DOCTORAL (PhD) DISSERTATION

EMESE SZEGEDI-HALLGATÓ

METHODOLOGICAL AND THEORETICAL CONSIDERATIONS IN IMPLICIT LEARNING RESEARCH

2019

EÖTVÖS LÓRÁND UNIVERSITY

FACULTY OF EDUCATION AND PSYCHOLOGY

Emese Szegedi-Hallgató

Methodological and theoretical considerations in implicit learning research

Doctoral School of Psychology

Head of the School: Prof. Zsolt Demetrovics (PhD, DSc), professor, Eötvös Loránd University

Cognitive Psychology Program

Head of the Program: Prof. Ildikó Király (PhD), professor, Eötvös Loránd University

Supervisor

Prof. Dezső Németh (PhD, DSc), professor, Eötvös Loránd University

Budapest, 2019

EÖTVÖS LORÁND TUDOMÁNYEGYETEM

ADATLAP a doktori értekezés nyilvánosságra hozatalához

I. A doktori értekezés adatai

A szerző neve: **Szegedi-Hallgató Emese**

A doktori értekezés címe és alcíme: **Módszertani és elméleti megfontolások az implicit tanulás kutatásában**

Angol cím: Methodological and theoretical considerations in implicit learning research

A doktori iskola neve: **Pszichológia Doktori Iskola**

A doktori iskolán belüli doktori program neve: **Kognitív Pszichológia Program**

A témavezető neve és tudományos fokozata: **Prof. Németh Dezső (PhD, DSc), egyetemi tanár**

A témavezető munkahelye: **ELTE PPK Pszichológiai Intézet**

MTA Adatbázis-azonosító: **10032697**

DOI-azonosító[1]: **10.15476/ELTE.2019.251**

II. Nyilatkozatok

1. A doktori értekezés szerzőjeként[2]

a) hozzájárulok, hogy a doktori fokozat megszerzését követően a doktori értekezésem és a tézisek nyilvánosságra kerüljenek az ELTE Digitális Intézményi Tudástárban. Felhatalmazom a ELTE PPK Pszichológiai Doktori Iskola hivatalának ügyintézőjét, **Barna Ildikót**, hogy az értekezést és a téziseket feltöltse az ELTE Digitális Intézményi Tudástárba, és ennek során kitöltse a feltöltéshez szükséges nyilatkozatokat.

b) kérem, hogy a mellékelt kérelemben részletezett szabadalmi, illetőleg oltalmi bejelentés közzétételéig a doktori értekezést ne bocsássák nyilvánosságra az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban;[3]

---

[1] A kari hivatal ügyintézője tölti ki.
[2] A megfelelő szöveg aláhúzandó.

c) kérem, hogy a nemzetbiztonsági okból minősített adatot tartalmazó doktori értekezést a minősítés (*dátum*)-ig tartó időtartama alatt ne bocsássák nyilvánosságra az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban;[4]

d) kérem, hogy a mű kiadására vonatkozó mellékelt kiadó szerződésre tekintettel a doktori értekezést a könyv megjelenéséig ne bocsássák nyilvánosságra az Egyetemi Könyvtárban, és az ELTE Digitális Intézményi Tudástárban csak a könyv bibliográfiai adatait tegyék közzé. Ha a könyv a fokozatszerzést követőn egy évig nem jelenik meg, hozzájárulok, hogy a doktori értekezésem és a tézisek nyilvánosságra kerüljenek az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban.[5]

2. A doktori értekezés szerzőjeként kijelentem, hogy
a) a ELTE Digitális Intézményi Tudástárba feltöltendő doktori értekezés és a tézisek saját eredeti, önálló szellemi munkám és legjobb tudomásom szerint nem sértem vele senki szerzői jogait;
b) a doktori értekezés és a tézisek nyomtatott változatai és az elektronikus adathordozón benyújtott tartalmak (szöveg és ábrák) mindenben megegyeznek.

3. A doktori értekezés szerzőjeként hozzájárulok a doktori értekezés és a tézisek szövegének plágiumkereső adatbázisba helyezéséhez és plágiumellenőrző vizsgálatok lefuttatásához.

Kelt: **Szeged, 2019.10.07.**

a doktori értekezés szerzőjének
aláírása

[3] A doktori értekezés benyújtásával egyidejűleg be kell adni a tudományági doktori tanácshoz a szabadalmi, illetőleg oltalmi bejelentést tanúsító okiratot és a nyilvánosságra hozatal elhalasztása iránti kérelmet.
[4] A doktori értekezés benyújtásával egyidejűleg be kell nyújtani a minősített adatra vonatkozó közokiratot.
[5] A doktori értekezés benyújtásával egyidejűleg be kell nyújtani a mű kiadásáról szóló kiadói szerződést.

Acknowledgements

It has been a long journey that seems to come to an end now. When I stepped on this road, in 2012, I had no clue how much joy and tears, success and fears, ups and downs will follow. Frankly, I thought it was going to be easier. But I learned a lot. A lot about science and cognitive psychology in particular; about managing life-work balance, and last but not least, a lot about myself. I am grateful for all of these lessons.

I would first like to thank my supervisor, **PROF. DEZSŐ NÉMETH (PhD, DSc)**, for everything he taught me, for his support of me and for the opportunities he provided me in the past ten years. Without you, it couldn't have happened.

I would also like to thank my colleagues and coauthors, **ANNA BÁLINT, DÓRA GYŐRI-DANI, EMŐKE ADRIENN HOMPOTH, LEILA KEREPES, DR. ZSUZSANNA LONDE, JUDIT PEKÁR, TÍMEA SÁNDOR, LIA ANDREA TASI, TEODÓRA VÉKONY** and especially **DR. KAROLINA JANACSEK**. It has been a pleasure to work with you on these projects.

I am grateful to **DR. ÉVA SZABÓ** (and the management of the Faculty of Humanities at the University of Szeged) for their moral and financial support of my doctoral studies.

I would like to express my very great appreciation to **DR. ATTILA KRAJCSI** for everything he taught me during my student years at the University of Szeged. His way of thinking and doing research had always inspired me, and it still does.

