Original Article

Measuring Individual Differences in Implicit Learning with Artificial Grammar Learning Tasks

Conceptual and Methodological Conundrums

Daniel Danner,[1] Dirk Hagemann,[2] and Joachim Funke[2]

[1]GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany
[2]Institute of Psychology, Heidelberg University, Heidelberg, Germany

Abstract: Implicit learning can be defined as learning without intention or awareness. We discuss conceptually and investigate empirically how individual differences in implicit learning can be measured with artificial grammar learning (AGL) tasks. We address whether participants should be instructed to rate the grammaticality or the novelty of letter strings and look at the impact of a knowledge test on measurement quality. We discuss these issues from a conceptual perspective and report three experiments which suggest that (1) the reliability of AGL is moderate and too low for individual assessments, (2) a knowledge test decreases task consistency and increases the correlation with reportable grammar knowledge, and (3) performance in AGL tasks is independent from general intelligence and educational attainment.

Keywords: implicit learning, artificial grammar learning task, reliability, validity

Implicit learning has been defined as learning without intention or as acquiring complex information without awareness of what has been learned (e.g., Mackintosh, 1998; Reber, 1992). One example of implicit learning is the learning of grammatical rules: in many situations we know whether a sentence is grammatically right or wrong but we cannot report the underlying grammatical rule. Several authors suggest that implicit learning is a fundamental aspect of learning in real life. For example, Gomez and Gerken (1999) suggest that implicit learning is crucial for learning languages, Funke and Frensch (2007) suggest that implicit learning is a determinant of success in solving complex real life problems, and Mackintosh (1998) suggests that implicit learning may even be a predictor of educational attainment. In line with that, there are empirical findings which suggest that implicit learning is a meaningful individual difference variable. First, several studies reported low associations between implicit learning and general intelligence (e.g., Gebauer & Mackintosh, 2007; Reber, Walkenfeld, & Hernstadt, 1991). These results suggest a good divergent validity of implicit learning. In addition, Kaufman et al. (2010) and Pretz, Totz, and Kaufman (2010) reported a significant relation between implicit learning and academic performance which also suggests the predictive validity of implicit learning. However, investigations on how individual differences in implicit learning can be measured have been sparse. This is the aim of the present study.

The Artificial Grammar Learning Task

During the last 45 years, the artificial grammar learning (AGL) task has become a standard paradigm of implicit learning (e.g., Reber, 1967). During a *learning phase*, the participants are asked to memorize a set of artificial letter strings (like WNSNXT). After that learning phase, the participants are informed that these letter strings were constructed according to a specific grammatical rule. In the subsequent *testing phase*, the participants are asked to judge new strings as grammatical or nongrammatical. One half of these strings are constructed according to the grammar and the other half are not. 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 are not able to report the grammar rules, which suggests that they learned the rules implicitly. The logic of such a task is

intuitively plausible. However, to serve as a meaningful individual difference variable, three criteria have to be met:

(1) The performance variable has to be reliable. The reliability is important because only a variable that can be measured reliably allows making inferences about individuals' ability. In particular, a low reliability results in a large confidence interval of an individual score whereas a high reliability allows an accurate estimate of an individual's ability. In addition, the reliability of a variable is important for evaluating the validity of a variable. A low reliability limits correlations with other variables and hence, correlations between variables with low reliability cannot be interpreted properly.
(2) The performance variable has to be task consistent. Task consistency means that several AGL tasks measure the same construct. In a research context, the task consistency may be important for investigating whether implicit learning is a trait-like ability that is stable over time. In an applied context, the task consistency may be important for an individual assessment (e.g., if a participant is tested more than one time).
(3) The performance variable has to be independent from reportable grammar knowledge. If AGL tasks measure implicit learning, there should be no correlation between the judgment accuracy and reportable grammar knowledge.

The usefulness of a performance measure may further be evaluated by its divergent validity and its predictive value. We will replicate the finding that implicit learning is independent from general intelligence and we will investigate its relation with educational attainment, but first, we will discuss previous findings and conceptual challenges.

Previous Findings

Reliability

There have only been sparse investigations of the reliability of implicit learning variables. Reber et al. (1991) examined $N = 20$ students and reported a Cronbach's alpha of $\alpha = .51$ for 100 grammaticality judgments. Likewise, Salthouse, McGuthry, and Hambrick (1999) assessed $N = 183$ participants between 18 and 87 years of age and reported a reliability of $\alpha = .40$ for an AGL task. These results demonstrate that it is possible to measure individual differences in implicit learning although this measurement is not very consistent. However, a limitation of these studies is that only a single grammar was used. So in sum, there is only weak support for the measurement of reliable individual differences in artificial grammar learning yet. Thus, we will systematically investigate the reliability of the performance in AGL tasks using Cronbach's alpha, the split-half correlation, and the retest correlation.

Task Consistency

Estimating the task consistency requires the same participants to complete at least two AGL tasks which causes a conceptual obstacle: during a first AGL task, the participants are asked to memorize a set of letter strings but they do not know that these strings are constructed according to a grammatical rule. Thus, they cannot learn the grammar intentionally. However, during a second AGL task, the participants already do know that there is a grammar and accordingly they may try to discover the grammar with intention. Having this in mind, can we be sure that a second or third AGL task still measures implicit learning? To avoid this problem, Gebauer and Mackintosh (2007) refined the standard paradigm. In the learning phase, they asked their participants to memorize a set of letter strings. In the subsequent testing phase, they asked their participants not to judge whether a letter string is grammatical but to judge whether a letter string was presented before ("old"). Even though, none of the strings were previously presented, they scored a grammatical letter string which was classified as "old" as a correct decision. The cunning idea behind this procedure was 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. From a conceptual point of view, novelty judgments and grammaticality judgments may be seen as similar. However, from an empirical point of view, it is unclear whether asking participants to rate the novelty of letter strings measures the same construct as asking participants to rate the grammaticality of letter strings.

Using novelty judgments, Gebauer and Mackintosh (2007) reported a correlation of $r = .15$ between two AGL tasks. This points toward a low task consistency. However, it is not known at present if this result indicates a low consistency of AGL in general or just in the case that the participants are asked to rate the novelty instead of the grammaticality of the strings. Thus, we will investigate the task consistency of AGL tasks in three experiments. A great correlation between two tasks suggests a good task consistency; a small correlation between two tasks suggests a poor task consistency.

Reportable Grammar Knowledge

Reber (1967) suggested that the participants in an AGL 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 performance in AGL tasks. Without doubt, this is a thorny question and, 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 and 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 (e.g., the bigram NX occurs more often in grammatical letter strings) was sufficient to explain the above chance accuracy of grammaticality judgments.

