Figure 4). The system variables are connected and dynamic. For example, the ability to fly with the space ship depends on the state of the propulsion, the heat shield, the landing gear, and the state of the ground vehicle. The performance was measured with 22 items where the participants have to reach a specified target (e.g., land the space ship on a particular planet). The number of solved items was taken as the performance variable. Cronbach's alpha for the 22 items was $\alpha=.93$.

Please insert Figure 4 about here

Tailorshop. The participants were given information about the meaning of the variables in the Tailorshop (e.g., "The account status is the amount of money in your account that is available anytime. A negative value signifies that you took a loan."). Further, the participants were instructed to maximize the company value within twelve simulated months. For the purpose of the present study we measured (1) the changes of the company value and (2) the trends of the company value after every simulated month (English and German

versions of the Tailorshop simulation software are available from the website <http://www.atp.uni-hd.de/tools/tailorshop>).

Professional success. The participants' professional success was measured by supervisor ratings (Higgins, Peterson, Pihl, & Lee, 2007) with five items on a six point scale ("The employee achieves arranged and set objectives", "The employee demonstrates competence in all job-related tasks", "The employee meets all my expectations in his roles and responsibilities", "How do you rate the quality of his work?", "How do you rate the overall level of performance that you observe for this employee?"). Cronbach's alpha for these five item was $\alpha=.91$. In addition, the participants' yearly income was measured with thirteen categories (1 = "under €2,500", 2 = "€2,50 to €5,000", 3 = "€5,000 to €7,500", 4 = "€7,500 to €10,000 €", 5 = "€10,000 to €12,500", 6 = "€12,500 to €15,000", 7 = "€15,000 to €20,000", 8 = "€20,000 to €25,000", 9 = "€25,000 to €30,000", 10 = "€30,000 to €37,500", 11 = "€37,500 to €50,000", 12 = "€50,000 to €125,000", 13 = "over €125,000").

Results

Measurement models

The τ -equivalent measurement model was specified according to Figure 2. The measurement model for the *change variables* was estimated using the maximum likelihood procedure implemented in Mplus 5. The measurement model for the *trend variables* was estimated using the means and variance adjusted weighted least square estimator (WLSMV) implemented in Mplus 5 (Muthén & Muthén, 2007). In a first step, we estimated the measurement models for the performance indicators of all twelve months. However, the first assessment in a study may be unreliable and sometimes may not measure what is intended. Therefore, we also estimated the measurement models for the last eleven months, then for the last ten months and so on.

Neither measurement model for the *change variables* fitted with the data, all $\chi^2 > 714.41$, all RMSEA > 0.71 , all CFI < 0.45 . However, the measurement models for the *trend variables* fitted better with the data. The results are reported in Table 1. As can be seen, the measurement model for the last eleven trend variables revealed an acceptable model fit and the measurement models for the last nine or fewer trend variables fitted even better. However, the fewer months were included, the smaller the covariance matrix was and the fewer covariances had to be fitted with the parameters of the model. Therefore, the better model fit might also be a consequence of the smaller covariance matrix. Furthermore, the dynamics during twelve months is greater than the dynamics in only the last few months. Therefore, the more months are captured by a performance measure, the greater the content validity of the measure will be. Therefore, we decided to accept the measurement model for the last eleven trend variables and use it for reliability estimation.

Please insert Table 1 about here

The estimated variance of the latent τ -variable was 0.70, $p < .001$. Therefore, the reliability of each trend variable may be estimated by

$$\text{reliability trend}_i = \frac{\text{var}(\tau)}{\text{var}(\text{trend}_i)} = \frac{0.70}{1.00} = 0.70. \text{ Applying the Spearman-Brown formula to}$$

estimate the reliability of the sum score of these eleven items reveals a reliability estimate of 0.96.

Correlation between performance in the Tailorshop and other variables

To evaluate the convergent, predictive, and divergent validity of (1) the change and (2) the trend of the company value, we computed the correlations between these performance variables and the performance in the Heidelberg Finite State Automaton, the participants' income, the participants' supervisor ratings, and the performance in the APM. The sum of the

change variables was used as the performance indicator *change of the company value* and the sum of the trend variables (between the second and twelfth month) was used as the performance indicator *trend of the company value*.

The correlations between these variables are reported in Table 2. As can be seen, the correlation between the change variable and the trend variable was neither substantial nor significant, which suggests that both performance variables measure different performance aspects. The *change of the company value* only correlated significantly with the APM, which suggests a low overall validity of this performance variable.

On the other hand, there was a significant and substantial correlation between the *trend of the company value* and the Heidelberg Finite State Automaton, which points towards the convergent validity of the trend variable. Furthermore, there was a significant correlation between the trend variable and the supervisor ratings, which points towards the predictive validity of this measure.

Please insert Table 2 about here

There was also a substantial correlation between the trend of the company value and the APM. Therefore, we additionally computed partial correlations that were adjusted for the performance in the APM. The partial correlation between the trend variable and the Heidelberg Finite State Automaton was $r=.20$, $p=.023$, the partial correlation between the trend variable and the participants' income was $r=.05$, $p=.525$, and the partial correlation between the trend variable and the supervisor ratings was $r=.22$, $p=.010$.

Outlier analysis

The measurement models for the trend values fitted better with the data than the measurement models for the change variables. One reason for this may be that the trend

variables are less sensitive to outliers. To investigate the role of outliers in greater detail, we z -transformed the change variables for each month. There were $N=7$ participants with $|z|>3$ in at least one month. These z -values were trimmed to a maximum of $z=3$ and a minimum of $z=-3$ and the measurement models were estimated again. However, the measurement model for the trimmed change values also did not fit with the data, $\chi^2(65)=1963.52$, $p<.001$, RMSEA=0.43, CFI=0.30.

In addition, we computed the correlations between the (sum of the) trimmed change values and the participants' scores of the Heidelberg Finite State Automaton, income, supervisor ratings, and APM. The correlation with the Heidelberg Finite State Automaton was $r=.24$, $p=.003$, the correlation with the participants' income was $r=.02$, $p=.807$, the correlation with the supervisor ratings was $r=.14$, $p=.102$, and the correlation with the APM was $r=.38$, $p<.001$. Meng, Rosenthal, and Rubin's (1992) method for comparing correlated correlations revealed that none of these correlations was significantly greater than the correlation with the trend variable.

Discussion

The aim of the present study was to evaluate the reliability and the validity of performance variables in the Tailorshop simulation. Therefore, we investigated (1) the change of the company value and (2) the trend of the company value.

Reliability and measurement models

The measurement models for the changes of the company value did not fit with the data. This suggests that the single change values are not suitable for the reliability estimation. One reason for this may be that the τ -equivalent measurement model makes rather strong assumptions about how the company value develops over the months. In particular, the model states that the "true" change of the company value in the month i is the same than the "true" change in the month j , $\tau_i = \tau_j$.² However, this assumption may be violated because different

persons may use different strategies to maximize their company value. For example, one participant may make great investments in the first month and therefore has little gain first and great gain later. Another participant may make constant investments and therefore have a constant gain across the months. Hence, investigating individual differences in dynamic decision making *processes* may be a worthwhile issue for future research. Nonetheless, the structural equation model analysis of the present study revealed that the sum of the trends between the second and twelfth month is a reliable performance variable.

Content validity

The Tailorshop was developed according to Dörner's (1980) definition of dynamic decision making. In particular, the simulation may be seen as *complex* and *connected* because it consists of many variables that are connected. The tasks may also be seen as *non-transparent* because the participants do not know how the variables in the simulation are connected and the tasks may be seen as *dynamic* because each intervention in the simulation influences the following state of the simulation. Therefore, the structure of the present dynamic decision making task can be seen as a valid representation of general dynamic decision making demands. Furthermore, the participants were instructed to maximize their company value and therefore, the changes in the company value as well as the trends of the company can be seen as content valid performance measures.

Convergent validity

The correlation between the trend of the company value and the performance in the Heidelberg Finite State Automaton was substantial and significant, which indicates the convergent validity of this variable. Furthermore, this correlation remained significant when adjusted for general intelligence, which indicates that the relation between both dynamic decision making tasks is incremental to the overlap with general intelligence.

On the other hand, the correlation between the change of the company value and the performance in the Heidelberg Finite State Automaton was close to zero and not significant. After controlling for outliers this correlation increased. However, controlling for outliers may be difficult, especially in small samples or in individual assessments. Furthermore, none of the correlations with the trimmed change variable was significantly greater than the correlation with the trend variable.

Predictive validity

The correlation between the change of the company value and the participants' supervisor ratings was not significant. However, there was a significant correlation between the trend of the company value and the supervisor ratings, which remained significant after controlling for individual differences in general intelligence. This indicates the incremental predictive validity of the trend measure. This replicates the findings of Kersting (2001), who also reported an incremental predictive value of a dynamic decision making measures on participants' superior ratings. Furthermore, this result points towards the practical value of dynamic decision making measures and suggests that they may provide insights into aspects of professional success, which cannot be predicted by general intelligence.

There was no relationship with participants' income.³ This may be due to two reasons. First, income may measure a different aspect of professional success than supervisor ratings. This is supported by the low and non-significant correlation between income and supervisor rating. Second, income may just be a valid indicator for professional success within an occupational category and not between. For example, a priest may earn less than a broker, even if the priest does his job better than the broker.

Divergent validity and the relationship between dynamic decision making and general intelligence

Dörner and colleagues (e.g., Dörner, 1980; Dörner & Kreuzig, 1983), who introduced the construct of dynamic decision making (or complex problem solving respectively), proposed that general intelligence and dynamic decision making are independent abilities. They reported several studies where low relations between measures of general intelligence and dynamic decision making were observed (Dörner, Kreuzig, Reither, & Staudel, 1983; Putz-Osterloh, 1981; Putz-Osterloh & Lürer, 1981). However, following studies revealed rather heterogeneous findings. Kluwe, Misiak, and Haider (1991) presented an overview of early studies and reported a broad range of correlation (between $r=-.52$ and $r=.46$), whereas subsequent studies found stronger associations (Kröner, Plass, & Leutner, 2005; Wittmann & Hatrup, 2004). One study even found a correlation between a latent intelligence and a latent dynamic decision making variable of $r=.84$ (Wirth & Klieme, 2003).

In the present study, there was a significant correlation of $r=.31$ between the performance in the APM and the performance in the Tailorshop. In addition, there was a significant correlation of $r=.57$ between the performance in the APM and the performance in the Heidelberg Finite State Automaton. Thus, general intelligence could explain 10% (or 32% respectively) of the variance in dynamic decision making performance which suggests that there is a partial but not a complete overlap between the constructs.