Finally, I am very grateful to my **PARENTS** for teaching me that being a problem-solver is sexy; that girls can do math and programming; and that hard work pays off in the end. I am now aware that you were part of the reason why I started the whole thing. I wanted to make you proud. And I'm in deep pain because, when finally in the finish, I have to celebrate without you. I love you both.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Supplementary Tables

List of Supplementary Figures

List of abbreviations

AM	Ante Meridiem (before midday *in Latin*)
ANOVA	Analysis of Variance
ASRT	Alternating Serial Reaction Time
AUC	Area Under the Curve
AUROC	Area Under the Reciever Operating Characteristic Curve
CC	Contextual Cueing
CV	Coefficient of Variation
EPKEB	Egyesített Pszichológiai Kutatásetikai Bizottság (United Ethical Committee for Research in Psychology *in Hungarian*)
H	High frequency triplets
H1	First subset of the High frequency triplets
H1P	Pattern-ending H1 trials
H1R	Random-ending H1 trials
H2	Second subset of the High frequency triplets
H2P	Pattern ending H2 trials
HH	„High-High", High frequency triplet – High frequency triplet
HL	„High-Low", High frequency triplet – Low frequency triplet
HP	Pattern-ending High frequency triplet
HR	Random-ending High frequency triplet
KS	Kolmogorov-Smirnov test
L	Low frequency triplets
LH	„Low-High", Low frequency triplet – High frequency triplet
LL	„Low-Low", Low frequency triplet – Low frequency triplet
LR	Random-ending Low frequency triplet

MSE	Mean Squared Error
MW	Mann-Whitney test
NF	No Filter
P	Pattern
P1-P6	Pattern1 – Pattern6
PM	Post Meridiem (after midday *in Latin*)
QF	Quad Filter
R	Random
RSI	Response to Stimulus Interval
RT	Reaction Time
SD	Standard Deviation
SEM	Standard Error of Mean
SLE	Sequence Learning Effect
SF	Supplementary Figure
SM	Supplementary Methods
SR	Supplemetary Results
SRT	Serial Reaction Time
ST	Supplementary Table
TF	Triplet Filter
WP	Weather Prediction

List of publications that the dissertation is based upon

Study	Publication	Impact Factor
1	Nemeth, D., **Hallgató, E.**, Janacsek, K., Sándor, T., & Londe, Z. (2009). Perceptual and motor factors of implicit skill learning. NeuroReport, 20(18), 1654. https://doi.org/10.1097/WNR.0b013e328333ba08	**1.805**
2	**Hallgató, E**., Győri-Dani, D., Pekár, J., Janacsek, K., & Nemeth, D. (2013). The differential consolidation of perceptual and motor learning in skill acquisition. *Cortex*, *49*(4), 1073–1081. https://doi.org/10.1016/j.cortex.2012.01.002	**6.042**
3	**Szegedi-Hallgató, E.,** Janacsek, K., Vékony, T., Tasi, L. A., Kerepes, L., Hompoth, E. A., … Németh, D. (2017). Explicit instructions and consolidation promote rewiring of automatic behaviors in the human mind. Scientific Reports, 7(1), 1–7. https://doi.org/10.1038/s41598-017-04500-3	**4.122**
4	**Szegedi-Hallgató, E**., Janacsek, K., & Nemeth, D. (2019). Different levels of statistical learning — Hidden potentials of sequence learning tasks. *PloS One*, *14*(9), e0221966. https://doi.org/10.1371/journal.pone.0221966	**expected 2.776**

Each co-author has granted permission for the given publication to be included in the current dissertation.

I. GENERAL INTRODUCTION

Skill acquisition, habit formation, and development of behavioral automatisms are all results of learning processes, sharing a unique combination of features that makes them different from other kinds of learning. According to one point of view, these learning types are forms of non-declarative learning, underscoring thus that learning is not dependent on the mediotemporal brain structures (Squire & Zola, 1996). Another viewpoint emphasizes the fact that learners are usually not fully aware of the information that had been acquired, and it is only their improving performance that implies learning, thus emphasizing conscious awareness (or the lack of it) as a defining criterion. Learning that occurs without awareness is called implicit - in contrast with explicit learning where conscious awareness accompanies learning (A. S. Reber, 1967; Graf & Schacter, 1985). A third approach, by contrast, relies on three variables: the speed of encoding (rapid vs. slow); whether a single item is encoded or associations among multiple items; and the compositionality (vs. rigidity) of the resulting memory (Henke, 2010). According to this view, skill acquisition and habit formation is a form of slow encoding of rigid associations (as is classical conditioning and semantic memory). And finally, there is a separate research tradition, namely the investigation of statistical learning abilities originating from Saffran, Aslin, & Newport (1996) that also deals with the unsupervised, incidental learning of an inherent structure present in the to-be-learned material; Perruchet & Pacton (2006) went as far as suggesting that implicit learning and statistical learning is actually the same phenomenon (see also Christiansen, 2018). In a similar vein, Reber (2013) proposed that implicit memory manifests as an improvement from experience based on mechanisms of cortical plasticity; the extraction of the underlying statistical structure is incremental, and it allows for a distributed representation of information.

Despite the similarities between these research traditions, and the substantial overlap of their proposed constructs, their notions are not synonyms. For example, the term *implicit learning* is broader than the term *skill learning*, as other types of implicit learning phenomena also exist, e.g. priming, classical conditioning and habituation/sensitization (Squire & Zola, 1996). On the other hand, skill learning does not only rely on implicit processes but also on explicit learning (Ghilardi, Moisello, Silvestri, Ghez, & Krakauer, 2009; Taylor, Krakauer, & Ivry, 2014). Third, although statistical learning is thought to be an implicit learning process (e.g. Kim, Seitz, Feenstra, & Shams, 2009; Perruchet & Pacton, 2006; Turk-Browne, Scholl, Chun, & Johnson, 2008), there is evidence that explicit knowledge can also emerge after the encounter with statistically structured stimuli (e.g. Perruchet, Bigand, & Benoit-Gonin, 1997; Rünger & Frensch, 2008; Goujon, Didierjean, & Poulet, 2014). The narrow field I was interested in (which is summarized in this work) is the *implicit statistical learning*, not implicit learning or skill learning in general.

I/1. Tasks of implicit statistical learning

A typical test of implicit (statistical) learning is the Artificial Grammar Learning (AGL) Task (A. S. Reber, 1967; or more recently, Danner, Hagemann, & Funke, 2017), in which words of a non-existent, fictional language are created by an algorithm (based on conditional probabilities, e.g. the letter A is followed by the letter B or letter C, but never with the letter D; thus the words AB and AC are legal in that language, but AD is not). The algorithm is never explicitly uncovered, it can only be inferred from the shown examples, that, according to the instruction, need to be memorized. After the learning phase, new words are shown which either obey the rules of the algorithm or not; participants are asked to guess whether particular words are legal in the artificial language. The percentage of correct guesses informs us whether learning occurred or not.

Another type of task is the Weather Prediction (WP) Task (or more generally the Probabilistic Classification tasks) (e.g. Knowlton, Squire, & Gluck, 1994) which differ from the AGL in that instead of showing concrete examples resulting from the underlying statistical structure (algorithm) and then testing the knowledge via a forced-choice task, participants int he WP are asked to guess the „outcomes” (rainy or sunny weather) based on the shown cards from the beginning of the task, and learning is aided by the feedback that is provided about the correctness of the guesses. Again, performance is assessed by computing the percentage of correct guesses (and comparing it to the baseline of chance level).