Jamieson and Mewhort (2009) used an episodic memory model to demonstrate that grammaticality judgments can also be explained by the similarity of letter strings with previously learned strings (see also Knowlton & Squire, 1996). Other authors suggested that participants make grammaticality judgments based on the chunks (e.g., Servan-Schreiber & Anderson, 1990), fluency (e.g., Kinder, Shanks, Cock, & Tunney, 2003), or fragment overlap (e.g., Boucher & Dienes, 2003). 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. A knowledge test meets the information criterion if it captures all kinds of relevant knowledge. It also meets the sensitivity criterion if it is as similar as possible in terms of retrieval context and instruction. Thus, to investigate the relation between implicit learning performance and reportable knowledge, we developed a knowledge test that was designed to meet the information as well as the sensitivity criterion. In particular, we selected those bi- and trigrams (*n*-grams) that occurred in the learning phase and which also occurred in the testing more frequently in grammatical than in nongrammatical strings. These *n*-grams allow participants to identify grammatical strings based on *n*-gram knowledge. We also selected those *n*-grams that did not occur in the learning phase but that did occur in the testing phase more frequently in nongrammatical strings than grammatical ones. These *n*-grams allow participants to identify nongrammatical strings based on *n*-gram knowledge. We presented these bigrams to the participants and asked them to rate whether the *n*-gram occurred more often in grammatical or nongrammatical letter strings. The test meets the information criterion because it is a direct test of participants' *n*-gram knowledge and also an indirect test of other, correlated knowledge such as similarity, chunks, or fragment overlap. Hence, an above chance classification of *n*-grams indicates that an above change classification of letter strings can be explained by reportable knowledge such as *n*-grams, similarity, chunks, or fragment overlap. However, the *n*-gram test does not indicate which source of knowledge was used. The test further meets the sensitivity criterion because the presentation of the stimuli and the response format were identical with the testing phase. Thus, a great correlation between performance in the testing phase and performance in the knowledge test suggests that participants use reportable knowledge to make their judgments. A small correlation between performance in the testing phase and performance in the knowledge test suggests that participants do not use reportable knowledge.

Relation With General Intelligence and Educational Attainment

Implicit learning is often described as part of an unconscious, intuitive learning system that is independent from explicit, declarative learning (e.g., Mackintosh, 1998; Reber & Allen, 2000). In line with this, several studies report a weak association between AGL performance and general intelligence (e.g., Gebauer & Mackintosh, 2007; Reber et al., 1991). However, even if these findings seem appealing at first glance, they may be criticized. For example, some studies did not report reliability estimates and therefore, a low correlation may also be explained by a low reliability. Other studies modified the standard AGL task and it is unclear whether this finding may be generalized to the standard AGL task. Therefore, a further aim of the present study was to replicate the finding that AGL performance and general intelligence are only weakly related.

From a practical point of view, the most important characteristic of a measure may be its predictive value. Mackintosh (1998) hypothesizes that performance in AGL may be a predictor of educational attainment. However, there is no empirical evidence for this hypothesis yet. Therefore, we will close this gap and test whether performance in an AGL task can predict educational success.

The Present Study

The present study investigates if individual differences in implicit learning can be measured with AGL tasks. In Experiment 1, we will investigate the reliability, the task consistency, and the relation with reportable knowledge when the participants are asked to rate the *novelty* of letter strings. In Experiment 2, we will investigate the reliability, the task consistency, and the relation with reportable knowledge when the participants are asked to rate the *grammaticality* of letter strings. In Experiment 3, we will investigate whether a knowledge test affects the task consistency and the relation with reportable knowledge. In this latter experiment we will further investigate how AGL performance is associated with general intelligence and educational attainment.

Experiment 1

Estimating the task consistency requires that participants complete two AGL tasks. Because this may cause a validity problem, Gebauer and Mackintosh (2007) asked their participants to rate the novelty of letter strings. Even though this idea is theoretically sound, there is no empirical evidence for the similarity of grammaticality and novelty ratings. Therefore, the aim of this experiment was (1) investigating the reliability, the task consistency, and the relation with reportable knowledge and (2) investigating whether asking the participants to rate the novelty of letter strings measures the same construct as asking the participants to rate the grammaticality of letter strings. The participants completed three AGL tasks. In Task 1 and Task 2 the participants rated the novelty of letter strings; in Task 3 the participants rated the grammaticality of letter strings.

Method

Participants

The participants were N = 21 students from Heidelberg University who were recruited from the campus and were paid €5 for their participation. This sample size was chosen because it allows 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.

Stimulus Material

There were three grammars. The strings of Grammar 1 (Figure 1) and Grammar 2 (Figure 2) were the same as those used by Gebauer and Mackintosh (2007). The strings of Grammar 3 were constructed as shown in Figure 3. For each grammar, there were 30 grammatical strings in the learning phase and 40 grammatical and 40 nongrammatical strings in the testing phase (complete lists of the

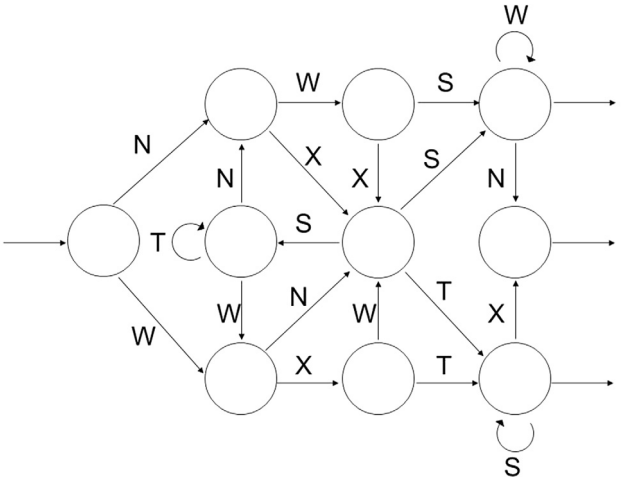

Figure 1. Grammar 1 (string construction identical to Gebauer & Mackintosh, 2007).

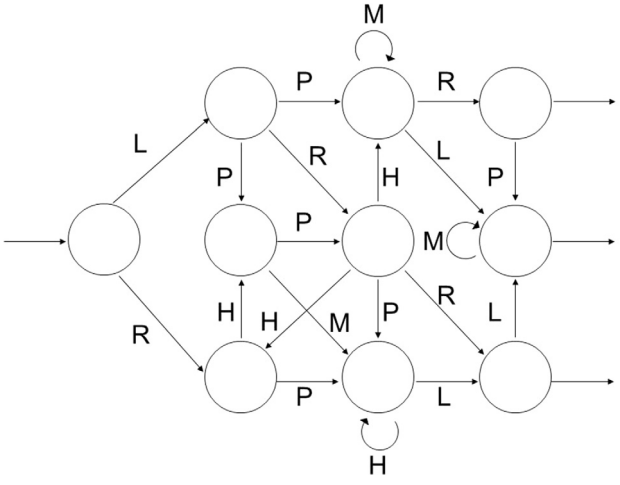

Figure 2. Grammar 2 (string construction identical to Gebauer & Mackintosh, 2007).