However, our results do not allow to draw final conclusions about the relation between general intelligence and dynamic decision making. In particular, Wittmann (1988; Wittmann & Süß, 1999) suggested that the relation between two indicators only allows conclusions about the relation between underlying constructs if the indicators are symmetric. For example, the APM may be seen as an intelligence test that particularly captures individual differences in figural reasoning. In a similar vein, the Tailorshop may particularly capture

individual differences in economy related dynamic decision making. Therefore, both measures may contain not only systematic construct variance (e.g., general intelligence variance) but also “unwanted” but reliable and specific variance (e.g., specific figural reasoning variance in the APM). However, investigating the symmetry of the variables would require to measure each construct with several indicators and at several measurement occasions. Following this reasoning, the present findings can not provide a final answer to the question on how general intelligence and dynamic decision making are related.

Performance differences between men and women

Wittmann and Hatrupp (2004) reported that men showed a better performance in the Tailorshop than women ($d=0.70$). This finding was replicated in the present study. The number of months with a positive trend in the company value (between the second and the twelfth month) was greater for men ($M=3.60$) than for women ($M=2.25$), $t(156)=2.49$, $p=.014$, $d=0.46$. Wittmann and Hatrupp (2004) suggested that women may behave more risk-averse than men and therefore construct themselves a less favorable learning environment in the Tailorshop and accordingly show a lower performance. Furthermore, there were no significant performance differences between women and men in the Heidelberg Finite State Automaton or the APM, which suggests that these differences are task specific for the Tailorshop.

Conclusion

The sum of the trends between the second and the twelfth month is a reliable and valid performance indicator in the Tailorshop simulation. Hence, this score may be used for the study of individual differences as well as for individual assessments. For example, dynamic decision making tasks may be a useful complement for the selection of job applicants as suggested by Kersting (2001).

References

- Baker, E. L., & O'Neil, H. F. (2002). Measuring problem solving in computer environments: Current and future states. *Computers in Human Behavior, 18*, 609-622. doi: 10.1016/S0747-5632(02)00019-5
- Barth, C. M., & Funke, J. (2010). Negative affective environments improve complex solving performance. *Cognition and Emotion, 24*, 1259-1268. doi: 10.1080/02699930903223766
- Dörner, D. (1980). On the difficulty people have in dealing with complexity. *Simulation & Gaming, 11*, 87-106.
- Dörner, D., & Kreuzig, H. W. (1983). Problemlösefähigkeit und Intelligenz. *Psychologische Rundschau, 34*, 185-192.
- Dörner, D., Kreuzig, H. W., Reither, F. Stäudel, T. (1983). *Lohausen. Vom Umgang mit Unbestimmtheit und Komplexität*. Bern: Hans Huber.
- Funke, J. (1983). Einige Bemerkungen zu Problemen der Problemlöseforschung oder: Ist Testintelligenz doch ein Prädiktor? *Diagnostica, 29*, 283-302.
- Gonzalez, C., Vanyukov, P., & Martin, M. K. (2005). The use of microworlds to study dynamic decision making. *Computers in Human Behavior, 21*, 273-286. doi: 10.1016/j.chb.2004.02.014
- Higgins, D. M., Peterson, J. B., Pihl, R. O., & Lee, A. G. M. (2007). Prefrontal cognitive ability, intelligence, Big Five personality, and the prediction of advanced academic and workplace performance. *Journal of Personality and Social Psychology, 93*, 298-319. doi: 10.1037/0022-3514.93.2.298
- Kersting, M. (2001). Zur Konstrukt- und Kriteriumsvalidität von Problemlöseszenarien anhand der Vorhersage von Vorgesetztenurteilen über die berufliche Bewährung. *Diagnostica, 47*, 67-76. doi: 10.1026//0012-1924.47.2.67

- Kluwe, R. H., Misiak, C., & Haider, H. (1991). The control of complex systems and performance in intelligence tests. In H. A. H. Rowe (Ed.), *Intelligence: Reconceptualization and measurement*. (pp. 227-244). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Kröner, S., Plass, J. L., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence*, *33*, 347-368. doi: 10.1016/j.intell.2005.03.002
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Oxford: Addison-Wesley.
- Meng, X., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*, 172-175. doi: 10.1037/0033-2909.111.1.172
- Muthén, L. K., & Muthén, B. O. (2007). *Mplus User's Guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.
- Putz-Osterloh, W. (1981). Über die Beziehung zwischen Testintelligenz und Problemlöseerfolg. *Zeitschrift für Psychologie mit Zeitschrift für Angewandte Psychologie*, *189*, 79-100.
- Putz-Osterloh, W., Bott, B., & Köster, K. (1990). Modes of learning in problem solving: Are they transferable to tutorial systems? *Computers in Human Behavior*, *6*, 83-96. doi: 10.1016/0747-5632(90)90032-c
- Putz-Osterloh, W., & Lüer, G. (1981). The predictability of complex problem solving by performance on an intelligence test. *Zeitschrift für Experimentelle und Angewandte Psychologie*, *28*, 309-334.
- Raven, J. C., Court, J. H., & Raven, J. (1994). *Manual for Raven's Progressive Matrices and Mill Hill Vocabulary Scales. Advanced Progressive Matrices*. Oxford: Oxford Psychologists Press.

- Rigas, G., Carling, E., & Brehmer, B. (2002). Reliability and validity of performance measures in microworlds. *Intelligence, 30*, 463-480. doi: 10.1016/s0160-2896(02)00121-6
- Roszkowski, M. J., & Grable, J. E. (2010). Gender differences in personal income and financial risk tolerance: How much of a connection? *The Career Development Quarterly, 58*, 270-275.
- Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability, and testability. *Methodika, 3*, 25-60.
- Strohschneider, S. (1986). Zur Stabilität und Validität von Handeln in komplexen Realitätsbereichen. *Sprache & Kognition, 5*, 42-48.
- Süß, H.-M. (1996). *Intelligenz, Wissen und Problemlösen*. Göttingen: Hogrefe.
- Süß, H.-M. (1999). Intelligenz und komplexes Problemlösen: Perspektiven für eine Kooperation zwischen differentiell-psychometrischer und kognitionspsychologischer Forschung. *Psychologische Rundschau, 50*, 220-228. doi: 10.1026//0033-3042.50.4.220
- Süß, H.-M., Kersting, M., & Oberauer, K. (1993). Zur Vorhersage von Steuerungsleistungen an computersimulierten Systemen durch Wissen und Intelligenz. *Zeitschrift für Differentielle und Diagnostische Psychologie, 14*, 189-203.
- Wirth, J., & Funke, J. (2005). Dynamisches Problemlösen: Entwicklung und Evaluation eines neuen Messverfahrens zum Steuern komplexer Systeme. In E. Klieme, D. Leutner & J. Wirth (Eds.), *Problemlösekompetenz von Schülerinnen und Schülern* (pp. 55-72). Wiesbaden: Verlag für Sozialwissenschaften.

- Wirth, J., & Klieme, E. (2003). Computer-based Assessment of Problem Solving Competence. *Assessment in Education: Principles, Policy & Practice, 10*, 329-345. doi: 10.1080/0969594032000148172
- Wittmann, W. (1988). Multivariate reliability theory: Principles of symmetry and successful validation strategies. In J. R. Cattell (Ed.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 505-560). New York: Plenum Press.
- Wittmann, W., & Hattrup, K. (2004). The relationship between performance in dynamic systems and intelligence. *Systems Research and Behavioral Science, 21*, 393-409. doi: 10.1002/sres.653
- Wittmann, W. W., & Süß, H.-M. (1999). Investigating the paths between working memory, intelligence, knowledge, and complex problem-solving performances via Brunswik symmetry. In P. L. Ackerman, P. C. Kyllonen & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants*. (pp. 77-108). Washington, DC: American Psychological Association.
- Zaccaro, S. J., Mumford, M. D., Connelly, M. S., Marks, M. A., & Gilbert, J. A. (2000). Assessment of leader problem-solving capabilities. *The Leadership Quarterly, 11*, 37-64. doi: 10.1016/S1048-9843(99)00042-9

Footnotes

¹ Due to the complex relations between the variables it is very unlikely to obtain a change in the company value of exactly zero. In the present study, there was always either a positive or a negative change in the company value.

² We additionally investigated the change variables with a τ -congeneric measurement model, which makes weaker assumptions than the τ -equivalent measurement model. In particular, the model states that the “true” change of the company in a month i can be linearly transformed into the true score of another month j , $\tau_j = \gamma^* \tau_i$. (Lord & Novick, 1968; Steyer, 1989). However, the τ -congeneric measurement model fitted neither with the non-trimmed change variables ($\chi^2(54)=4582.79$, $p<.001$, RMSEA=0.73, CFI=0.16) nor with the trimmed change variables ($\chi^2(54)=1605.81$, $p<.001$, RMSEA=0.43, CFI=0.42).

³ Some studies (e.g. Roszkowski & Grable, 2010) report, that women earn less than men. Therefore, we additionally calculated this correlation separately for women and men. There were no significant differences.

Table 1

Model fit indices for the measurement models for the trend of the company value

Trend	χ^2	<i>df</i>	<i>p</i>	RMSEA	CFI
1 to 12	79.85	22	<.001	0.13	0.94
2 to 12	40.10	23	.015	0.07	0.98
3 to 12	38.08	21	.013	0.07	0.98
4 to 12	25.35	18	.116	0.05	0.99
5 to 12	17.77	16	.337	0.03	1.00
6 to 12	11.93	14	.612	0.00	1.00
7 to 12	8.51	11	.667	0.00	1.00
8 to 12	6.19	8	.626	0.00	1.00
9 to 12	2.29	5	.808	0.00	1.00
10 to 12	1.24	2	.538	0.00	1.00












Table 2

Correlations between performance variables (p-values in brackets)

	Change	Trend	HFA	Income	Supervisor rating
Trend	.13 (.098)				
HFA	.03 (.255)	.31 (<.001)			
Income	.01 (.923)	.08 (.323)	.05 (.561)		
Supervisor rating	.15 (.085)	.19 (.025)	.09 (.292)	-.02 (.801)	
APM	.19 (.020)	.31 (.001)	.55 (<.001)	.16 (.054)	-.03 (.706)

Note. change = sum of changes of the company value, trend = sum of trends of the company value (between second and twelfth month), HFA = Heidelberg Finite State Automaton, income = participants' yearly income, APM = Advanced Progressive Matrices.

Round 1 of 12

Variable	Value	Planning	
Account status	165775		
Number of shirts sold	407		
Price of raw material	3.99		
Shirts in stock	81		
Workers 50	8	<input type="text"/>	
Workers 100	0	<input type="text"/>	
Salary	1080	<input type="text"/>	
Price of shirts	52	<input type="text"/>	
Shops	1	<input type="text"/>	
Worker satisfaction %	57.7		
Loss of production %	0.0		











Variable	Value	Planning	
Company value	250685		
Demand	767		
Raw material in stock	16	<input type="text"/>	
Machines 50	10	<input type="text"/>	
Machines 100	0	<input type="text"/>	
Repair & service costs	1200	<input type="text"/>	
Social costs per worker	50	<input type="text"/>	
Advertising costs	2800	<input type="text"/>	
Business location	suburb	<input type="text" value="suburb"/>	
Machine damage %	5.9		

Figure 1. Screenshot of the graphical user interface of the Tailorshop (labels translated).