A further typical task is the Sugar Factory (or more generally the Dynamic Systems Control tasks) (Berry & Broadbent, 1984) in which participants are required to learn to control a complex system, where the relationship between participants' settings and the outcomes is governed by a hidden algorithm. On every trial, a goal is defined that participants need to achieve by setting the input variables, and if they approximate the goal close enough, the trial is considered to be completed. Similarly to the previously described tasks, accuracy is the only measure of implicit learning, as there is no time limit for accomplishing the goals.

In the Contextual Cueing (CC) paradigm (Chun, 2000) complex spatial layouts are shown to participants and their task is to find a target among the distractors (and indicate its direction with one of the two possible keypresses). Some of the layouts are repeated, and participants are getting progressively more efficient in reacting to targets in these layouts despite not being able to recognize that they have completed these trials before. In other words, performance is mediated by global repetition statistics of the displays (Zang, Zinchenko, Jia, Assumpção, & Li, 2018). Participants' efficacy in responding to the targets is measured by assessing reaction times and/or accuracy.

Finally, in the Serial Reaction Time task (Nissen & Bullemer, 1987) - or more generally the Sequence Learning tasks – participants have to respond to the location of consecutive stimuli, which, unbeknownst to them, follows a deterministic or probabilistic sequence. With determinisitic sequences, learning is usually measured by inserting random or pseudo-random blocks of stimuli, and assessing the worsening of performance on these blocks (Nissen & Bullemer, 1987); with probabilistic sequences, on the other hand, performance is measured by contrasting performance on probable outcomes with performance on less probable outcomes (J. H. Howard & Howard, 1997). Similarly to the Contextual Cueing paradigm, the efficacy of responding can be measured via reaction times and/or accuracy measures.

I/2. The relationship between the different tasks measuring implicit statistical learning

The previously described tasks differ in many ways; e.g. whether the regularity is present temporally or spatially, whether the exposure of the regularity is passive or requires some activity from the participant, etc. Nevertheless, they all rely on the detection of statistical regularities which are covertly present in the task (Arciuli & Conway, 2018). It is thus somewhat surprising that learning scores gained from different tests do not correlate with each other (Gebauer & Mackintosh, 2007; Sævland

& Norman, 2016; Siegelman & Frost, 2015) or even if they do, the correlation is weak (Kalra, Gabrieli, & Finn, 2019). Conversely, dissociations within implicit memory tests were observed in dyslexics (Bennett, Romano, Howard, & Howard, 2008; J. H. Howard, Howard, Japikse, & Eden, 2006) and children with Attention Deficit Hyperactivity Disorder (Barnes, Howard, Howard, Kenealy, & Vaidya, 2010).

The lack of correlation (and the dissociations) between the different measures of implicit statistical learning is alarming, and it is important to find the reasons behind it. First, it is possible that there is truly no relationship between these measures and hence research is (rightfully) unable to find one. Theoretically, this scenario would question the domain-generality (opposed to domain-specificity) and/or the unitary nature (opposed to multicomponentiality) of implicit statistical learning. In other words, it would mean that there is no such thing as „the implicit statistical learning", only different types of it. Practically, it would highlight the need to find the factors that differentiate between different types of implicit statistical learning, and this knowledge – in turn – would be used for designing new tasks and/or help us to choose from the existing tasks to fulfill our purposes.

In a second scenario, there *is* a positive relationship between these different measures, but – for some reason – researchers have been unable to find it. The reason behind this could be methodological and/or related to the psychometric properties of the tasks. In spite of bearing the hope that we could somehow overcome these obstacles in the future, this scenario would also mean that our knowledge about the nature of implicit statistical learning is seriously biased (possibly wrong in many aspects). If our tests are so weak in terms of reliability, for example, that they barely correlate with each other, how could we interpret the lack of correlation with other kinds of tests?

In the following sections, I will briefly discuss the possible factors behind both scenarios (i.e. no relationship between tasks, or the difficulty of finding them). I will also indicate how we considered these factors in our research.

I/3. Implicit Statistical Learning – One or Many?

I/3.1. Modality Specificity

Accumulating evidence suggests that there are qualitative differences in patterns of implicit statistical learning in the auditory, visual and tactile modalities, which corroborate the notion of modality specificity of implicit statistical learning (Emberson, Conway, & Christiansen, 2011; Li, Zhao, Shi, Lu, & Conway, 2018; Walk & Conway, 2016). A putative explanation puts forward that encoding of information follows different constraints that are determined by the specific properties of the input in the respective brain cortices (despite similar sets of computational principles) (Conway & Christiansen, 2005). For example, the auditory cortex might be more sensitive to the temporal accumulation of information than the visual cortex (Frost, Armstrong, Siegelman, & Christiansen, 2015). In line with this, it was found that timing parameters affect the visual statistical learning more than auditory learning, and visual learning of temporally structured information is worse than the visual learning of spatially structured information or the auditory learning of temporally structured information (Conway & Christiansen, 2009).

I/3.2. Independency from other cognitive abilities

Arciuli (2017) reviewed evidence that statistical learning is sometimes found to be better in younger than in older participants, while sometimes the opposite pattern can be observed. As a resolution for the mixed findings, he suggested that implicit statistical learning is a multicomponent ability (being comprised of certain types of attention, processing speed, and memory, etc.); and performance on different tasks might depend on the way they draw on particular underlying components (Arciuli, 2017; Arciuli & Conway, 2018).

Although one might question whether attention or processing speed, for example, should be regarded as parts of implicit statistical learning, it is certainly true that the different tasks vary in terms of cognitive demands (comprising statistical learning and other abilities), which could result in very divergent results. Even with equivalent statistical learning abilities, there might be significant individual differences in performance on different tasks (e.g. it is necessary for one to be able to motorically respond quickly to an event in case of reaction time tasks, otherwise learning can not be detected even if it occurs).

Additionally, even if implicit and explicit processes dissociate, it does not exclude the possibility of interplay between these memory systems; and although the evidence is not unequivocal, some results do point towards this possibility (Boyd & Winstein, 2003; Arnaud Destrebecqz et al., 2005; Dew & Cabeza, 2011; Lagarde, Li, Thon, Magill, & Erbani, 2002; Sun, Zhang, Slusarz, & Mathews, 2007; but see Sanchez & Reber, 2013; and Curran & Keele, 1993). In a related field of research, assessing the performance of skilled behavior under stress, it was found that explicit processing (but not implicit learning of the same skill) hampered performance of that skill under stressful conditions (Masters, 1992; Maxwell, Masters, & Eves, 2000; Gucciardi & Dimmock, 2008). Thus, performance on implicit statistical learning tasks might also be mediated by explicit processes.

I/3.3. Type of statistics – Does it matter?