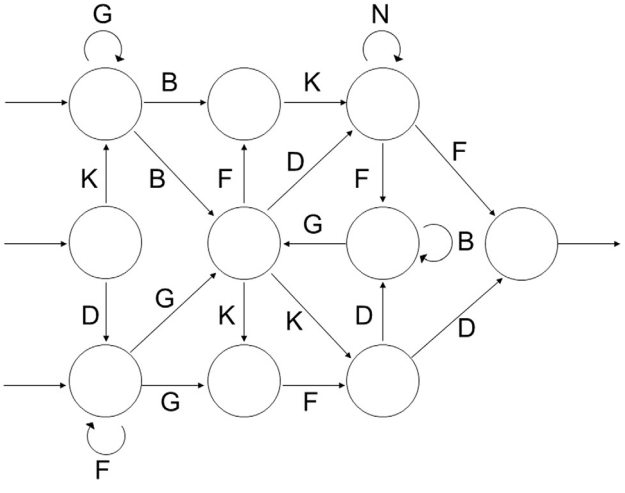

Figure 3. Grammar 3.

strings used in each phase are provided in the Electronic Supplementary Material, ESM 1). The nongrammatical strings contained one violation of the grammar at random positions of the strings. The length of the strings varied between three and eight letters.

To test the reportable 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 in nongrammatical strings (see ESM 1). These *n*-grams were chosen because 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 nongrammatical strings than in grammatical ones. Those strings were chosen because they may help to identify nongrammatical strings.

Procedure

Each participant completed three AGL 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. Each string was presented individually for 3 s. 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 10 strings correctly, these 10 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. In the testing phase, 80 new strings were presented (see ESM 1). Ten grammatical and 10 nongrammatical 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 (not 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 (e.g., NWS) 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.

After a short break, 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 AGL task.

After a short break, the *third artificial grammar learning task* was run with Grammar 3. 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, the participants were instructed to rate the strings as grammatical or nongrammatical. Otherwise, the procedure was identical with the first and the second task.

Measures

The *judgment accuracy* was quantified as the percentage of correct classifications of the strings in the testing phase. As suggested by Gebauer and Mackintosh (2007), grammatical strings which were rated as "old" strings and nongrammatical strings which were rated as "new" strings were counted as correct classifications. 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 *n*-grams which were rated as "old" and nongrammatical *n*-grams which were rated as "new" were counted as correct classifications.

Results

Judgment Accuracy

The judgment accuracy, the reliability estimates, and the correlation between tasks are shown in Table 1. In line with previous studies, the judgment accuracy was significantly above chance in all tasks (all $M \geq 57.29\%$, all $t \geq 5.95$, all $p < .001$). The reliability was estimated with Cronbach's alpha, the split-half correlation (odd-even-split, Spearman-Brown corrected), and the retest correlation. Because the retest correlation was based on only 20 out of the 100 presented strings, we de-attenuated the correlation coefficient by $\frac{5r}{1+4r}$ (Lord & Novick, 1974, p. 86). As can be seen, the reliability estimates were rather heterogeneous. Some estimates were even negative which clearly indicates that the assumptions of the underlying measurement model were violated. The greatest reliability estimate was the retest correlation of 0.87 in Task 2. There was no significant or

Table 1. Judgment accuracy in Experiment 1

	Task 1	Task 2	Task 3
Instruction	Novelty	Novelty	Grammaticality
Mean (%)	64.00	63.62	57.29
SD (%)	4.06	6.26	5.60
Cronbach's α	−.16	.46	.30
Split-half correlation[1]	−.59	.67	−.22
Retest correlation[2]	.52	.87	.27
Correlation with Task 1		−.18	.58**
Correlation with Task 2			−.08

Note. **$p < .01$. [1]Spearman-Brown corrected by $\frac{2r}{1+r}$; [2]corrected by $\frac{5r}{1+4r}$.

Table 2. *N*-gram knowledge in Experiment 1

	Task 1	Task 2	Task 3
Instruction	Novelty	Novelty	Grammaticality
Mean (%)	72.94	62.70	71.63
SD (%)	11.18	8.17	11.53
Cronbach's α	.45	−.29	.37
Split-half correlation[1]	.39	−.15	.47
Retest correlation	.50	.49	.64
Correlation with judgment accuracy	.27	−.10	.16

Note. No correlation with the judgment accuracy was significant. [1]Spearman-Brown corrected by $\frac{2r}{1+r}$.

substantial correlation between Task 1 and Task 2 ($r = -.18$, $p = .443$) or Task 2 and Tasks 3 ($r = -.08$, $p = .728$), but a significant correlation between Task 1 and Task 3 ($r = .58$, $p = .006$).

N-Gram Knowledge

Performance in the *n*-gram knowledge test and correlation with judgment accuracy are shown in Table 2. *N*-gram knowledge was significantly above chance in all tasks (all $M \geq 62.70\%$, all $t \geq 8.44$, all $p < .001$). There was no significant correlation between *n*-gram knowledge and judgment accuracy in Task 1, Task 2, or Task 3.

Discussion

This first experiment addressed the conceptual obstacle that arises when an AGL task is completed for a second time. After participants have completed a standard AGL task, they do know that there is a grammar constituting the letter strings and this may change their learning strategies in a second task. To avoid this problem, we followed Gebauer's and Mackintosh's (2007) approach and asked the participants to rate the *novelty* instead of the grammaticality. Then, we investigated whether novelty ratings indicate implicit learning success and whether novelty ratings bear an incremental value over grammaticality ratings.

The present results do suggest that novelty ratings indicate implicit learning success. First, the judgment accuracy in all tasks was significantly above chance in all three tasks. This replicates the results of Gebauer and Mackintosh (2007) and suggests that learning took place. Second, in Task 1 and Task 2 there are individual differences in implicit learning. The retest correlations were $r = .52$ and $r = .87$ which correspond with the reliability estimates reported by Gebauer and Mackintosh (2007) and Reber et al. (1991). Third, the correlations with the *n*-gram knowledge test were nonsignificant. This suggests that judgment accuracy cannot be explained by the participants' *n*-gram knowledge.