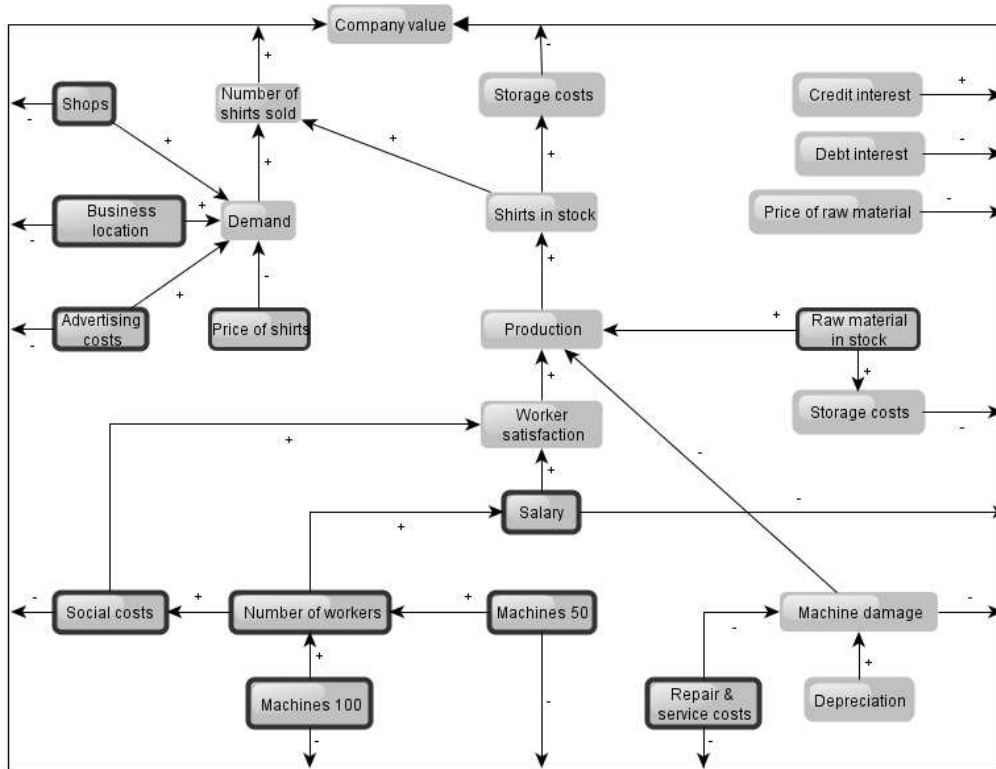


Figure 2. Schematic relation between the variables in the Tailorshop. The marked variables can be manipulated directly.

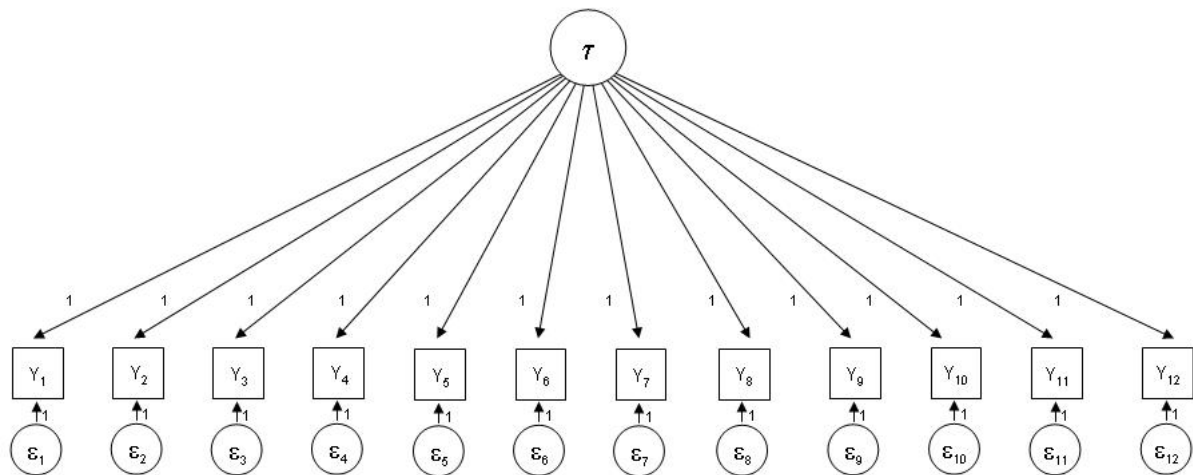


Figure 3. τ -equivalent measurement model. τ = true score variable, ε = measurement error variable.

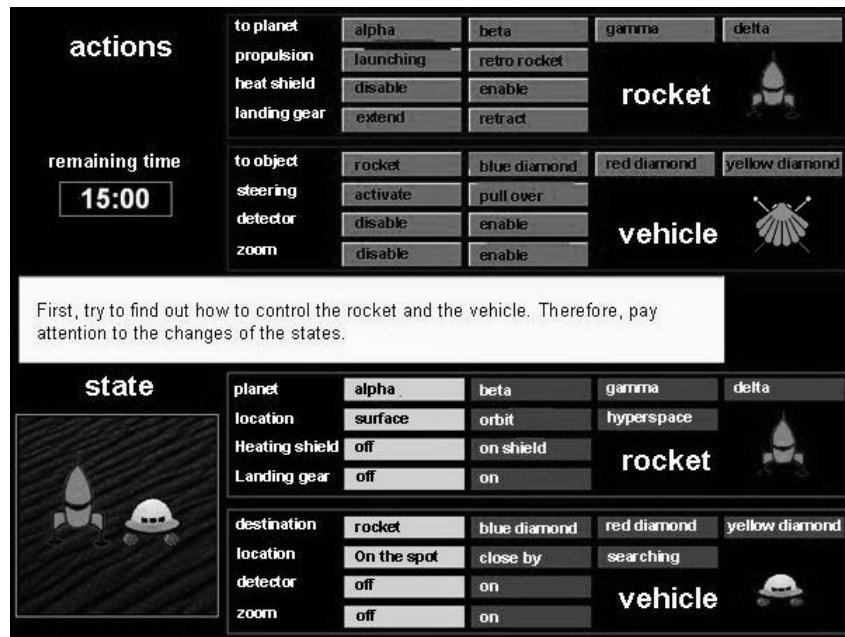


Figure 4. Screenshot of the graphical user interface of the Heidelberg Finite State Automaton (labels translated).

Appendix A4

Manuscript 4: Beyond IQ. A latent state trait analysis of general intelligence, dynamic decision making, and implicit learning

Contents lists available at [ScienceDirect](#)

Intelligence

journal homepage:



Beyond IQ: A latent state-trait analysis of general intelligence, dynamic decision making, and implicit learning[☆]

Daniel Danner^{*}, Dirk Hagemann, Andrea Schankin, Marieke Hager, Joachim Funke

Institute of Psychology, University of Heidelberg, Germany

ARTICLE INFO

Article history:

Received 31 January 2011
 Received in revised form 6 May 2011
 Accepted 3 June 2011
 Available online xxx

Keywords:

Dynamic decision making
 Complex problem solving
 Implicit learning
 Latent state-trait theory
 Professional success

ABSTRACT

The present study investigated cognitive performance measures beyond IQ. In particular, we investigated the psychometric properties of dynamic decision making variables and implicit learning variables and their relation with general intelligence and professional success. $N = 173$ employees from different companies and occupational groups completed two standard intelligence tests, two dynamic decision making tasks, and two implicit learning tasks at two measurement occasions each. We used structural equation models to test latent state-trait measurement models and the relation between constructs. The results suggest that dynamic decision making and implicit learning are substantially related with general intelligence. Furthermore, general intelligence is the best predictor for income, social status, and educational attainment. Dynamic decision making can predict supervisor ratings even beyond general intelligence.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

General intelligence is one of the most successful psychological constructs. Since Spearman's (1904) early investigations, there is a wealth of evidence for the reliability, stability, and validity of intelligence measures (Carroll, 1993). Furthermore, general intelligence is a powerful predictor of success in many domains of real life (Ng, Eby, Sorensen, & Feldman, 2005; Salgado et al., 2003; Schmidt & Hunter, 2004). Beside its undisputed usefulness, some researchers have suggested to use additional constructs for characterizing individuals' cognitive ability such as dynamic decision making and implicit learning (Dörner, 1980; Mackintosh, 1998).

The concept of dynamic decision making was developed by Dörner (1980, 1986) who proposed that situations in real life are complex and solving problems in real life requires managing complex information. He criticized that standard measures of general intelligence only assess whether individuals perform accurately and quickly in rather simple tasks but not whether they show intelligent behavior in complex tasks. Therefore, he suggested to measure performance in computer based scenarios that simulate complex, connected, dynamic, and non-transparent environments. Further on, he hypothesized that individual differences in dynamic decision making are unrelated to general intelligence but are substantially related to professional success.

Mackintosh (1998) suggested to consider another construct. He proposed that there are two independent mental systems: an explicit, hypothesis generating and testing system and an implicit, associative learning system. In particular, the explicit learning system is necessary for discovering regularities with intention and awareness (like in a numerical series task). The implicit learning system, on the other hand, detects contingencies without awareness or intention (like judging whether a sentence is grammatically

[☆] This research was funded by German Research Foundation Grant DFG, Ha3044/7-1. We gratefully thank Andreas Neubauer, Anna-Lena Schubert, and Katharina Weskamp for conducting the assessment and Neil Patrick Harris and two anonymous reviewers for their helpful suggestions.

^{*} Corresponding author at: University of Heidelberg, Institute of Psychology, Hauptstrasse 47-51, D-69117 Heidelberg, Germany. Tel.: +49 6221 547354; fax: +49 6221 547325.

E-mail address: daniel.danner@psychologie.uni-heidelberg.de (D. Danner).

right or wrong without being able to report the respective grammatical rule). Mackintosh suggested that standard intelligence tests capture individual differences in the explicit system but not individual differences in the implicit learning system. Therefore, he suggested to take individual differences in implicit learning into account. He hypothesized that these differences are independent from general intelligence measures but are nevertheless important predictors of educational and professional success.

Dörner and Mackintosh's proposals raise two interesting questions. Are there reliable individual differences in dynamic decision making and implicit learning which are independent from general intelligence? Can these differences predict real life performance beyond IQ? Investigating these issues will be the aim of the present study.