It has been recognized that humans are capable of learning at least two types of statistics: joint probaibilities (i.e. distributional statistics of chunks of information), and conditional probabilities (i.e. the predictability of a target event given its antecedents) (J. H. Howard, Howard, Dennis, & Kelly, 2008; Thiessen, Kronstein, & Hufnagle, 2013; Thiessen, 2017) and it has been suggested that those are results of independent processes (Thiessen, 2017). However, the relative contribution of different types of statistics in a specific learning task is rarely discussed (but see J. H. Howard et al., 2008).

Additionally, the complexity of the embedded statistical structure might also contribute to differences observed with different statistical learning tasks. For example, in sequential tasks, when the previous element predicts the next element, it is called a first-order sequential structure; when the N-2th trial has predictive power on the current target, the sequence has a second-order structure, and so on. It has been shown that humans are capable of learning up to fourth-order statistical regularities (Remillard, 2008, 2011), or even fifth- and sixth-order regularities (Remillard, 2010). At the same time it has been shown that learning of higher-order information can be selectively impaired (in dyslexia: W. Du & Kelly, 2013; J. H. Howard et al., 2006; in Parkinson's disease: Smith & McDowall, 2004; in Schizophrenia: Schwartz, Howard, Howard, Hovaguimian, & Deutsch, 2003; with age: J. H. Howard, Howard, Dennis, & Yankovich, 2007; D. V. Howard et al., 2004; Feeney, Howard, & Howard, 2002; J. H. Howard & Howard, 1997; Urry, Burns, & Baetu, 2018). It is a matter of question, though, whether lower- and higher-order sequence learning should be thought of as worse or better performance on the same measure, or as different abilities.

In sum, the lack of correlation between different measures of implicit statistical learning may indicate that implicit statistical learning is not a unitary process but rather many processes or a multicomponential one, which possibly vary for different kinds of statistics that can be learned. That being said, it would be of outstanding importance to define the type of statistical learning for every previously used task (e.g. a visuomotor sequence learning task with second-order conditional probabilities) instead of just referring to „implicit statistical learning" in general. Additionaly, extensive work is required to determine to what extent do different types of statistical learning share characteristics, e.g. resistance to interference, on-line and/or offline consolidation, sleep-dependent consolidation, sensitivity to instructions (interaction with explicit processes), etc.

I/4. The psychometric properties of the tasks

As noted earlier, it is also possible that there is a relationship between different types of implicit statistical learning (or between different tasks, assuming a single ability behind every task), and the reason for not being able to see the relationship is related to the psychometric properties of the resulting learning scores.

I/4.1. Low reliability

Other things being equal, the correlation between two variables will be low when the reliability of the measures are low (i.e. measurement error is high). Since reliability is the correlation of a test with itself, therefore it is easy to see that a measure that does not correlate with itself can not correlate with other variables either (Goodwin & Leech, 2006).

Unfortunately, implicit measures are generally considered less reliable than explicit measures (Lebel & Paunonen, 2011). A possible explanation blames the often vague and ambiguous instructions (for example in the Weather Prediction task participants have no solid idea on what basis should they guess, translating to very diverse cognitive and noncognitive processes contributing to performance for a given individual) (Buchner & Wippich, 2000). Additionaly, many measures are based on reaction times which vary considerably from one testing situation to the next as a function of psychological, hormonal, emotional or other factors, leading to high variability (Lebel & Paunonen, 2011), although Buchner & Wippich (2000) also speculated that speeded responding leads to better reliability than responding with no time limits. Furthermore, learning scores computed as difference scores (which often is the case) lead to another problem: such aggregate scores suffer in reliability in direct proportion to the correlation between the two components the difference score was computed from (Edwards, 2001); Kaufman et al. (2010) even suggested that RT difference scores tend to be too unstable to provide rank-ordering between individuals.

I/4.2. Low individual variability

It is hypothesized that implicit learning is evolutionarily older than explicit learning, implying that it is also more robust and results in less inter- and intra-species variability (A. S. Reber & Allen, 2000). It has been assumed that individual differences in implicit cognition are minimal relative to individual differences in explicit cognition (A. S. Reber, 1993). In line with this assumption, the individual differences in implicit cognition remained largely unexplored (A. S. Reber & Allen, 2000; but see Kaufman et al., 2010; and Kalra et al., 2019).

Although the assumption of low individual variability is far from being empirically proven, it may give us a concern because (other things being equal) the value of the correlation coefficient is greater if there is more variability among the observations (Goodwin & Leech, 2006). Additionally, low variability may also stem from floor effects, ceiling effects or artifacts that contaminate the measures of implicit learning. If any of these factors applies (for at least some of the measures), it may explain the lack of correlation between different measures of implicit learning.

I/4.3. Issues related to reaction-time based measures

As noted earlier, difference scores derived from reaction times are thought to be unstable (Kaufman et al., 2010), and it was suggested that accuracy (Urry, Burns, & Baetu, 2015; Urry et al., 2018) or reaction time ratio measures (Kaufman et al., 2010) provide better measures of learning, and are less prone to result in floor effects (Urry et al., 2015). The fact that difference scores based on reaction times and difference scores based on accuracy do not show correlation (Hedge, Powell, Bompas, Vivian-Griffiths, & Sumner, 2018) also implies that the choice between the two types of measures should not be based on convenience or traditions only, but should be a matter of theoretical consideration. Accordingly, tasks that are based on accuracy percentages (responses being made without time limit, such as the Weather Prediction task) could be uncorrelated with reaction-time based measures (such as the SRT or ASRT) for methodological reasons rather than theoretical ones.

Second, there is an often-overlooked factor that might influence serial reaction time tasks, namely that different series of responses are not equally easy to be performed, e.g. responding to the same stimuli many times in a row is easier than responding to an unsystematic order of stimuli. This is sometimes referred to as „pre-existing sequential effects" and „preexisting biases" in the context of serial reaction time tasks (Song, Howard, & Howard, 2007a) or, more generally, „sequential effects" in the context of the broader category of forced-choice reaction time tasks (e.g. Remington, 1969). Complementary to these cognitive effects, there are also biomechanical constraints of the body that also affect serial reaction times (Y. Du & Clark, 2017), as not all effectors (e.g. fingers) are equally efficient in responding. Apart from manifesting as an artifact, and thus influencing our interpretations of the results, these biases might also mask the individual variability of implicit learning (given that they are robust and similar in direction for every participant), and seemingly increase the reliability of the task. Low variability, makes it harder to detect any relationship of implicit learning measures with each other or with other measures of cognitive abilities; seemingly higher reliability, on the other hand, gives the illusion that the results are more trustworthy than they actually are (since it stems from the artifact rather than from the effect we intended to measure).