However, the correlation between judgment accuracy in Task 1 and Task 2 was small and nonsignificant ($r = -.18$). This suggests that novelty ratings are not task consistent: performance in the first AGL tasks indicates something different than performance in the second AGL task. When participants complete an AGL task for the first time, they are asked to learn a list of letter strings but they do not know that they will be asked to rate letter strings as new or old afterwards. When participants complete an AGL task for the second time, they are asked to learn a list of letter strings again, but they already know that they will be asked to rate letter strings as new or old afterwards. This may cause the participants to use different strategies or heuristics to remember the strings and the performance in the second task may reflect not only implicitly learned knowledge but also a change in cognitive processing. In sum, the results of the first experiment suggest that novelty ratings are moderately reliable and independent from *n*-gram knowledge but not task consistent. Therefore, novelty ratings seem to create no incremental value over grammaticality ratings.

Experiment 2

The results of the first experiment suggest that novelty ratings are not task consistent. In Experiment 2 we will investigate whether *grammaticality* ratings are a better performance indicator for implicit learning. The participants complete three AGL tasks. We estimate the reliability, the task consistency, and the association with *n*-gram knowledge when the participants are asked to rate the *grammaticality* of letter strings. In addition, we add the order of presentation of the grammars as a between-subject factor to make sure that a feature of a specific grammar does not bias the results.

Method

Participants

The participants were $N = 42$ students from the Heidelberg University who were recruited from the campus and were paid €5 for their participation. The order of presentation of the grammars was added as a between-participant

variable. One participant already had participated in Experiment 1 and therefore was excluded from the analysis.

Stimulus Material

The stimuli were the same as used in Experiment 1.

Procedure

All participants completed three AGL 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 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 as in Experiment 1 with two exceptions. First, the participants were asked to rate the grammaticality of the letter strings in all three tasks. Second, the *n*-grams were only presented once instead of twice, because of a software problem.

Measures

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

Results

The pattern of results was the same for order 1 and order 2. Therefore, we present the results pooled over both groups.

Judgment Accuracy

The judgment accuracy, the reliability estimates, and the correlation between tasks are shown in Table 3. Again, judgment accuracy was significantly above chance in all tasks (all $M \geq 59.15\%$, all $t \geq 8.88$, all $p < .001$). As in Experiment 1, the reliability estimates were moderate (between 0.49 and 0.80). There was no significant or substantial correlation between Task 1 and Task 2 ($r = .05$, $p = .715$) or between Task 1 and Task 3 ($r = .08$, $p = .631$), but a significant correlation between Task 2 and Task 3 ($r = .38$, $p = .014$).

N-Gram Knowledge

Performance in the *n*-gram knowledge test and the correlation with the judgment accuracy are shown in Table 4. *N*-gram knowledge was significantly above chance in all tasks (all $M \geq 64.90\%$, all $t \geq 7.83$, all $p < .001$). There was no significant correlation between *n*-gram knowledge and the judgment accuracy in Task 1 ($r = .06$, $p = .715$). However, there was a small correlation in Task 2 ($r = .30$, $p = .060$) and a significant correlation in Task 3 ($r = .34$, $p = .023$).

Discussion

In Experiment 2 we investigated whether grammaticality ratings can be used to measure individual differences in implicit learning. As in Experiment 1, the judgment accuracy was above chance in all three tasks which suggests that learning took place. In Task 1, there was no association

Table 3. Judgment accuracy in Experiment 2

	Task 1	Task 2	Task 3
Instruction	Grammaticality	Grammaticality	Grammaticality
Mean (%)	61.22	61.76	59.15
SD (%)	7.12	7.00	6.59
Cronbach's α	.55	.54	.49
Split-Half correlation[1]	.55	.75	.62
Retest correlation[2]	.79	.80	.52
Correlation with Task 1		.05	.08
Correlation with Task 2			.38*

Note. *$p < .05$. [1]Spearman-Brown corrected by $\frac{2r}{1+r}$; [2]corrected by $\frac{5r}{1+4r}$.

Table 4. N-gram knowledge in Experiment 2

	Task 1	Task 2	Task 3
Instruction	Grammaticality	Grammaticality	Grammaticality
Mean (%)	66.12	67.21	64.90
SD (%)	8.94	11.90	11.53
Cronbach's α	−.42	.24	.20
Split-Half correlation[1]	−.25	.40	.17
Correlation with judgment accuracy	.06	.30	.34*

Note. *$p < .05$. [1]Spearman-Brown corrected by $\frac{2r}{1+r}$.

between judgment accuracy and *n*-gram knowledge. In Task 2, there was a small correlation, and in Task 3 there was a significant correlation between judgment accuracy and *n*-gram knowledge. This suggests that performance in Task 1 captures individual differences in implicit learning, but performance in Task 2 and Task 3 captures the performance in a different learning process. In line with this interpretation, there was no association between performance in Task 1 and Task 2 or between Task 1 and Task 3 but there was a significant correlation between Task 2 and Task 3. Comparing the results of Experiment 1 and Experiment 2 further reveals that there was a substantial and significant correlation between the first and the third task in Experiment 1 but not in Experiment 2. This suggests that novelty ratings and grammaticality ratings are not equivalent even though both indicators measure aspects of implicit learning.

At first glance, this pattern of results appears to demonstrate that grammaticality ratings may only be used once to measure individual differences in implicit learning. During a second task, the participants may direct their attention toward *n*-grams and judgment accuracy is not a valid indicator for implicit learning any more. However, isn't there an alternative explanation? The participants completed a knowledge test (containing *n*-grams of letter strings) after every AGL 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. In Experiment 3, we will follow up on this possible explanation and investigate whether a knowledge test affects the task consistency and the relation with reportable knowledge.

Experiment 3

In Experiment 3 we investigate whether an *n*-gram knowledge test affects the task consistency of AGL tasks. Half of the participants completed two AGL tasks and an *n*-gram knowledge test after each AGL task (*n*-gram group). Half of the participants completed an *n*-gram knowledge test after the second AGL task only (control group). In addition, we investigated the relation between AGL performance, general intelligence, and educational attainment.

Method

Participants

The participants were $N = 106$ students from the Heidelberg University who were recruited from the campus and were paid €5 for their participation. The participants were randomly assigned to either the *n*-gram group ($N = 53$) or the control group ($N = 53$).

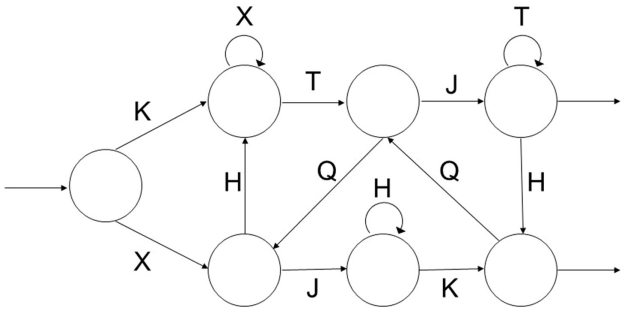

Figure 4. Grammar 4.