1.1. Previous findings

1.1.1. Dynamic decision making

Dörner's (1980, 1986) critique of standard intelligence tests laid the foundation for a field of research, which has been called *dynamic decision making* (Gonzalez, Vanyukov, & Martin, 2005) or *complex problem solving* (Funke, 2010). Over the years, several dynamic decision making tasks have been developed. For example, the Tailorshop scenario (Funke, 1983) simulates a fictional company where the participants have to control many variables like the number of workers or the costs for advertising to maximize their company value. Other tasks simulate a forestry (Wagener, 2001), a power plant (Wallach, 1998), or a space flight (Wirth & Funke, 2005) where the participants have to control several variables to reach a given goal state. Recently, dynamic decision making tasks have also been included in the Programme for International Student Assessment (PISA; Wirth & Klieme, 2003).

Over the years, there have been many studies investigating the relation between dynamic decision making and general intelligence. Whereas several studies found non-significant or only small correlations (for an overview see Kluwe, Misiak, & Haider, 1991), other studies reported significant standardized path coefficients between $\beta = 0.38$ and $\beta = 0.54$ from latent intelligence to latent dynamic decision making variables (Kröner, Plass, & Leutner, 2005; Rigas, Carling, & Brehmer, 2002; Wittmann & Hatstrup, 2004). One study even found a correlation between a latent intelligence and a latent dynamic decision making variable of $r = 0.84$ (Wirth & Klieme, 2003).

There are only two studies that investigated the predictive validity of dynamic decision making measures. Wagener and Wittmann (2002) assessed a sample of $N = 35$ trainees and reported correlations between $r = 0.16$ and $r = 0.40$ between the performance in a dynamic decision making task and the performance in different assessment center tasks. However, the study did not report whether these relationships were incremental or due to an overlap between dynamic decision making and general intelligence. Kersting (2001) reported a correlation of $r = 0.37$ between the performance in a dynamic decision making task and supervisor ratings in a sample of $N = 73$ policemen. He further reported that this correlation remained significant after controlling for individual differences in general intelligence, $r = 0.29$, which points towards

the incremental predictive validity of this dynamic decision making measure.

Taken together, these findings draw a rather heterogeneous picture of the relation between dynamic decision making and general intelligence and there is only preliminary evidence for the predictive validity of dynamic decision making variables.

1.1.2. Implicit learning

Mackintosh (1998) suggested to use artificial grammar learning tasks (Reber, 1967) to measure performance differences in implicit learning. In such a task, the participants are asked to learn a list of apparently arbitrary letter strings (like WNSNXS). Afterwards, they are told that these strings were constructed according to a complex rule system (a grammar) and they are asked to judge newly presented strings as grammatical or non-grammatical. Typically, the participants show above chance performance but are not able to report the grammar rules. Therefore, Reber (1967) suggested that the participants learned the grammar implicitly. Although Reber's interpretation released a long and fertile discussion about implicit learning processes, there have been only a few studies investigating the relation between performance in artificial grammar learning tasks and general intelligence.

Reber, Walkenfeld, and Hernstadt (1991) reported a correlation of $r = 0.25$ between the performance in an artificial grammar learning task and IQ, and Gebauer and Mackintosh (2007) reported respective correlations between $r = -0.03$ and $r = 0.17$ depending on the task and the instruction. To our knowledge, there is no published study investigating the relation between educational or professional success and the performance in an artificial grammar learning task. Thus, there is a paucity of evidence on the relation between implicit learning and general intelligence as well as on the relation between implicit learning and success in real life.

1.2. Some psychometric considerations

Previous studies that investigated the relation between general intelligence, dynamic decision making, and implicit learning treated the performance measures as trait-like variables. A trait may be defined as a variable that is stable over several measurement occasions, consistent across different situations, and consistent across different methods. However, the variance of a performance measure may capture additional factors beyond individual differences in a trait.

First, a performance measure may also be influenced by the specific measurement situation even in standardized experiments. For example, one person may be well rested whereas another person may already have worked several hours before testing. One person may be motivated to show maximum performance whereas another person may have gotten a stinging rebuke by his or her supervisor that day and may not be motivated to show performance at all. Because these effects may contribute unwanted variance, it may be beneficial to take this *occasion specificity* of performance variables into account.

Second, a performance measure may be influenced by the *specific method* that is used for the assessment. Hence, there may be individual differences in a performance measurement which are triggered by the method. For example, a verbal intelligence test may capture individual differences in general intelligence as well as individual differences in speech comprehension whereas a figural intelligence test may capture individual differences in general intelligence and visual thinking. Thus, individual differences in speech comprehension or visual thinking are method specific because they can only be assessed with verbal or figural test material. Similarly, a particular dynamic decision making task may measure performance differences, which are specific for this particular task but not for dynamic decision making in general.

Third, a performance measure may be influenced by *unsystematic measurement error*. For example, instructions may be ambiguous or persons may accidentally make mistakes, which may result in a low reliability of performance measures. Because these effects may contribute unwanted variance, it seems worthwhile to investigate these factors with respect to dynamic decision making and implicit learning variables in greater detail.

These considerations have been formalized in Steyer et al.'s latent state-trait theory (Steyer, Schmitt, & Eid, 1999). In a nutshell, latent state-trait theory proposes that the measurement i of a variable Y can be decomposed into a trait ξ_i , a state residual ζ_i , a method residual η_i , and an unsystematic error residual ε_i , thus $Y_i = \xi_i + \zeta_i + \eta_i + \varepsilon_i$. Given the independence of these factors (Steyer et al., 1999), the variance of this measurement can be decomposed as $\sigma^2(Y_i) = \sigma^2(\xi_i) + \sigma^2(\zeta_i) + \sigma^2(\eta_i) + \sigma^2(\varepsilon_i)$, and the factor variances may be estimated with a structural equation model as shown in Fig. 1. As can be seen in this figure, the latent trait factor is defined as a variable that is consistent across several measurement occasions and methods, whereas the latent state residual and the method factor are specific for the individual measurement occasion and the assess-

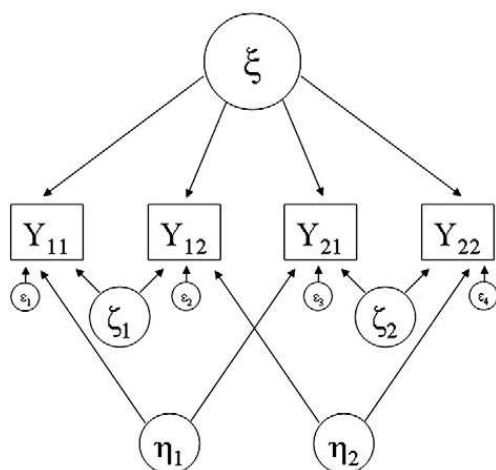


Fig. 1. Latent state-trait structural equation model. Y_{11} = variable at measurement occasion 1 with method 1, Y_{12} = variable at measurement occasion 1 with method 2, Y_{21} = variable at measurement occasion 2 with method 1, Y_{22} = variable at measurement occasion 2 with method 2, ξ = trait variable, ζ_1 = state residual 1, ζ_2 = state residual 2, η_1 = method residual 1, η_2 = method residual 2, ε_1 = error 1, ε_2 = error 2, ε_3 = error 3, ε_4 = error 4.

ment method, respectively. Hence, these models allow to separate the different contributions of the trait, the measurement occasion, and the measurement method to the manifest variables.

There have been many applications of latent state-trait models in different domains of personality research, which demonstrated substantial effects of the measurement occasion or the method on behavioral variables (e.g., Eid, Notz, Steyer, & Schwenkmezger, 1994; Schmitt & Steyer, 1993; Steyer, Schwenkmezger, & Auer, 1990; Yasuda, Lawrenz, Whitlock, Lubin, & Lei, 2004; Ziegler, Ehrlenspiel, & Brand, 2009) and physiological variables (e.g., Hagemann, Hewig, Seifert, Naumann, & Bartussek, 2005; Hermes et al., 2009). However, there have been no applications of latent state-trait models on performance variables yet, even if some findings suggest that it may be instructive to consider the occasion specificity and method specificity of these variables.

For example, in some studies the participants completed the same dynamic decision making task for several times (Süß, Kersting, & Oberauer, 1993; Wittmann & Hatrup, 2004) and the performance between subsequent task correlated only moderately (between $r = 0.37$ and $r = 0.62$). This points either towards a low reliability or towards a substantial occasion specificity of the variables. Moreover, Wirth and Klieme (2003) reported structural equation models, which implied a correlation of $r = 0.33$ between two dynamic decision making tasks ($r = 0.47$ when corrected for attenuation) and Gebauer and Mackintosh (2007) reported a correlation of $r = 0.15$ between two artificial grammar learning tasks ($r = 0.21$ when corrected for attenuation). These findings suggest a substantial method specificity of performance measures. Therefore, a further aim of the present study was to investigate the occasion specificity and the method specificity of dynamic decision making and implicit learning variables.

1.3. The present study

The present study investigated the psychometric properties of general intelligence, dynamic decision making, and implicit learning measures within the framework of latent state-trait theory. Therefore, each construct was measured with two methods at two measurement occasions. A further scope of this study was the relation between the respective trait variables and real life performance. We expected that general intelligence is a powerful predictor of professional success and we further expected that there are individual differences beyond IQ that are also able to predict professional success.

2. Method

2.1. Participants

There were $N = 173$ employees (113 females, 47 males, 13 not reported) completing the first measurement occasion and $N = 151$ completing the second measurement occasion. The participants were recruited via newspaper announcement from different branches and different companies around Heidelberg. The participants' jobs were rated according to the

International Standard Classification of Occupations (ISCO-88 COM). 6% rated themselves as legislators, senior officials, and managers, 25% as professionals, 11% as technical and associate professionals, 14% as clerks, 40% as service workers and shop and market sales workers, 1% as craft and related trade workers, 1% as plant and machine operators and assemblers, and 1% as elementary occupations. The participants' mean age was $M = 43.34$ ($SD = 11.22$).

2.2. Measures

2.2.1. Advanced progressive matrices (APM)

The APM (Raven, Court, & Raven, 1994) were used as an indicator for participants' general intelligence. A computer adapted version of the test was administered. According to the test manual, the number of solved items of the second set was taken as a performance indicator. These raw scores were transformed to z-scores for further analysis, because the APM and the Berlin Intelligence Structure Test were scaled differently.

2.2.2. Berlin intelligence structure test (BIS)

The short version of the BIS (Jäger, Süß, & Beauducel, 1997) was used as a second indicator of general intelligence. The BIS consists of a variety of tasks like an analogical reasoning task, a visual memory task, and a numerical series task (for an English description, see Süß, Oberauer, Wittmann, Wilhelm, & Schulze, 2002). The test was administered and the raw scores were computed according to the test manual. We did not compute IQ scores because there is no adult normative sample for the BIS. For further analysis the raw scores were transformed to z-scores.