I/5. Questions and aims of the studies

Taken together, there is a myriad of questions regarding the methodology and analysis methods in the research of implicit statistical learning that needs to be clarified. The nature of the resulting statistical knowledge should be assessed for each (possible) subtype of statistical learning – considering modality, the type of statistics embedded in the task, etc. so that we could get to a conclusion about the theoretical questions (what factors matter and how). Also, psychometric properties of the tasks used should be routinely reported, along with the observed individual variability in a particular experiment and the assessment of possible artifacts biasing the results. Only this way could we be sure that the theory that we build is not the by-product of questionable methodology.

Admittedly, this is a very ambitious goal requiring lots of investment. In the present Dissertation, I present four studies covering only a tiny slice of these goals: to increase our knowledge about the nature of implicit statistical learning that could be measured with the ASRT task, to learn about the psychometric properties of the task, and to improve the analysis methods to overcome its flaws.

I/5.1. About the ASRT task

The ASRT task was introduced in 1997 as a means of measuring implicit memory (J. H. Howard & Howard, 1997). In the original task, visual stimuli are presented on a computer screen in one of four possible locations, and the subject's task is to react as fast and as accurately as possible to the location of the stimuli by pressing the corresponding response button (usually aligned to stimuli to allow for a simple 1:1 stimulus-response mapping). Thus, due to necessity of a collaboration between visual and motor components, one might consider the ASRT a **visuomotor task**.

The stream of stimuli is not entirely random: a pre-defined four-element long pattern (P) is embedded in a stream of random (R) trials so that P and R trials alternate (hence the name of the task). This alternation is crucial as it allows for the comparison of performance on the predetermined (P) and random (R) trials continuously, in contrast with the SRT task (Nissen & Bullemer, 1987) in which the uninterrupted stream of pattern trials is occasionally followed by an uninterrupted stream of random (or pseudo-random) trials, and learning can only be assessed at these occasions by comparing performance on the random chunk to the performance on the surrounding pattern chunks.

Learning on the task may not (entirely) rely on subjects ability to differentiate between pattern and random trials. The structure that results from their alternation is a second-order probabilistic sequence. A second-order structure means that the basic units of the statistical structure are three consecutive trials, so-called triplets; some triplets are frequent and others are infrequent. In this particular case, after encountering any two consecutive trials, a prediction could be made of what to expect next. The term *probabilistic* refers to the fact that sometimes the following trial is „unexpected", not very probable. Learning can be derived from the comparison of performance on probable versus improbable trials (e.g. Nemeth et al., 2011; Nemeth, Janacsek, Londe, et al., 2010; Nemeth, Janacsek, Polner, & Kovacs, 2013). Thus ASRT may be thought of as a measure of **statistical learning**. It is an open question whether it can be also considered as a measure of pattern learning (i.e. whether humans are capable of learning to differentiate between P and R trials in addition to being able to discriminate between statistical properties of trials; the two are heavily confounded).

As a difference to the aforementioned SRT task, learning on the ASRT task is thought to be more clearly implicit. The authors introducing the task reported that not a single subject became aware of the hidden pattern (J. H. Howard & Howard, 1997), and our experience with the task corroborates their notion. Thus, the ASRT task measures **implicit learning**.

In summary then, the ASRT task is an implicit visuomotor statistical learning task measuring the ability to acquire second-order probabilistic information.

I/5.2. Open questions about the ASRT task and the resulting knowledge

As noted above, the ASRT is typically considered a visuomotor task, but the contribution of the visual and motor components has not been systematically studied before. Since statistical learning is thought to be modality specific (Emberson, Conway, & Christiansen, 2011; Li, Zhao, Shi, Lu, & Conway, 2018; Walk & Conway, 2016), it is possible than ASRT measures more than one type of statistical learning (i.e. in the visual and motor domains). If so, the relative contribution of the two is to be determined. Second, if learning of the visual and motor stream is separable, it is also possible that the resulting representations differ in some aspects (e.g. the magnitude of learning, consolidation (and sleep) effects, etc.), which are to be determined.

Independently from the question of modality, one also needs to explore the nature of the statistical knowledge that results from the experience with the ASRT task more generally. Is it prone to interference effects? If so, to what extent? Is there an interaction with other cognitive abilities (e.g. can we „boost" implicit learning by providing explicit information)?

Finally, the last line of quesitons addresses the of the utility of the task itself. Is the ASRT task reliable? Is it possible to differentiate between different types of statistical learning using the ASRT task? Are the learning scores affected by pre-existing biases – and if so, to what extent? How can we overcome these obstacles?

I/5.3. Aims of the studies

In **Study 1** the main question was whether perceptual information is learned in a temporally structured visuomotor sequence such as the ASRT (in addition to motor sequencing), and if so, then whether perceptual learning is comparable to motor learning in the paradigm. In order to assess this question, we modified the ASRT task so that stimuli always appeared in the center of the screen (and their identity was differentiated based on perceptual features rather than the location of appearance), this way eye movements were minimized. In **Study 2** we extended our findings with assessing consolidation of these different learning types with the inclusion of off-line periods that either included sleep or not. This way the question of modality-specificity was assessed.

In **Study 3** we addressed the question of interference between similar (but different) sequences learned in succession; whether the sequence learned in the first place could be „overwritten" with a second sequence, whether there are costs associated with the proactive interference caused by the first sequence; and whether it really gets „overwritten" (rewired) or the knowledge for both sequences is accessible later. Additionally, consolidation was also addressed, since the experiment took place on three consecutive days allowing for the assessment of benefits of these off-line periods. Finally, an important question related to the effect of explicit (top-down) knowledge about the rule (but not about the statistical structure) embedded in the sequence, and whether this knowledge – or the differences in participants mindsets owing to this knowledge – results in differences in implicit statistical learning measured on trials on which the explicit knowledge could not be utilized. This way, the interaction between implicit and explicit processes was assessed.

In **Study 4** our goal was two-fold. First, we wanted to show that the ASRT task makes it possible to assess the learning of both second-order and third-order statistical structure without any modification to the task (just by refining the analysis methods), and also assess the question of pattern (rule) learning, i.e. whether participants learn about the alternating structure of the sequence in addition to its statistical properties. We have also compared the currently/typically used analysis methods with the proposed method (in terms of goodness of fit). Second, we assessed the psychometric properties of the task (both with the typical analysis methods and with the newly proposed method), and we suggested the application of a filter to lessen the impact of pre-existing (cognitive or biomechanical) biases to certain stimulus combinations which could result in artifacts in the learning scores.

II. PERCEPTUAL AND MOTOR FACTORS OF IMPLICIT SKILL LEARNING

(Study 1)[6]

II/1. Abstract

Implicit skill learning underlies not only motor but also cognitive and social skills, and represents an important aspect of life from infancy to old age. Earlier research examining this fundamental form of learning has shown that learning relies on motor and perceptual skills, along with the possible role of oculomotor learning. The goals of this study were to determine whether motor or perceptual cues provide better prompts to sequence learning and to remove the possibility of oculomotor learning during the task. We used a modified version of the probabilistic alternating serial reaction time task, which allowed the separation of motor and perceptual factors. Our results showed that motor and perceptual factors influenced skill learning to a similar extent.