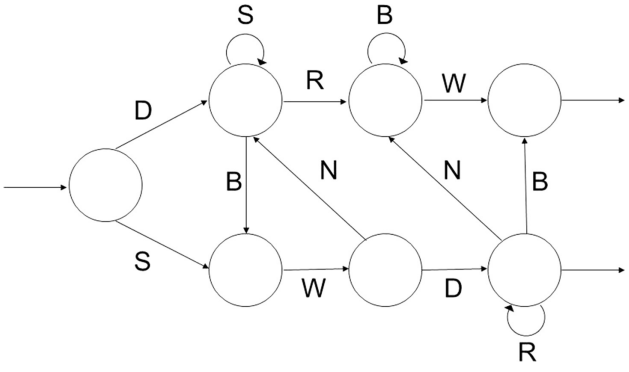

Figure 5. Grammar 5.

Stimulus Material

The stimuli for the first AGL task were constructed according to Grammar 4 (Figure 4). The stimuli for the second AGL task were constructed according to Grammar 5 (Figure 5). We created these two new grammars to ensure that our results are generalizable to different grammatical structures.

Procedure

All participants completed (1) a first AGL task, then (2) a knowledge test, then (3) the Culture Fair Intelligence Test (CFT; Cattell, Krug, & Barton, 1973), then (4) an additional AGL task, and then (5) a further knowledge test.

(1) *The first artificial grammar learning task*. The procedure of the AGL task was identical with Experiment 2, except that the learning phase consisted of 39 letter strings and that the testing phase consisted of 78 letter strings which were all repeated.

(2) *The first knowledge test*. Immediately after the testing phase, the participants completed a knowledge test. The *n*-gram group completed an *n*-gram knowledge test and the control group completed a dummy knowledge test. The *n*-gram knowledge test assessed participants' knowledge of *n*-grams. To judge an *n*-gram as grammatical, the participants had to press the A-key. To judge an *n*-gram

as nongrammatical, the participants had to press the L-key. There were 34 different *n*-grams for Grammar 4 (see ESM 1). All *n*-grams were presented twice so that there were a total of 68 items in the *n*-gram 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 *n*-gram knowledge test was taken as an indicator for the amount of reportable knowledge.

In order to make the procedure for the *n*-gram 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 Peru from 1990 to 2000" (which is correct, 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.

(3) The Culture Fair Intelligence Test (Cattell et al., 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 min. The number of correctly solved items was taken as the performance indicator for participants' general intelligence. In the present sample, mean IQ was $M = 128$ (range 91–150, $SD = 12.8$).

(4) *The second artificial grammar learning task*. The second AGL task also consisted of a learning phase and a testing phase. The procedure was identical to the first AGL task.

(5) *The second knowledge test*. The procedure for the second knowledge test was identical to the first with the exception that all participants completed an *n*-gram knowledge test after the testing phase and the *n*-gram knowledge test consisted of 36 items for Grammar 5. All *n*-grams were presented twice so that there was a total of 72 items in the *n*-gram knowledge test. The stimuli are shown in ESM 1.

Results

Judgment Accuracy

Judgment accuracy, reliability estimates, and correlation between tasks are shown in Table 5. In line with previous studies, judgment accuracy was significantly above chance in all tasks (all $M \geq 56.98\%$, all $t \geq 11.05$, all $p < .001$. Reliability estimates were generally moderate. In the control group, there was a significant correlation between Task 1 and Task 2 ($r = .39$, $p = .004$). In contrast, there was no significant correlation between Task 1 and Task 2 ($r = .22$, $p = .109$) in the *n*-gram group.

N-Gram Knowledge

Performance in the *n*-gram knowledge test and correlation with judgment accuracy are shown in Table 6. *N*-gram knowledge was significantly above chance in all tasks (all $M \geq 55.03\%$, all $t \geq 5.50$, all $p < .001$). In the *n*-gram group, there was no significant correlation between *n*-gram knowledge and judgment accuracy in Task 1 ($r = .01$, $p = .942$), but there was a significant correlation in Task 2 ($r = .30$, $p = .029$). In the control group, there was no knowledge test after the first AGL task. There was no significant correlation ($r = .02$, $p = .884$) between *n*-gram knowledge and judgment accuracy in Task 2.

Relation With General Intelligence and Educational Attainment

To investigate the relation between implicit learning, general intelligence, and educational attainment, we took the performance in the first AGL task as an indicator for implicit learning performance. The data of the *n*-gram group and the control group were analyzed together because the procedure for both groups was identical until the completion of the first AGL task. The number of solved items in the Culture Fair Intelligence Test served as a measure of participants' general intelligence. Cronbach's alpha for the 48 items was $\alpha = .73$. The correlation between performance in the first AGL task and the CFT3 was low and not significant ($r = .16$, $p = .111$). We also computed the correlation corrected for attenuation (r^*) across both groups. The retest

Table 5. Judgment accuracy in Experiment 3

	With knowledge test		Without knowledge test	
	Task 1	Task 2	Task 1	Task 2
Instruction	Grammaticality	Grammaticality	Grammaticality	Grammaticality
Mean (%)	58.09	56.98	59.62	57.28
SD (%)	7.96	4.61	6.55	4.32
Cronbach's α	.77	.33	.66	.21
Split-Half correlation[1]	.71	−.13	.49	.12
Retest correlation[2]	.75	.52	.68	.43
Correlation with Task 1		.22		.39*

Note. *$p < .05$. [1]Spearman-Brown corrected by $\frac{2r}{1+r}$; [2]Spearman-Brown corrected by $\frac{5r}{1+4r}$.

Table 6. *N*-gram knowledge in Experiment 3

	With knowledge test		Without knowledge test	
	Task 1	Task 2	Task 1	Task 2
Instruction	Grammaticality	Grammaticality	Grammaticality	Grammaticality
Mean (%)	55.45	55.03	–	56.40
SD (%)	7.92	6.66	–	6.87
Cronbach's α	.45	.33	–	.35
Split-Half correlation[1]	.49	−.01	–	.00
Retest correlation	.55	.32	–	.33
Correlation with judgment accuracy	.01	.30*	–	.02

Note. *$p < .05$. [1]Spearman-Brown corrected by $\frac{2r}{1+r}$.

correlation of Task 1 for both groups was $r = .72$ and hence, the correlation corrected for attenuation was $r^* = .22$.

We asked the participants to report their final school exams' grade point average (1 = "*very good*" to 6 = "*failed*"). The grades ranged between 1.0 and 3.1 with a mean of $M = 1.81$. The correlation between school grades and performance in the Culture Fair Intelligence Test was significant ($r = -.35$, $p < .001$, $r^* = -.40$) but the correlation between school grades and performance in the first AGL task was not significant ($r = -.10$, $p = .320$, $r^* = -.12$).