2.2.3. Artificial grammar learning tasks

Implicit learning was measured with two artificial grammar learning tasks (Reber, 1967). The procedure and the stimuli were adopted from Gebauer and Mackintosh (2007). The artificial grammar learning tasks consisted of a learning phase and a testing phase. In the *learning phase*, 30 letter strings were presented and the participants were instructed to memorize them. Each string was presented individually for 3 s on a 17 in. screen of a personal computer (e.g., WNSNXS). The participants were asked to repeat the strings correctly by pressing the respective letters on the keyboard. When a string was repeated correctly, the feedback "correct" was given and the next string occurred. When a string was repeated incorrectly, the feedback "false" was given and the string was displayed again until repeated correctly. After a participant repeated ten strings correctly, these ten strings were simultaneously displayed for 90 s on the screen and the participant was asked to repeat them silently. After a participant repeated all 30 strings correctly the learning phase was finished and the participant was informed that all strings in the learning phase were constructed according to a complex rule system. In the *testing phase*, 80 new strings were presented (see Appendix A). There were 40 grammatical strings that were constructed according to the same rule system as the strings in the learning phase (e.g., WNSWW). In addition, there were 40 non-grammatical strings that contained one letter at a

position that violated the rule system (e.g., NTSWWN). The participants were instructed to judge the letter strings as grammatical or non-grammatical. To judge a string as grammatical, the participants had to press the A-key of the keyboard, to judge a string as non-grammatical, the L-key. The order of presentation of the strings was fixed across the participants in a random order. The percentage of correct judgments in the testing phase was taken as the performance indicator. The stimuli for the first artificial grammar learning task were constructed according to Fig. 2. The stimuli for the second artificial grammar learning task were constructed according to Fig. 3.

2.2.4. Tailorshop

The Tailorshop simulation (Funke, 1983) was used as a dynamic decision making task. The Tailorshop is a computer based scenario and requests the participants to lead a fictional company which produces and sells shirts for twelve simulated months. Several variables can be manipulated like the number of workers, the expenses for advertising etc. (see Fig. 4). The state of a variable in a given month influences the state of the same and other variables in the following month

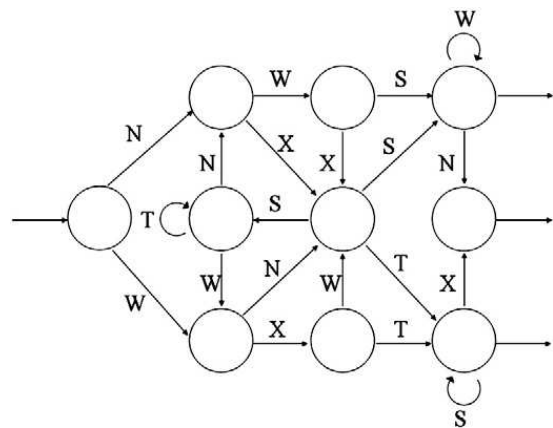


Fig. 2. Grammar 1 that was used in the first artificial grammar learning task.

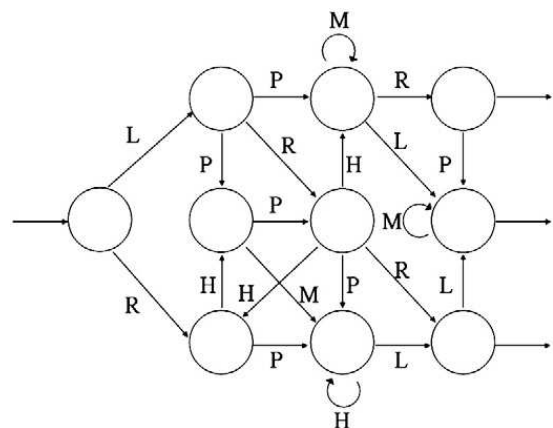


Fig. 3. Grammar 2 that was used in the second artificial grammar learning task.

but the participants do not know how the variables are connected (for a more detailed description see Funke, 1983, 2010). The participants completed a training phase, a knowledge test, and a control phase. In the training phase the participants controlled the system for six simulated months and were instructed to find out as much as possible about the scenario. The knowledge test consisted of twelve questions that measured how much the participants learned about the Tailorshop so far. In the control phase the participants were instructed to maximize their company value during twelve simulated months. For the purpose of the present study only data from the control phase were analyzed. The percentage of months with an increase in the company value between the second and the twelfth month was taken as the performance indicator, because Danner et al. (2011) have shown that this is a reliable and valid performance indicator.

2.2.5. Heidelberg finite state automaton (HFA)

The HFA (Wirth & Funke, 2005) was taken as a second indicator for dynamic decision making. The scenario is computer based and simulates a space flight where the participants can control a space ship and a vehicle with a user interface (see Fig. 5). The scenario consists of a training phase, a knowledge test, and a control phase. During the 15 minute training phase the participants were instructed to find out how to control the space ship and the vehicle. The knowledge test consists of 16 items and measures how much the participants have learned about the system so far. The control phase consists of 22 items where a target state is given which the participants have to reach by controlling the system (e.g., landing the space ship on a specified planet). For the purpose of the present study, only data from the control phase were analyzed. The percentage of correctly solved items was taken as the performance indicator.

2.2.6. Professional success

The participants' professional success was measured with two instruments. *Objective professional success* was

measured by the participants' income (thirteen categories), self-rated social status (seven categories), and the participants' highest educational attainment (nine categories). To adjust for different scaling, the three variables were z-transformed ($M=0, SD=1$) for further analysis. In addition, professional success was measured by *supervisor ratings* with five items (e.g., "The employee demonstrates competence in all job-related tasks") on a six-point Likert scale.

2.3. Procedure

There were two measurement occasions. The first measurement occasion started in July 2009 (till September 2009) and consisted of session 1 and session 2. Both sessions took place within one week for each participant. The second measurement occasion started in December 2009 (till February 2010) and consisted of session 3 and session 4, which also took place within one week. The participants were assessed in small groups of not more than four persons. Each session took approximately 2.5 h.

The participants completed the same tasks at both measurement occasions. During session 1 (and session 3) the participants completed an artificial grammar learning task with grammar 1, the APM, and the Heidelberg Finite State Automaton. During session 2 (and session 4), the participants completed an artificial grammar learning task with grammar 2, the short version of the BIS, and the Tailorshop simulation. After the first session, each participant received an envelope with a questionnaire for his or her supervisor. During the third session, the participants additionally completed a questionnaire about their professional success.

2.4. Statistical analysis

To investigate the relations between the variables, we used structural equation models. The parameters of the models were estimated using the maximum likelihood

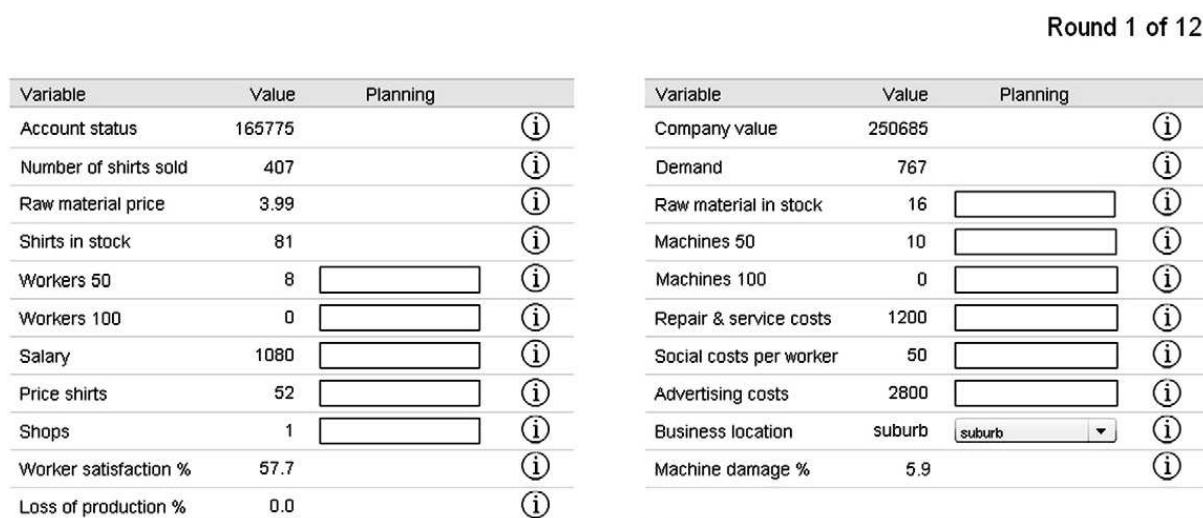


Fig. 4. Screenshot of the graphical user interface of the Tailorshop (labels translated).

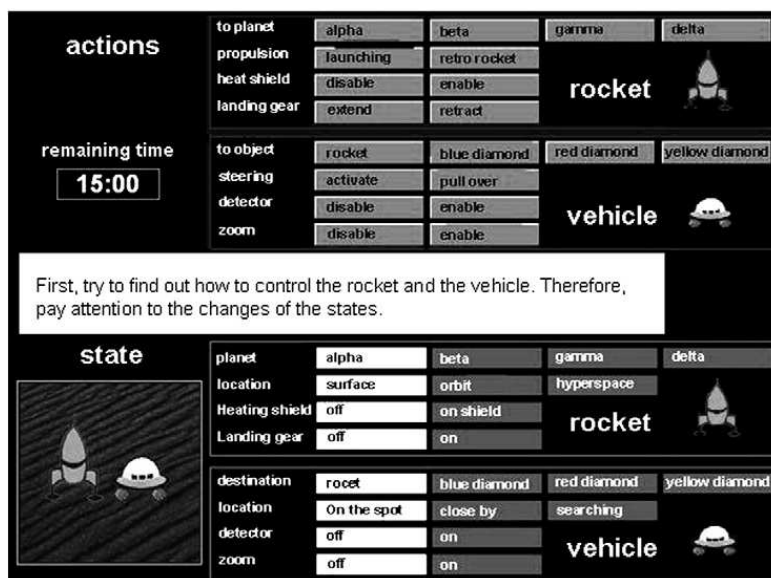


Fig. 5. Screenshot of the graphical user interface of the Heidelberg Finite State Automaton (labels translated).

algorithm implemented in Amos 18 (Arbuckle, 2006). In a first step, we investigated latent state-trait measurement models separately for intelligence, dynamic decision making, and implicit learning. In a second step, we investigated the correlation between the latent trait variables. In a third step, we performed a latent regression analysis to investigate relations between the constructs in greater detail.