Keywords: alternating serial reaction time, implicit Skill learning, motor learning, oculomotor learning, perceptual learning

---

6 Nemeth, D., **Hallgató, E**., Janacsek, K., Sándor, T., & Londe, Z. (2009). Perceptual and motor factors of implicit skill learning. NeuroReport, 20(18), 1654. https://doi.org/10.1097/WNR.0b013e328333ba08

II/2. Introduction

Implicit skill learning occurs when information is acquired from an environment of complex stimuli without conscious access either to what was learned or to the fact that learning had occurred (A. S. Reber, 1993). In everyday life, this learning mechanism is crucial for adapting to the environment and to evaluate events. The most important models of skill learning in cognitive neuroscience and neuropsychological studies emphasize the role of the basal ganglia and the cerebellum (Doyon, Bellec, et al., 2009; Hikosaka et al., 1999; Hikosaka, Nakamura, Sakai, & Nakahara, 2002), although the role of the hippocampus remains inconclusive (Albouy et al., 2008; Schendan, Searl, Melrose, & Stern, 2003). Skill learning can be differentiated into phases (an initial rapid phase and a subsequent slower phase), into types (motor, visuomotor, or perceptual such as visual, auditory, etc.), and into consciousness types (implicit and explicit) (Doyon, Bellec, et al., 2009). Implicit motor skill learning tasks have been used for decades, but there is no agreement about how these tasks reflect motor versus perceptual learning, and what their proportions are.

The most widely used task to measure skill learning is the serial reaction time (SRT) task (Nissen & Bullemer, 1987). In this task, the stimulus appears in one of four possible positions on the screen and the participant has to press the appropriate response key as fast as possible. The stimuli follow a predefined sequence, and although the research subjects are not aware of this, they perform better on these trials than in corresponding random trials. In most SRT tasks, the location of the stimulus corresponds to the location of the response key. Therefore, learning can be influenced by the sequence of stimuli locations on the screen (perceptual learning), by the correct answer button sequence in the egocentric space (answer-based learning) or by the finger movement patterns (effector-based learning) (Remillard, 2003).

Another disadvantage of these paradigms (classical SRT and finger-tapping tasks) is that after a short training session, the participants often recognize the stimulus pattern, which causes significant limitations in studying implicit learning (J. H. Howard & Howard, 1997). In contrast, using the alternating SRT (ASRT) task (J. H. Howard & Howard, 1997) allows researchers to overcome this aforementioned problem by using an eight-element sequence, whereby random elements alternate with sequence elements (e.g.: 2–R–3–R–1–R–4–R, where R refers to random).

In these research paradigms, it is difficult to isolate perceptual learning. Specifically, motor learning cannot be eliminated in both observation-based and transferbased studies because it is the motor response reaction time (RT) that gives the informative measurements (Dennis, Howard, & Howard, 2006). Perceptual learning in these paradigms can be observed only if it can be shown in addition to implicit skill learning. For example, Robertson et al. (E. M. Robertson, Tormos, Maeda, & Pascual-Leone, 2001) showed that if perceptual and motor sequences are combined (e.g. color and location), it leads to a greater level of learning than either one of the sequences alone.

In the case of first-order probability sequences, motor learning is not necessary to learn patterns. However, in second-order probability sequences (e.g. ASRT), perceptual learning is, at best, minimal (Remillard, 2003). Nevertheless, previous studies have been able to isolate perceptual learning based on second-order or higher-order probability sequences (Deroost, Coomans, & Soetens, 2009). For example, Dennis and colleagues (2006) found that young adults showed implicit skill learning in higher-order sequences even without motor learning. Moreover, when no motor response was requested, deterministic sequence learning (e.g. SRT) led to explicit learning by simply observing the stimuli, whereby participants revealed the hidden sequence explicitly (J. H. Howard & Howard, 1997; Willingham, Nissen, & Bullemer, 1989). In the case of second-order sequences, explicit knowledge has been shown to be minimal or totally eliminated (J. H. Howard & Howard, 1997). Song et al. (Song, Howard, & Howard, 2008) showed perceptual learning using similar tasks and found that learning took place even without a motor response to the observed stimuli. After the observation, participants were able to transfer the sequence knowledge to the testing (motor) condition. The concern with this study was that the stimuli appeared on four different areas of the screen. Hence, skill learning could have reflected oculomotor learning as well (for example, Song et al., 2008). The question remains whether learning is purely perceptual when it is accompanied with eye movements. Remillard (2003) found that perceptual learning was not influenced by the distance between the stimuli (i.e. the amplitude of the eye-movement). In contrast, Willingham et al. (1989) were not able to show perceptual learning without eye movements.

Willingham et al. (Willingham, Wells, Farrell, & Stemwedel, 2000) changed the conditions of the SRT task after the learning phase in one of the two following ways:

either the stimulus sequence (perceptual information) remained the same as in the learning phase while the sequence of the answers (motor information) was changed, or the motor response sequence remained the same and the response locations changed (participants had to answer crossing their hands during the testing phase). Participants were able to transfer their knowledge only when the sequence of response locations was maintained, not the sequence of finger movements (Willingham et al., 2000). These findings suggest that the sequence of response locations must have been retained for implicit knowledge to transfer, whereas the contribution of motor and perceptual information was less considerable. It is important to note that Willingham et al. (Willingham et al., 2000) did not eliminate the possibility of oculomotor learning as the sequence occurred perceptually in the locations of the stimuli.

The goal of this study was to investigate the role of perceptual learning in implicit sequence learning through a modified ASRT task. In this modified paradigm, the sequence followed a second-order regularity that eliminated the possibility of oculomotor learning because the stimuli always appeared in the same, central position. Similar to the study by Willlingham et al. (Willingham et al., 2000) in the learning phase, the sequence of stimuli and their responses were different. In the second phase (testing or transfer phase), the sequence of stimuli (perceptual information) remained the same and the response sequence (motor information) changed or vice versa.

Our hypothesis was that, unlike Willingham et al. (Willingham et al., 2000), we would be able to show perceptual learning or perceptual transfer with a task that eliminated oculomotor learning. In addition, our goal was to create a task that would distinguish between perceptual and motor factors of implicit sequence learning.

II/3. Methods

II/3.1. Participants

Thirty-four healthy right-handed individuals took part in the experiment. Half of the participants were randomly assigned to the perceptual condition (mean age M = 21.76 years, SD = 2.02; 7 male/10 female), and the other half were assigned to the motor condition (mean age M = 21.76 years, SD = 1.64; 8 male/9 female). Participants did not suffer from any developmental, psychiatric, or neurological disorders. All participants provided signed informed consent agreements and received no financial compensation for their participation.