In addition, we also ran a multiple regression analysis and predicted school grade by the Culture Fair Intelligence Test and the performance in the first AGL task. There was a significant association between school grades and the Culture Fair Intelligence Test ($\beta = -.34$, $p < .001$) but not between school grades and the AGL task ($\beta = -.05$, $p = .600$).

Discussion

In Experiment 3, we investigated the hypothesis that AGL tasks are consistent if there is no *n*-gram test between subsequent tasks but not consistent if there is an *n*-gram test between tasks. The present results support this hypothesis. There was a significant correlation between two successive AGL tasks in the control group (no *n*-gram test after the first task) but not in the *n*-gram group (*n*-gram test after the first task).

The present results further suggest that the decrease in task consistency was due to an attention shift toward *n*-grams. In the *n*-gram group, there was no correlation between judgment accuracy and *n*-gram knowledge in Task 1, but there was a significant correlation between judgment accuracy and *n*-gram knowledge in Task 2. This suggests that the participants started to base their grammaticality judgments on *n*-grams after completing an *n*-gram test. In contrast, in the control group, there was no correlation between judgment accuracy and *n*-gram knowledge in Task 2. This suggests that the *n*-gram test and not the awareness that there is a grammar constituting the letter strings decreases the task consistency.

Buchner and Wippich (2000) discuss that the typical higher reliability of explicit measures compared to implicit measures makes it more likely to observe significant correlation between two explicit measures than between an implicit and an explicit measure. This suggests that the low and nonsignificant correlation between implicit learning and school grade could be explained by the lower reliability of the implicit learning measure. However, the reliability estimates for the first implicit learning task (α = .73 across both conditions) and the CFT (α = .73) were identical and thus, the reliabilities of the measures cannot have biased the correlation.

General Discussion

Implicit learning has stimulated research in various fields of psychology. In cognitive psychology implicit learning tasks have spawned fertile discussions about cognitive strategies and conscious and unconscious processes (e.g., Pothos, 2007). In psychophysiological research AGL tasks have been used to investigate physiological foundations of implicit learning (e.g., Schankin, Hagemann, Danner, & Hager, 2011). Recently, implicit learning has also attracted attention as an individual difference variable (e.g., Danner, Hagemann, Schankin, Hager, & Funke, 2011). The present work addressed conceptual and methodological conundrums with measuring individual differences in implicit learning. For one thing, we discussed and investigated the obstacles that arise when participants complete an AGL task more than once. For another thing, we addressed how reportable knowledge can be measured and how a knowledge test impacts the psychometric properties of an implicit learning performance variable. In the coming section, we will summarize the core findings and discuss how they are linked with theoretical and practical implications. The three core findings of our research are: (1) overall, the reliability of

performance indicators is moderate, (2) an *n*-gram knowledge test decreases the task consistency and increases the correlation with reportable grammar knowledge, and (3) performance in AGL tasks is independent from general intelligence and educational attainment.

Moderate Reliability of Performance Indicators

A glance over Tables 1–6 shows that both Cronbach's alpha and the split-half correlations have repeatedly negative values (e.g., Tables 1, 2, 4–6). Because reliability coefficients cannot be negative by definition, this points to a violation of assumptions that underlay the interpretation of these statistics as estimates of reliability. In particular, these statistics can be interpreted as point estimates of reliability only if all items (in the case of Cronbach's alpha) or both test-halves (in the case of split-half correlations) measure exactly the same true score and if the measurement errors are uncorrelated (and in the case of the split-half correlations, if the error variances of both test-halves are equal). The negative values of these statistics imply that at least one assumption is violated and therefore these statistics cannot be interpreted as estimates of reliability. On the other hand, all retest correlations were positive and therefore in the admissible range. Therefore, there is no direct indication that the assumptions underlying this statistics are violated and therefore we can interpret the retest correlations as coefficients of reliability.

As an estimate of reliability, the retest correlation requires the average of the repeated items to have the same true score and the same error variance (Lord & Novick, 1974). In particular, the method requires that the true score of the average judgment accuracy during the first presentation is the same true score as the average judgment accuracy during the second presentation. On the one hand, participants may feel more familiar with the strings and thus are more likely to rate them as grammatical. This would artificially increase the proportion of grammaticality ratings during the second presentation and thus decrease the retest correlation. On the other hand, participants may also explicitly remember the string as well as their prior rating which would artificially increase the correlation. However, as shown by Reber and Allen (1978) and our own pretests, the participants are not able to remember specific letter strings or their responses to specific letter strings.[1] Hence, the retest correlation may provide the most accurate reliability estimate in the present study and repeating items in an implicit learning task may be the most promising method to obtain an accurate reliability estimate.

In the present experiments, the average retest correlation was $r = .58$. The magnitude of this estimate is in line with previous research. Reber et al. (1991) reported a Cronbach's alpha of $\alpha = .50$ and Gebauer and Mackintosh (2007) reported split-half correlations of $r = .70$. These findings suggest that the manifest performance score is too inaccurate to make inferences regarding individuals' abilities and that AGL tasks should not be used for individual assessments. Buchner and Wippich (2000) suggest that participants use various cognitive processes when they complete implicit learning tasks and that this may be one reason for their typical low reliability.

Beyond the generally moderate reliability estimates, the results show a further conspicuity: there were substantial differences in the reliability estimates between tasks. For example, in Experiment 1, Task 2, the retest correlation was $r = .87$ whereas in Experiment 3, Task 2, the retest correlation was only $r = .21$. How can these differences be explained?

First, the reliability estimates of the judgment accuracies were strongly associated with the variance of judgment accuracies. We computed Spearman's ρ between the *SD*s of learning scores and the respective retest correlations across all tasks and experiments. This correlation was $\rho = .64$ ($p = .045$), that is, larger *SD*s of learning scores are positively related to larger reliabilities of learning scores. Some grammars are associated with larger individual differences in performance than others, and the former ones are particularly well suited for a reliable measurement of individual differences in AGL task performance.

Second, the properties of the specific letter strings can affect the reliability estimates because different letter strings may indicate implicit learning to a different extent which in turn can decrease the (true score) variance in implicit learning for a specific set of letter strings. One way of increasing the reliability may be selecting the items with the highest item-total correlation. On the one hand, this will yield a homogeneous set of letter strings and greater reliability estimates. On the other hand, the selected items may not be representative of the underlying grammar any more. For example, selecting items with the highest item-total correlation may produce a set of grammatical and nongrammatical strings that do not only differ in grammaticality but also in superficial features such as string length or fluency. Accordingly, the judgment accuracy in such a set of items may no longer indicate implicit learning but rather the use of fluency, string complexity, or string length as a heuristic. In addition, an increasing number of items may decrease participants' concentration

[1] One limitation of this argument must be mentioned. Reber and Allen (1978) used introspective reports to examine whether the participants remember specific strings, but even if the participants were not able to recall specific strings, they may still feel more familiar with some and hence, rated them as grammatical (cf. Whittlesea & Leboe, 2000).

or motivation. Thus, increasing the number of items may not be the best way to increase measurement accuracy. Another way may be developing letter strings that are less susceptible to fragment knowledge, fluency, or similarity. However, since we do not know which specific strings are affected by these effects, this will be a rocky road to greater reliability.