3. Results

3.1. Raw scores

The raw scores of the measurements are reported in Table 1. The number of solved items in the Advanced Progressive Matrices at the first measurement occasion was $M=21.64$ ($SD=5.80$), which corresponds to an IQ of $M=100.62$ ($SD=22.55$). There are no normative samples for the Berlin Intelligence Structure Test, the Tailorshop, the Heidelberg Finite State Automaton, or the artificial grammar learning tasks. However, the present scores are similar to previous results. The mean score of the BIS was $M=96.30$ ($SD=6.21$) at the first measurement occasion and $M=99.21$ ($SD=6.38$) at the second measurement occasion. According to Jäger et al. (1997), a mean score of $M=100$ corresponds to an average performance. In the present study, the participants solved $M=10.79$ ($SD=5.80$) HFA items at the first measurement occasion and $M=13.44$ ($SD=5.95$) HFA items at the second measurement occasion. This result is similar to Wirth and Klieme (2003), who reported that their participants solved $M=11$ HFA items on average. The judgment accuracy in the artificial grammar learning tasks varied between $M=61.58$ ($SD=7.11$) and $M=63.90$ ($SD=7.24$), which corresponds to the findings of Gebauer and Mackintosh, who reported mean accuracies between $M=59.16$ ($SD=8.59$) and $M=69.93$ ($SD=7.52$)

for the same artificial grammar learning tasks that were used in the present study.

3.2. Measurement models

We used a basic latent state-trait model (Steyer et al., 1999) with a state residual ζ for each measurement occasion and a method factor η for each instrument to control for effects of the measurement occasion and method effects (see Fig. 1). All path coefficients were fixed to one and the variances of all latent variables were estimated. If a first estimation revealed negative or non-significant variances, then these variances were fixed to zero and the model was estimated again.

3.2.1. Intelligence

A first analysis of the basic model revealed a good model fit, $\chi^2(1)=0.30$, $p=0.569$, $RMSEA=0.00$, $CFI=1.00$. However,

Table 1
Mean and standard deviation of raw scores.

Task	Measurement occasion 1		Measurement occasion 2	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
APM	21.64	5.80	22.94	7.02
BIS	96.30	6.21	99.21	6.38
Tailorshop	2.68	3.21	3.15	3.77
HFA	10.79	5.80	13.44	5.95
AGL1	61.58	7.11	62.83	6.87
AGL2	63.90	7.24	62.20	7.70

Note. APM = number of solved items in the Advanced Progressive Matrices, BIS = scores in Berlin Intelligence Structure Test, Tailorshop = number of months with an increase in the company value, HFA = number of items solved in the Heidelberg Finite State Automaton, AGL1 = percent of correct judgments in the artificial grammar learning task with grammar 1, AGL2 = percent of correct judgments in the artificial grammar learning task with grammar 2.

the estimated variance for ζ_1 was negative ($\zeta_1 = -14.14$, $p = 0.016$), and the estimated variance for ζ_2 was not significant ($\zeta_2 = 9.60$, $p = 0.125$). Therefore, these parameters were set to zero and the model was estimated again. The modified model fitted the data well, $\chi^2(3) = 5.59$, $p = 0.133$, RMSEA = 0.07, CFI = 1.00, and the difference in the fit of the models was not significant, $\Delta\chi^2(2) = 4.29$, $p = 0.117$. Therefore, this model could be accepted. The estimated model parameters are reported in Table 2.

3.2.2. Dynamic decision making

The basic latent state-trait model fitted well with the data, $\chi^2(1) = 0.9$, $p = 0.335$, RMSEA = 0.00, CFI = 1.00. However, the latent state residuals were negative ($\zeta_1 = -49.20$, $p = 0.183$) or non-significant ($\zeta_1 = 48.30$, $p = 0.257$). The modified model without latent state residuals also fitted well with the data, $\chi^2(3) = 3.25$, $p = 0.355$, RMSEA = 0.02, CFI = 1.00; $\Delta\chi^2(2) = 2.35$, $p = 0.309$. Thus, this model could be accepted. The estimated model parameters are presented in Table 2.

3.2.3. Implicit learning

The basic latent state-trait model fitted well with the data, $\chi^2(1) = 0.13$, $p = 0.719$, RMSEA = 0.00, CFI = 1.00. However, the variances of the latent state residual and the latent method variables were non-significant ($\zeta_1 = 6.57$, $p = 0.128$; $\zeta_2 = 2.35$, $p = 0.585$; $\eta_1 = -0.06$, $p = 0.988$; $\eta_2 = -4.96$, $p = 0.250$). Therefore, these variances were set to zero. This modified model fitted the data well, $\chi^2(5) = 3.19$, $p = 0.671$, RMSEA = 0.00, CFI = 1.00; $\Delta\chi^2(4) = 3.06$, $p = 0.548$, and this model was accepted. The estimated model parameters are presented in Table 2.

3.2.4. LST parameters

Based on these estimates, several latent state-trait parameters may be computed such as coefficients of reliability, trait-specificity (also referred to as consistency), occasion-specificity, and method-specificity. These parameters have a range between zero and one, and a greater value indicates a greater specificity. The reliability coefficient of a measurement i reveals how great the proportion of systematic variance in this measurement is. It is computed as $[\sigma^2(\xi_i) + \sigma^2(\zeta_i) + \sigma^2(\eta_i)]/\sigma^2(Y_i)$. The

trait-specificity coefficient of a measurement i reveals how great the proportion of trait differences in a measurement is. It may be computed as $\sigma^2(\xi_i)/\sigma^2(Y_i)$. The occasion-specificity coefficient of a measurement i indicates the effects of the situation and the interaction between the situation and the person on the measurement. It may be computed as $\sigma^2(\zeta_i)/\sigma^2(Y_i)$. The method-specificity coefficient of a measurement i reveals how great the proportion of individual differences is due to the method (e.g., task) used. This coefficient is computed as $\sigma^2(\eta_i)/\sigma^2(Y_i)$.

These parameters are presented in Table 3. As can be seen, the general intelligence measurements revealed great reliabilities, great trait-specificities, and low method-specificities. The Heidelberg Finite State Automaton measurements also showed great reliabilities, but smaller trait-specificities and greater method-specificities. The Tailorshop measurements revealed small reliabilities and small trait-specificities. All implicit learning measurements revealed very small reliabilities and trait-specificities. Since all measurement models fitted well without state residuals, the estimated occasion-specificity was zero for all measurements.

3.2.4. Professional success

Objective professional success was measured with three indicators at session 3. A measurement model with one latent success variable, equal path coefficients ($\beta = 1$), and a latent error variable for each manifest variable was specified. The model fitted the data well, $\chi^2(2) = 2.46$, $p = 0.293$, RMSEA = 0.04, CFI = 0.98. Therefore, this model was accepted. The composite reliability (Raykov, 1997) of the items' mean score was 0.71. The participants' supervisor ratings were measured with a five item questionnaire. A measurement model with one latent success variable, equal path coefficients ($\beta = 1$), and a latent error variable for each manifest variable fitted the data well, $\chi^2(9) = 11.93$, $p = 0.217$, RMSEA = 0.04, CFI = 0.99. Thus, this model was accepted. The composite reliability of the items' mean score was 0.95.

Table 2
Estimated variances for measurement models (p-values in brackets).

	Intelligence	Dynamic decision making	Implicit learning
ξ	0.73 (<0.001)	317.12 (<0.001)	14.87 (<0.001)
ζ_1	0 (fixed)	0 (fixed)	0 (fixed)
ζ_2	0 (fixed)	0 (fixed)	0 (fixed)
η_1	0.14 (0.015)	144.66 (0.046)	0 (fixed)
η_2	0.24 (<0.001)	257.37 (<0.001)	0 (fixed)
ε_1	0.14 (<0.001)	425.17 (<0.001)	35.92 (<0.001)
ε_2	0.18 (<0.001)	637.27 (<0.001)	33.38 (<0.001)
ε_3	0.11 (<0.001)	146.00 (<0.001)	35.29 (<0.001)
ε_4	0.06 (0.014)	145.21 (<0.001)	43.83 (<0.001)

Note. ξ = trait variable, ζ_1 = state residual 1, ζ_2 = state residual 2, η_1 = method residual 1, η_2 = method residual 2, ε_1 = error 1, ε_2 = error 2, ε_3 = error 3, ε_4 = error 4. The different scaling of the variables affects the magnitude of the variances estimates.

Table 3
Reliability, trait- and method-specificity of measurements.

Task	Measurement occasion	Reliability	Trait-specificity	Method-specificity
APM	1	0.86	0.72	0.14
APM	2	0.83	0.70	0.13
BIS	1	0.90	0.67	0.22
BIS	2	0.95	0.71	0.24
Tailorshop	1	0.52	0.36	0.16
Tailorshop	2	0.42	0.29	0.13
HFA	1	0.80	0.44	0.36
HFA	2	0.80	0.44	0.36
AGL1	1	0.29	0.29	0.00
AGL1	2	0.31	0.31	0.00
AGL2	1	0.30	0.30	0.00
AGL2	2	0.25	0.25	0.00

Note. APM = Advances Progressive Matrices, BIS = Berlin Intelligence Structure Test, HFA = Heidelberg Finite State Automaton, AGL1 = artificial grammar learning task with grammar 1, AGL2 = artificial grammar learning task with grammar 2.

3.3. Relations between intelligence, dynamic decision making, implicit learning, and professional success

We specified an omnibus model, which simultaneously tested all measurement models described above and allowed free correlations between the latent trait variables and the latent professional success variables. The specified model revealed a good model fit, $\chi^2(174) = 197.74$, $p = 0.105$, RMSEA = 0.03, CFI = 0.98 and thus was accepted. The correlations between the latent variables are shown in Table 4. As can be seen, there were significant and substantial correlations between all performance variables. The greatest correlation was between intelligence and dynamic decision making, $r = 0.86$, $p < 0.001$. There was also a correlation of $r = 0.78$, $p < 0.001$ between objective professional success and general intelligence. There were further substantial correlations between objective professional success and dynamic decision making, $r = 0.52$, $p < 0.001$, and between objective professional success and implicit learning, $r = 0.31$, $p = 0.030$. The only significant correlation with supervisor ratings was the correlation with dynamic decision making, $r = 0.25$, $p = 0.021$.

3.4. Prediction of objective professional success

To investigate the relation between performance variables and objective professional success in greater detail, we specified a latent regression model according to Fig. 6. As can be seen, dynamic decision making, implicit learning, and professional success were regressed on intelligence. The residuals of this regression are the proportions of trait variances which are independent from general intelligence. The dynamic decision making and implicit learning residuals were used to predict the proportion of construct variance in objective professional success that could not be explained by general intelligence.