II/3.2. Task and procedure

We used a modified version of the ASRT task (J. H. Howard & Howard, 1997), the socalled AS-RT-Race. We created a story about a car race for the task. The stimuli were the left, right, up, and down arrows (5 cm long and 3 cm wide), which appeared on the center of the screen. When the stimulus appeared on the screen, it represented the car's direction. For example, when the participants saw an up arrow, they had to press the up button on the keyboard to move the car forward, the left button to turn left, and so on. All participants pressed the keys with their dominant hand.

After the starting block of 85 random presses, they were told that there was a car crash and the steering wheel failed (**Fig. II/1/a**). The car now kept going to the left if they wanted to go straight, but by turning the steering wheel right they could correct this malfunction, and could continue to go straight. Thus participants had to mentally rotate the arrows (the steering wheel) by 90° to the right, and press the button corresponding to this rotated arrow.

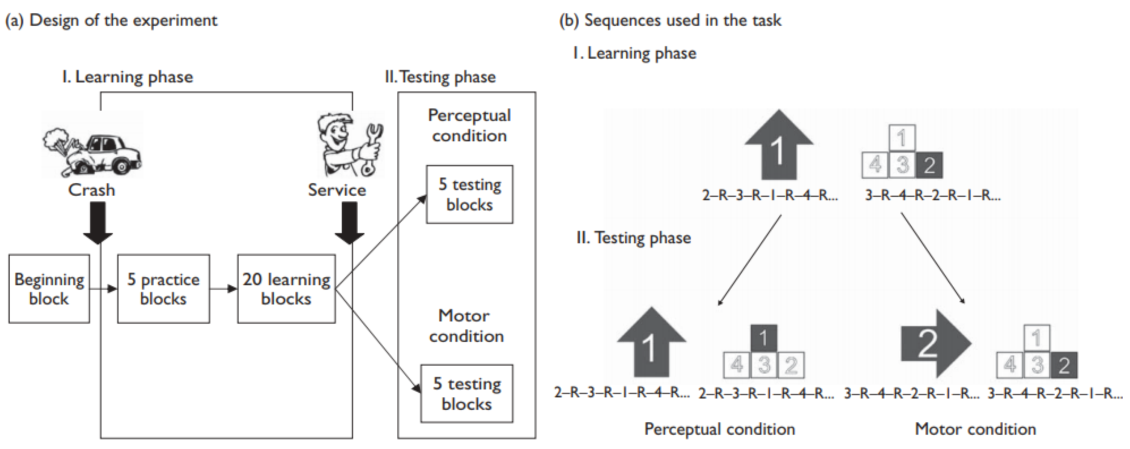


Figure II/1. a) Schematic diagram of the experiment. b) In the perceptual condition, the perceptual sequence was the same and the motor sequence (button presses) changed compared with the sequences in hte learning phase. In the motor condition, key presses followed the learned sequence and the preceptual information changed.

In the learning phase, five practice blocks were presented (these were excluded from the analysis), followed by 20 learning blocks with 85 key presses in each block. These 85 key presses included an initial five random presses (warm-up; excluded from the analysis), then an eightelement sequence alternated 10 times (2–R–3–R–1–R–4-R, where R represents random trials). The stimulus remained on the screen until the participant pressed the correct button. The next stimulus appeared after a 120-ms delay (response to stimulus interval) after the participant's correct response (following the parameters of the original task by Howard and Howard (J. H. Howard & Howard, 1997). During this delay, a fixation cross was displayed on the screen. Participants were told to respond as fast and as accurately as they could.

After the learning phase (and a 3-min-long break), the participants were told that the car had been taken to a service station and the steering wheel had been fixed. They were told to use the answer keys corresponding to the arrows that appeared on the screen (up button for up arrow, left button for left arrow, etc.). In the testing phase, half of the participants were assigned to the perceptual condition and the other half to the motor condition (**Fig. II/1/a**). In the perceptual condition, participants responded to the sequence seen during the learning phase (e.g. 2–R–3–R–1–R–4–R, **Fig. II/1/b**), and the appropriate key presses represented a new sequence (also 2–R–3–R–1–R–4–R), which they had not practiced before. In contrast, participants in the motor condition had to respond by key presses practiced before (e.g. 3–R–4–R–2–R–1–R, **Fig. II/1/b**) but the corresponding stimuli on the screen followed another sequence (also 3–R–4–R–2–R–1–R), which they had not seen before. Thus, in the perceptual condition, the perceptual sequence was the same but the motor sequence (key presses) changed compared with the previously practiced sequence. However, in the motor condition, key presses followed the previously learned sequence and the perceptual information (the sequence of the stimuli displayed on the screen) changed. By comparing the participant's performance between the two conditions, we could determine whether the perceptual and the motor component had the same or different effects on learning. The possible oculomotor aspect of learning was excluded by displaying all the stimuli in the same place (in the center) of the screen.

To explore how much explicit knowledge the participant acquired about the task, we used a short questionnaire after the testing phase. None of the participants reported noticing the sequences in the tasks.

II/3.3. Statistical analysis

We followed the procedures of the original ASRT task (Bennett, Howard, & Howard, 2007; Song, Howard, & Howard, 2007b) in our analysis because the core structure of the tasks was the same. Given that there was a fixed sequence in the AS-RT-Race task (and in the ASRT task as well), which included alternating random elements (e.g. 2–R–3–R–1–R–4–R), some triplets or runs of three events occurred more frequently than others. For example, in the above illustration, triplets such as 2_3, 3_1, 1_4, 4_2 could occur more frequently because the third element could be derived from the sequence or could also be a random element. In contrast, triplets such as 4_1, 4_4 would occur less frequently, because in this case, the third element could only be random. In other words, pattern trials were always high frequency, whereas one-fourth of random trials were high frequency by chance. Previous studies have shown that as participants practice, they come to respond more quickly to the highfrequency compared with the low-frequency triplets, thereby revealing sequence-specific learning (triplet type effect; (D. V. Howard et al., 2004; J. H. Howard & Howard, 1997; Song et al., 2007a)). In addition, general motor skill learning was revealed by the overall speed with which participants responded, irrespective of the triplet types. Thus, we obtained measures of both sequence-specific and general motor skill learning in the AS-RT-Race task.

The blocks of the AS-RT-Race task were organized into groups of five to facilitate data processing. A group of five blocks was referred to as an ‘epoch’ (a term given by the ASRT authors). The first epoch contained blocks 1–5, the second epoch contained blocks 6–10, etc. Our analysis focused only on RT data because participants’ accuracy remained very high during the entire test (the average was 97% for both conditions in both the learning and testing phases). Median RTs were calculated for each participant and in each epoch both for the high-frequency and low-frequency triplets.