Third, the heterogeneity of participants can influence the reliability of a variable. Reliability is defined as the true score variance of a variable relative to the observed variance. Thus, in a homogeneous sample with only minor true score differences, the reliability of a variable may be small even if the test or instrument itself allows an accurate measurement with small error variance. A small variation of implicit learning true scores is also consistent with Reber's evolutionary model of implicit learning (e.g., Reber & Allen, 2000). The model describes implicit learning as a mechanism that developed long before explicit learning. Such an unconscious learning system that, for example, allows detecting the association between climate and occurrence of particular food sources may have been crucial for success or survival. Accordingly, individuals with high implicit learning abilities would have a higher probability to survive whereas individuals with low implicit learning abilities would have a lower probability to survive (principle of success). Over a long period of time, only successful implicit learners would survive which would result in smaller individual differences in implicit learning (principle of conservation). Hence, the moderate reliability estimates in AGL may generally reflect small individual differences in implicit learning.

Effects of *N*-Gram Knowledge Test

In Experiment 1 and Experiment 2, there were low and nonsignificant correlations between the first and the second AGL task when the participants completed an *n*-gram knowledge test between tasks ($-.18 \leq r \leq .05$). Likewise, in Experiment 3, there was a low and nonsignificant correlation when the participants completed an *n*-gram knowledge test between tasks ($r = .22$). This suggests a low task consistency when the participants complete knowledge tests between subsequent AGL tasks. On the other hand, there was a substantial and significant correlation ($r = .39$) between subsequent tasks when the participants did not complete an *n*-gram knowledge test between tasks. Adjusting this correlation for unreliability even reveals a correlation of $r = \frac{0.39}{\sqrt{0.68 \times 0.43}} = .72$ between the true scores. This suggests that a knowledge test decreases the task consistency of AGL tasks – not the participants' awareness that there is a grammar constituting the letter strings.[2]

In other words, measuring *n*-gram knowledge appears to generate a Heisenberg effect: by measuring the phenomenon, we change the phenomenon. The findings suggest that the participants start to shift their attention toward *n*-grams after completing a knowledge test and the participants may start to pay attention to which *n*-grams occur in subsequent learning phases and base their grammaticality judgments on their *n*-gram knowledge. For example, after completing an *n*-gram knowledge test, a participant may pay more attention to the *n*-grams in a subsequent learning phase. Hence, the participants may notice that the *n*-grams WNS and NXT occur more frequently than other *n*-grams. Thus, in a subsequent testing phase, the participant will judge letter strings containing these *n*-gram as grammatical. In line with this interpretation, the correlation between *n*-gram knowledge and judgment accuracy rose across tasks in Experiment 2 ($r = .06$, $r = .30$, $r = .34$). Likewise, in Experiment 3, there was a significant correlation between *n*-gram knowledge and judgment accuracy after the participants completed an *n*-gram knowledge test ($r = .30$), but not when the participants did not complete an *n*-gram test before ($r \leq .02$). This interpretation is in line with Perruchet and Pacteau (1990) who demonstrated that participants can use *n*-gram knowledge to reach above chance accuracy. Therefore, we suggest avoiding *n*-gram knowledge tests if the same participants should complete another AGL tasks in the future.

Artificial Grammar Learning Not Related to General Intelligence or Educational Attainment

In Experiment 3, performance in AGL was independent from general intelligence. This replicates previous research and suggests the divergent validity of AGL. Individual differences in implicit learning do not overlap with general intelligence and can reveal insights into cognitive ability beyond IQ. Thus, implicit learning may be seen as a complementary construct to describe human ability.

The correlation between AGL performance and the participants' school grades was also low and nonsignificant. This does not suggest that implicit learning is not relevant for educational success. However, there have not been many investigations of the predictive value of implicit learning yet and the students' grade point average may only be seen as a rough indicator of success in students' real lives. Therefore, future research will help to understand the role of implicit learning in real life in greater detail.

Limitation

Before strong conclusions can be drawn, one limitation of the present work must be noted. In student samples,

[2] The adjusted correlation is based on the reliability estimates of the tasks. As discussed, these reliability estimates can be biased and thus, the adjusted correlation may overestimate the correlation between true scores.

cognitive performance variables may be biased toward the upper range and may be restricted in their variance. This problem can readily be demonstrated in Experiment 3 where we used the CFT to measure intelligence. According to the norm tables (that are based on samples from 1963 to 1970), the participants had an IQ range between 91 and 150 with an above-average IQ of $M = 128$ (instead of the expected population $M = 100$) and a reduced variability of $SD = 12.8$ (instead of the expected population $SD = 15$). When interpreting these data, one must keep in mind that according to the "Flynn effect" the performance in IQ tests increases over the years, which may in part explain the above-average IQ scores in the present sample (see Pietschnig & Voracek, 2015, for a recent meta-analysis). Nonetheless, our data point to a variance reduction of nearly 30% in IQ scores, which may mitigate the correlations between IQ and other variables in the present study such as AGL task performance and school grades. In particular, Roth et al. (2015) performed a meta-analysis of the association between IQ scores and school grades. Based on 240 independent samples, they corrected for sampling error, unreliability, and restrictions of range and estimated the population correlation to be $r = .54$. The homogeneity of our samples with regard to cognitive performance may be one important reason why the observed correlation between IQ scores and school grades was only $r = .35$. Taken together, the use of student samples may reduce the variance of cognitive performance measures (such as AGL task performance, IQ scores, and school grades), which in turn mitigates reliabilities and correlations of these measures.

Alternative Setups of the AGL Task

One straightforward solution to the problem of low reliabilities of AGL scores noted above might be the use of different grammars in a sequence of different AGL tasks and combine the learning score. From a classical test theory perspective it is a sound suggestion to measure a construct with a variety of independent tasks and aggregate across them to obtain a total test score of great reliability and generality (validity). This idea is exemplified in test batteries of general intelligence such as the Wechsler tests. With respect to AGL task performance, one might perform subsequent AGL tasks with different grammars. This immediately poses the problem of repeated AGL tasks, that is, after the first task the participants know the grammatical nature of the letter strings and therefore may change their learning strategy in subsequent tasks.