The specified model revealed a good model fit, $\chi^2(95) = 114.44$, $p = 0.085$, RMSEA = 0.03, CFI = 0.98. The standardized path coefficients are shown in Fig. 6. As can be seen, dynamic decision making as well as implicit learning revealed trait variances, which were independent from general intelligence. In addition, general intelligence was the only significant predictor of objective professional success. Neither the path coefficient from the residual dynamic decision variable to the residual professional success variable, nor the path coefficient from the residual implicit learning variable to the residual professional success variable was significant. Therefore, these path coefficients were set to zero and the model was estimated again. The modified model also revealed a good model fit, $\chi^2(97) =$

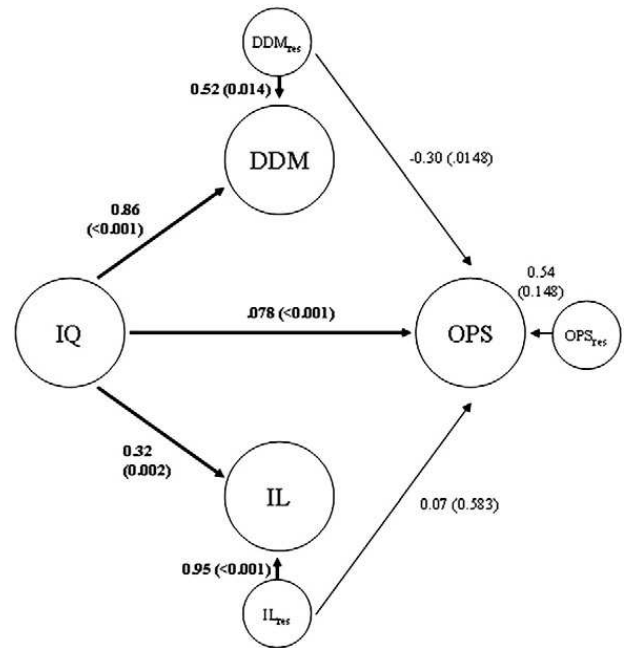


Fig. 6. Latent Regression Analysis with standardized path coefficients (p -values in brackets). IQ = latent general intelligence variable, DDM = latent dynamic decision making variable, IL = latent implicit learning variable, OPS = latent objective professional success variable, DDM_{res} = latent residual for dynamic decision making, IL_{res} = latent residual for implicit learning, OPS_{res} = latent residual for professional success.

117.62, $p = 0.076$, RMSEA = 0.04, CFI = 0.98; $\Delta\chi^2(2) = 3.18$, $p = 0.204$. Thus, this model was accepted.

3.5. Prediction of supervisor ratings

The relations between general intelligence, dynamic decision making, implicit learning, and supervisor ratings were investigated analogously to the analysis described above. The specified model fitted the data well, $\chi^2(126) = 125.86$, $p = 0.487$, RMSEA = 0.00, CFI = 1.00. The standardized path coefficients are shown in Fig. 7. As can be seen, dynamic decision making was the only significant predictor of participants' supervisor ratings. Neither the path coefficient from the general intelligence variable, nor the path coefficient from the residual implicit learning variable was significant. A modified model, which fixed these parameters to zero, revealed an adequate model fit, $\chi^2(128) = 126.00$, $p = 0.533$, RMSEA = 0.00, CFI = 1.00; $\Delta\chi^2(2) = 0.14$, $p = 0.932$. Therefore, this model was accepted.

Table 4 Correlation between latent success and latent trait variables (p -values in brackets).

	Intelligence	Dynamic decision making	Implicit learning	Objective professional success
Dynamic decision making	0.86 (<math>p < 0.001</math>)			
Implicit learning	0.32 (0.005)	0.26 (0.033)		
Objective professional success	0.78 (<math>p < 0.001</math>)	0.52 (<math>p < 0.001</math>)	0.31 (0.030)	
Supervisor ratings	0.03 (0.760)	0.25 (0.021)	-0.02 (0.871)	-0.07 (0.559)

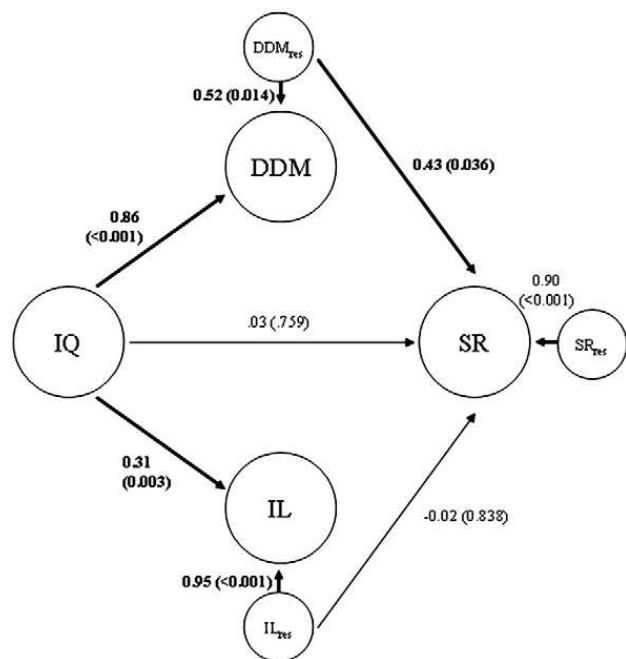


Fig. 7. Latent Regression Analysis with standardized path coefficients (*p*-values in brackets). IQ = latent general intelligence variable, DDM = latent dynamic decision making variable, IL = latent implicit learning variable, SR = latent supervisor rating variable, DDM_{res} = latent residual for dynamic decision making, IL_{res} = latent residual for implicit learning, SR_{res} = latent residual for supervisor ratings.

4. Discussion

The present study investigated Dörner's (1980) and Mackintosh's (1998) hypotheses that dynamic decision making and implicit learning are cognitive abilities that are independent from general intelligence.

In a first step, we analyzed the psychometric properties of intelligence variables, dynamic decision making variables, and implicit learning variables within the framework of latent state-trait theory. All measurement models fitted well without latent state residuals. This indicates that the performance measures were not affected by situational factors such as individual differences in fatigue or individual differences in the form of the day. Furthermore, the general intelligence variables revealed high trait specificities and low method specificities, which indicate a high proportion of trait differences in these performance measures. The dynamic decision making and implicit learning variables, on the other hand, revealed lower trait specificities and greater method specificities, which suggests that these variables capture task specific performance differences as well. However, even if the trait specificities were small, the variances of the latent trait variables were still significant. This indicates that there are true individual trait differences in dynamic decision making and implicit learning.

In a second step, we analyzed the relations between these latent trait variables. The present results suggest that there are substantial relations between general intelligence, dynamic decision making, and implicit learning. In particular, there was a great correlation ($r = 0.86$) between the

latent general intelligence variable and the latent dynamic decision making variable. This result goes in line with previous findings of Wirth and Klieme (2003), Wittmann and Hattrup (2004), and Kröner et al. (2005) who also reported great relations between measures of dynamic decision making and measures of general intelligence. Taken together, these findings contradict Dörner's hypothesis that dynamic decision making and general intelligence are independent variables.

The correlation between the latent implicit learning variable and the latent general intelligence variable was of medium size ($r = 0.32$). This goes in line with the findings of Reber et al. (1991) and Gebauer and Mackintosh (2007) who also reported low to medium correlations between measures of implicit learning and general intelligence. This finding does not support Mackintosh's hypothesis that implicit learning and general intelligence are independent constructs. However, general intelligence could only explain 10.24% of the implicit learning trait variance, which suggests that there are substantial individual differences in implicit learning beyond IQ.

Taken together, this pattern of result suggests that there are substantial relations between cognitive performance measures, which have been developed within very different domains. Measures of general intelligence have a long research tradition and were developed to measure persons' general mental ability. Measures of dynamic decision making arose in the domain of complex problem solving and were designed to explore persons' ability to deal with realistic problems. And measures of implicit learning were developed in the domain of cognitive psychology in order to study persons' ability in making intuitive decisions. The present findings suggests that these performance measures share a substantial proportion of common variance but also reveal variance proportions that are independent from each other. This fits well with hierarchical intelligence models like Carroll's (1993) three-stratum theory of cognitive abilities. In particular, Carroll suggested that the structure of human cognitive abilities may be explained by a hierarchical structure with three levels (three strata). On the lowest level (stratum 1) there are 64 different specific ability factors like reading comprehension, memory span, or general sound discrimination. According to Carroll, these specific abilities are not independent and therefore may be grouped together to eight more general ability factors (stratum 2), which are fluid intelligence, crystallized intelligence, general memory and learning, broad visual perception, broad auditory perception, broad retrieval ability, broad cognitive speediness, and processing speed. On the top of the hierarchy (stratum 3) there is a single general ability factor that explains the correlation between the stratum 2 factors. In Carroll's model there are no ability factors such as dynamic decision making or implicit learning. Accordingly, these constructs may be seen as supplementary aspects of human cognitive ability. However, the present results fit well with the concept of a hierarchical structure of human cognitive ability. In particular, the results of the structural equation models revealed that the overlap between the performance in the Tailorshop and the Heidelberg Finite Automaton may be

explained by a more general dynamic decision making ability factor. In the same vein, the overlap between the different artificial grammar learning tasks could be explained by an implicit learning ability factor. Furthermore, there were substantial correlations between general intelligence, dynamic decision making and implicit learning that could be explained by one single general ability factor. Taken together, these results suggest that dynamic decision making and implicit learning may be supplementary abilities that fit well into a hierarchical concept of human cognitive ability. However, the present findings do not sufficiently allow to draw a conclusion on which stratum these ability factors may be located. Investigating this may be an interesting issue for future research.

In a third step, we analyzed whether dynamic decision making and implicit learning are powerful predictors of professional success beyond IQ. The zero correlation between objective professional success and supervisor ratings ($r=0.07$) suggests that both variables capture different aspects of professional success. One reason for this may be that income, social status, and education attainment are rather profit-based indicators, whereas supervisor ratings may also capture social aspects. According to this, both aspects were analyzed separately.

There were substantial correlations between *objective professional success* and dynamic decision making ($r=0.52$) as well as between objective professional success and implicit learning ($r=0.31$). This suggests that both performance measures are able to predict objective professional success. However, when general intelligence was included as a predictor, then general intelligence remained the only significant predictor ($\beta=0.78$). This finding is consistent with the literature and emphasizes the meaningfulness and usefulness of IQ measures (e.g., Schmidt & Hunter, 2004).

There was a substantial relation between the participants' *supervisor ratings* and dynamic decision making even when general intelligence was simultaneously considered ($\beta=0.43$). This replicates findings of Kersting (2001) who also reported an incremental predictive value of dynamic decision making measures on participants' supervisor ratings. Furthermore, this result points towards the practical value of dynamic decision making measures and suggests that dynamic decision making measures may provide insights into aspects of professional success, which cannot be predicted by general intelligence. Therefore, Dörner's hypothesis that dynamic decision making has an incremental predictive value is partially supported. The relation between supervisor ratings and implicit learning was close to zero ($r=-0.02$) and not significant. Thus, this result may be seen as preliminary evidence against Mackintosh's hypothesis that implicit learning is a useful predictor of professional success. There was no significant correlation between supervisor ratings and general intelligence. At first sight, this finding is astonishing because there is a wealth of evidence for the relation between general intelligence and supervisor rating (e.g., Ng et al., 2005; Salgado et al., 2003; Schmidt & Hunter, 2004). However, the samples in these studies typically consist of employees within a single department or company whereas the sample in the present study consisted of

employees of different companies and occupational groups. In particular, there may be a relation between general intelligence and supervisor ratings within single companies or occupational groups but not between. For example, a broker with an IQ of 130 may be rated as more successful than a broker with an IQ of 100 but a journalist with an IQ of 130 may still be rated as less successful than the broker with the IQ of 100.