There may be one interesting option to circumvent the problem of repeated testing with the AGL task. In the present experiments we used a sequential setup of the tasks, that is, we conducted one AGL task with one grammar at one time, then we conducted the next AGL task with another grammar and so forth, each task having its own learning and testing phase. However, it would also be possible to use these different grammars to produce letter strings and mixing these strings from different grammars in one learning phase. This approach would effectively prevent the problem that knowledge of the grammatical nature of the letter strings may influence the next learning phase. Unfortunately, this approach has some limitations of its own. Using two or more different grammars to generate one set of letter strings is indistinguishable from constructing one hybrid grammar and using this to generate the strings. Such a hybrid grammar would start with a common starting node and then switch to one of the basic grammars. For example, consider Figures 1 and 2 which show two basic grammars. They can easily be combined by using the left-handed starting node of each grammar as a common start and allowing four paths N, W, L, and R, with the former two continuing with grammar 1 and the latter two continuing with grammar 2. Of course, this hybrid grammar would be much more complex than each of the basic grammars. Schiff and Katan (2014) have investigated the associations between grammar complexity and performance in the AGL tasks. They meta-analyzed data from 56 experiments that used 10 different grammars and showed that there is a negative correlation of $r = -.32$ between grammar complexity (quantified as the topological entropy of the grammar chart) and AGL performance across all experiments. From this result it may be inferred that following the idea of intermixing letter strings of several separate grammars will deteriorate performance, that is, the task difficulty will increase. A shift of task difficulty toward greater difficulty must shift each person's judgment accuracy toward chance level, which in turn must reduce the variance of judgment accuracy. In turn, a reduction of the variance of AGL task performance will mitigate reliability of the performance measures and their correlations with other variables (such as intelligence and school grades). Therefore, mixing up several items from different grammars in one learning phase may solve the problem of task knowledge in a sequential setup of the AGL task but will reduce reliability and correlations. If this route is a viable one may be target of future research.

Summary

Artificial grammar learning tasks can be used to measure individual differences in implicit learning. The low correlation with other ability constructs such as general intelligence suggests a good divergent validity of AGL. In line with previous research, the present results suggest that the reliability of the measurement is generally moderate and hence may be used for the study of individual differences but not for individual assessments. The present results further suggest

that *n*-gram knowledge tests should be avoided when the participants complete more than one AGL task because an *n*-gram test shifts attentions toward *n*-grams, decreases the task consistency, and increases the relation with reportable grammar knowledge. We hope that the present results and reflections stimulate and support this line of research in the field of implicit learning.

Acknowledgments

This research was supported by grants awarded by the Deutsche Forschungsgemeinschaft to Dirk Hagemann (HA3044/7-1). Part of this research was previously published as part of the Daniel Danner's dissertation at the Heidelberg University. We gratefully thank Andreas Neubauer, Anna-Lena Schubert, and Katharina Weskamp for administering the experiments and Saul Goodman and an anonymous reviewer for helpful comments on an earlier draft of this manuscript.

Electronic Supplementary Material

The electronic supplementary material is available with the online version of the article at http://dx.doi.org/10.1027/2151-2604/a000280

ESM 1. Tables (.doc).
Strings used in the experiments.

References

Boucher, L., & Dienes, Z. (2003). Two ways of learning associations. *Cognitive Science, 27*, 807–842. doi: 10.1207/s15516709cog2706_1

Buchner, A., & Wippich, W. (2000). On the reliability of implicit and explicit memory measures. *Cognitive Psychology, 40*, 227–259. doi: 10.1006/cogp.1999.0731

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.

Danner, D., Hagemann, D., Schankin, A., Hager, M., & Funke, J. (2011). Beyond IQ: A latent state-trait analysis of general intelligence, dynamic decision making, and implicit learning. *Intelligence, 39*, 323–334. doi: 10.1016/j.intell.2011.06.004

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

Funke, J., & Frensch, P. A. (2007). Complex problem solving: The European perspective – 10 years after. In D. H. Jonassen (Ed.), *Learning to solve complex scientific problems* (pp. 25–47). New York, NY: Erlbaum.

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

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70*, 109–135. doi: 10.1016/S0010-0277(99)00003-7

Jamieson, R. K., & Mewhort, D. J. K. (2009). Applying an exemplar model to the artificial-grammar task: Inferring grammaticality from similarity. *The Quarterly Journal of Experimental Psychology, 62*, 550–575. doi: 10.1080/17470210802055749

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. (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. doi: 10.1037/0278-7393.22.1.169

Lord, F. M., & Novick, M. R. (1974). *Statistical theories of mental test scores*. Oxford, UK: Addison-Wesley.

Mackintosh, N. J. (1998). *IQ and human intelligence*. New York, NY: 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. doi: 10.1037/0096-3445.119.3.264

Pietschnig, J., & Voracek, M. (2015). One century of global IQ gains: A formal meta-analysis of the Flynn effect (1909–2013). *Perspectives on Psychological Science, 10*, 282–306. doi: 10.1177/1745691615577701

Pothos, E. M. (2007). Theories of artificial grammar learning. *Psychological Bulletin, 133*, 227–244. doi: 10.1037/0033-2909.133.2.227

Pretz, J. E., Totz, K. S., & Kaufman, S. B. (2010). The effects of mood, cognitive style, and cognitive ability on implicit learning. *Learning and Individual Differences, 20*, 215–219. 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, 1*, 93–133. doi: 10.1016/1053-8100(92)90051-B

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* (Vol. 20, pp. 227–247). Amsterdam, The Netherlands: John Benjamins.

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

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Dominik, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence, 53*, 118–137. doi: 10.1016/j.intell.2015.09.002

Salthouse, T. A., McGuthry, K. E., & Hambrick, D. Z. (1999). A framework for analyzing and interpreting differential aging

patterns: Application to three measures of implicit learning. *Aging, Neuropsychology, and Cognition, 6*(1), 1–18.
Schankin, A., Hagemann, D., Danner, D., & Hager, M. (2011). Violations of implicit rules elicit an early negativity in the ERP. *NeuroReport, 13*, 642–645. doi: 10.1097/WNR.0b013e328349d146
Schiff, R., & Katan, P. (2014). Does complexity matter? Meta-analysis of learner performance in artificial grammar tasks. *Frontiers in Psychology, 5*, 1084. doi: 10.3389/fpsyg.2014.01084
Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 592–608.
Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences, 17*, 367–447. doi: 10.1017/S0140525X00035032
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

Received October 30, 2016
Revision received November 29, 2016
Accepted December 13, 2016
Published online July 12, 2017

Daniel Danner
GESIS – Leibniz Institute for the Social Sciences
PO Box 122155
68072 Mannheim
Germany
daniel.danner@gesis.org