4.1. Implications for assessment

The present results show that the APM as well as the Berlin Intelligence Structure Test yield measures with good trait specificities (0.67 to 0.72). Furthermore, there was a strong relation ($r=0.78$) between general intelligence and objective professional success. Therefore, general intelligence tests seem to be a good choice for measuring cognitive ability.

There was also a relation between the dynamic decision making trait variable and objective professional success ($r=0.52$) and between the dynamic decision making trait variable and supervisor ratings ($r=0.25$). However, the performance measures of the Tailorshop simulation and the Heidelberger Finite State Automaton showed trait specificities between 0.29 and 0.44. This suggests that less than half of the variance in these performance measures is due to trait differences in dynamic decision making. Therefore, the trait-specificity of both tasks should be improved before they are used for an individual assessment. A more theory-orientated development of dynamic decision making tasks may help to reach this goal.

There was a relation of medium size between the implicit learning trait variable and objective professional success ($r=0.31$). However, the latent regression analysis revealed that this relation was due to an overlap with general intelligence. This suggests that there is no incremental predictive value of implicit learning measures. The trait specificities of the artificial grammar learning measures were between 0.25 and 0.31. There was no method specificity of these variables, which suggests that the low trait specificity was due to unsystematic measurement error. Therefore, lengthening the test may help to enhance the trait-specificity. However, whether such an approach increases the reliability or rather causes fatigue effects is an open issue.

5. Conclusion

The present findings acknowledge the overall approval and usefulness of general intelligence measures. In addition, the results demonstrated that there are significant individual trait differences in cognitive performance beyond IQ. In particular, there was a large proportion of trait variance in implicit learning, which was independent from general intelligence and in addition, dynamic decision making revealed an incremental predictive validity. These findings make dynamic decision making as well as implicit learning attractive for the research of individual differences.

References

- Arbuckle, J. L. (2006). *Amos 7.0 user's guide*. Chicago: SPSS.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.
- Danner, D., Hagemann, D., Holt, D., Bechtold, M., Schankin, A., Wüstenberg, S., & Funke, J. (2011). *Measuring performance in dynamic decision making: reliability and validity of the Tailorshop simulation*. *Journal of Individual Differences*. doi:10.1027/1614-0001/a000055.
- Dörner, D. (1980). On the difficulty people have in dealing with complexity. *Simulation & Gaming*, 11, 87–106.
- Dörner, D. (1986). Diagnostik der Operativen Intelligenz. *Diagnostica*, 32, 290–308.
- Eid, M., Notz, P., Steyer, R., & Schwenkmezger, P. (1994). Validating scales for the assessment of mood level and variability by latent state-trait analyses. *Personality and Individual Differences*, 16, 63–76. doi:10.1016/0191-8869(94)90111-2.
- Funke, J. (1983). Einige Bemerkungen zu Problemen der Problemlöseforschung oder: Ist Testintelligenz doch ein Prädiktor? *Diagnostica*, 29, 283–302.
- Funke, J. (2010). Complex problem solving: A case for complex cognition? *Cognitive Processing*, 11, 133–142. doi:10.1007/s10339-009-0345-0.
- Gebauer, G. F., & Mackintosh, N. J. (2007). Psychometric intelligence dissociates implicit and explicit learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 33, 34–54. doi:10.1037/0278-7393.33.1.34.
- Gonzalez, C., Vanyukov, P., & Martin, M. K. (2005). The use of microworlds to study dynamic decision making. *Computers in Human Behavior*, 21, 273–286. doi:10.1016/j.chb.2004.02.014.
- Hagemann, D., Hewig, J., Seifert, J., Naumann, E., & Bartussek, D. (2005). The latent state-trait structure of resting EEG asymmetry: Replication and extension. *Psychophysiology*, 42, 740–752. doi:10.1111/j.1469-8986.2005.00367.x.
- Hermes, M., Hagemann, D., Britz, P., Lieser, S., Bertsch, K., & Naumann, E. (2009). Latent state-trait structure of cerebral blood flow in a resting state. *Biological Psychology*, 80, 196–202. doi:10.1016/j.biopsycho.2008.09.003.
- Jäger, A. O., Süß, H. -M., & Beauducel, A. (1997). *Berliner Intelligenzstruktur – Test. Form 4*. Göttingen: Hogrefe.
- Kersting, M. (2001). Zur Konstrukt- und Kriteriumsvalidität von Problemlöse-szenarien anhand der Vorhersage von Vorgesetztenurteilen über die berufliche Bewährung. *Diagnostica*, 47, 67–76. doi:10.1026//0012-1924.47.2.67.
- Kluwe, R. H., Misiak, C., & Haider, H. (1991). The control of complex systems and performance in intelligence tests. In H. A. H. Rowe (Ed.), *Intelligence: Reconceptualization and measurement* (pp. 227–244). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Kröner, S., Plass, J. L., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence*, 33, 347–368. doi:10.1016/j.intell.2005.03.002.
- Mackintosh, N. J. (1998). *IQ and human intelligence*. Oxford: Oxford University Press.
- Ng, T. W. H., Eby, L. T., Sorensen, K. L., & Feldman, D. C. (2005). Predictors of objective and subjective career success. A meta-analysis. *Personnel Psychology*, 58, 367–408. doi:10.1111/j.1744-6570.2005.00515.x.
- Raven, J. C., Court, J. H., & Raven, J. (1994). *Manual for Raven's progressive matrices and mill hill vocabulary scales. Advanced progressive matrices*. Oxford: Oxford Psychologists Press.
- Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, 21, 173–184. doi:10.1177/01466216970212006.
- Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning & Verbal Behavior*, 6, 855–863. doi:10.1016/s0022-5371(67)80149-x.
- Reber, A. S., Walkenfeld, F. F., & Hernstadt, R. (1991). Implicit and explicit learning: Individual differences and IQ. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 17, 888–896. doi:10.1037/0278-7393.17.5.888.
- Rigas, G., Carling, E., & Brehmer, B. (2002). Reliability and validity of performance measures in microworlds. *Intelligence*, 30, 463–480. doi:10.1016/s0160-2896(02)00121-6.
- Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European community. *Journal of Applied Psychology*, 88, 1068–1081. doi:10.1037/0021-9010.88.6.1068.
- Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology*, 86, 162–173. doi:10.1037/0022-3514.86.1.162.
- Schmitt, M. J., & Steyer, R. (1993). A latent state-trait model (not only) for social desirability. *Personality and Individual Differences*, 14, 519–529. doi:10.1016/0191-8869(93)90144-r.
- Spearman, C. (1904). 'General intelligence', objectively determined and measured. *The American Journal of Psychology*, 15, 201–293. doi:10.2307/1412107.
- 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. doi:10.1002/(sici)1099-0984(199909/10)13:5<389::aid-per361>3.0.co;2-a.
- Steyer, R., Schwenkmezger, P., & Auer, A. (1990). The emotional and cognitive components of trait anxiety: A latent state-trait model. *Personality and Individual Differences*, 11, 125–134. doi:10.1016/0191-8869(90)90004-b.
- Süß, H. -M., Kersting, M., & Oberauer, K. (1993). Zur Vorhersage von Steuerungsleistungen an computersimulierten Systemen durch Wissen und Intelligenz. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 14, 189–203.
- Süß, H. -M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability – and a little bit more. *Intelligence*, 30, 261–288. doi:10.1016/s0160-2896(01)00100-3.
- Wagener, D. (2001). *Psychologische Diagnostik mit komplexen Szenarios*. Taxonomie, Entwicklung, Evaluation. Lengerich: Pabst Science Publishers.
- Wagener, D., & Wittmann, W. W. (2002). Personalarbeit mit dem komplexen Szenario FSYS: Validität und Potential von Verhaltensskalen. *Zeitschrift für Personalpsychologie*, 1, 80–93. doi:10.1026//1617-6391.1.2.80.
- Wallach, D. (1998). *Komplexe Regelungsprozesse*. Eine kognitionswissenschaftliche Analyse. Wiesbaden: Deutscher Universitäts-Verlag.
- Wirth, J., & Funke, J. (2005). Dynamisches Problemlösen: Entwicklung und Evaluation eines neuen Messverfahrens zum Steuern komplexer Systeme. In E. Klieme, D. Leutner, & J. Wirth (Eds.), *Problemlösekompetenz von Schülerinnen und Schülern*. Wiesbaden: Verlag für Sozialwissenschaften.
- Wirth, J., & Klieme, E. (2003). Computer-based assessment of problem solving competence. *Assessment in Education: Principles, Policy & Practice*, 10, 329–345. doi:10.1080/0969594032000148172.
- Wittmann, W., & Hattrup, K. (2004). The relationship between performance in dynamic systems and intelligence. *Systems Research and Behavioral Science*, 21, 393–409. doi:10.1002/sres.653.
- Yasuda, T., Lawrenz, C., Whitlock, R. V., Lubin, B., & Lei, P. -W. (2004). Assessment of intraindividual variability in positive and negative affect using latent state-trait model analyses. *Educational and Psychological Measurement*, 64, 514–530. doi:10.1177/0013164403258445.
- Ziegler, M., Ehrlenspiel, F., & Brand, R. (2009). Latent state-trait theory: An application in sport psychology. *Psychology of Sport and Exercise*, 10, 344–349. doi:10.1016/j.psychsport.2008.12.004.

Erklärung

Erklärung gemäß § 8 Abs. 1 Buchst. b) der Promotionsordnung der Universität Heidelberg für die Fakultät für Verhaltens- und Empirische Kulturwissenschaften

Ich erkläre, dass ich die vorgelegte Dissertation selbstständig angefertigt, nur die angegebenen Hilfsmittel benutzt und die Zitate gekennzeichnet habe.

Erklärung gemäß § 8 Abs. 1 Buchst. c) der Promotionsordnung der Universität Heidelberg für die Fakultät für Verhaltens- und Empirische Kulturwissenschaften

Ich erkläre, dass ich die vorgelegte Dissertation in dieser oder einer anderen Form nicht anderweitig als Prüfungsarbeit verwendet oder einer anderen Fakultät als Dissertation vorgelegt habe.

Heidelberg, 12.Oktober 2011

Daniel Danner