II/4. Results

II/4.1. Learning phase

The 2 (triplet: high and low) 4 (epochs: 1–4) repeatedmeasures analysis of variance with condition (perceptual vs. motor) as the between-subject factor revealed sequence-specific learning [indicated by a significant main effect of the triplet: $F(1,23) = 124$, mean square error $MSE = 56.65$, $p < 0.001$, $\eta_p^2 = 0.63$, as well as general motor skill learning. shown by the significant main effect of the epoch: $F(4,20) = 8.85$, $MSE = 32.53$, $p < 0.001$, $\eta_p^2 = 0.72$, thereby suggesting that the more the participants practiced, the faster their responses became (**Fig. II/2/a** and **Fig. II/2/b**). The two groups (perceptual and motor conditions) did not differ either in sequencespecific or in general motor skill learning ($p > 0.31$).

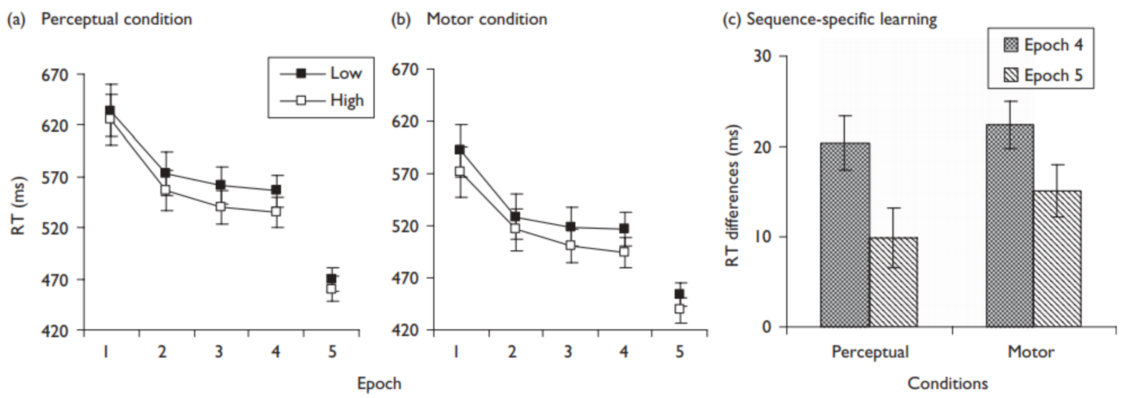


Figure II/2. Results of the learning phase (Epochs 1-4) and testing phase (Epoch 5) for perceptual (a) and motor (b) conditions. Filled squares represent low-frequency triplets; open squares represent high-frequency triplets. Comparing the sequence-specific knowledge (the reaction time (RT) differences between high-frequency and low-frequency triplets) of perceptual and motor conditions (c). Error bars indicate standard error of mean (SEM).

II/4.2. Testing phase

To compare the perceptual and motor conditions in the testing phase, a 2 (triplet: high and low) 2 (epochs: 4–5) repeated-measures analysis of variance was conducted with condition (perceptual vs. motor) as the betweensubject factor. The main effect of the triplet was significant, $F(1,32) = 69.72$, $MSE = 139.36$, $p < 0.001$, $\eta_p^2 = 0.69$, such that participants responded faster for high-frequency than for low-frequency triplets

(**Fig. II/2/c**). The main effect of the epoch was also significant, $F(1,32) = 115.4$, $MSE = 1448.27$, $p < 0.001$, $\eta_p^2 = 0.78$, whereby participants were faster in the testing phase (455 ms) than in the learning phase (525 ms). Interestingly, the triplet epoch interaction was also significant, $F(1,32) = 5.75$, $MSE = 117.79$, $p = 0.02$, $\eta_p^2 = 0.15$, thereby suggesting that the sequence-specific knowledge decreased between the learning and the testing phases (the RT difference between the high-frequency and low-frequency triplets was 21 ms in epoch 4 and 12 ms in epoch 5). However, despite this decrease, participants still showed a significant triplet type effect in epoch 5, indicated by a one-sample t-test: $t(33) = 4.52$, $p < 0.001$. In addition, there was no difference between the conditions either in sequence-specific ($p = 0.38$) or in general motor skill ($p = 0.10$).

II/5. Discussion

Our research investigated the role of perceptual and motor learning in implicit skill learning. We addressed the possibility of demonstrating perceptual transfer beyond motor learning in a testing situation where, after the learning phase, the task continues either with motor sequence or with perceptual sequence while eliminating oculomotor learning. We were able to show learning after the learning phase both in the perceptual and motor conditions. We focused on the perceptual sequence transfer under the former condition, and the motor sequence in the latter. Our results showed that under this research paradigm, both motor and perceptual transfer was significant. These results support the different methods of Song et al. (Song et al., 2008), which showed perceptual learning with probabilistic sequence learning tasks. In contrast, our results partly differ from that of Willingham et al. (Willingham et al., 2000), which did not find perceptual learning to be an important element of learning. Their research design, however, did not eliminate the possibility of oculomotor learning, whereas this study did. Furthermore, our findings also indicated that there was motor transfer, thereby supporting the results of Willingham et al. (Willingham et al., 2000) and their implicit motor sequence learning model.

Our findings well complement motor skill learning models (Doyon, Bellec, et al., 2009; Hikosaka et al., 1999, 2002), as well as the neuropsychological and neuroimaging studies that suggest the basal ganglia and the primary and secondary motor cortices play a role in implicit skill learning (Doyon, Bellec, et al., 2009; Grafton, Hazeltine, & Ivry, 1995; E. M. Robertson, Press, & Pascual-Leone, 2005; Willingham

& Koroshetz, 1993). The task developed in this study separated motor and perceptual learning, thereby allowing researchers to conduct more detailed studies in cognitive neuroscience for various pathologies affecting implicit skill learning and the underlying mechanisms of motor and perceptual learning.

II/6. Conclusion

In our study, we constructed a novel task (AS-RT-Race) to separate the perceptual and motor factors of implicit skill learning. We found that these components underlie the mechanisms behind skill learning to nearly the same extent. Our results draw attention to the fact that skill learning is not a single process. Instead, there are multiple mechanisms in this fundamental learning process. The novel task we developed was shown to be an appropriate method to investigate the components of skill learning in different neuropsychological pathologies (e.g. basal ganglia disorders, Alzheimer's disease, etc.), and for examining the effects of development, aging, and sleep on the motor and perceptual factors contributing to skill learning.

II/7. Acknowledgements

This work was supported by the Bolyai Scholarship Program and the Hungarian National Research Fund (OTKA F 61943). The authors thank Ágnes Lukács and Tamás Kincses for their helpful comments. The authors report no conflict of interest and have no financial disclosure.

III. THE DIFFERENTIAL CONSOLIDATION OF PERCEPTUAL AND MOTOR LEARNING IN SKILL ACQUISITION

(Study 2)[7]

III/1. Abstract

Implicit skill learning is an unconscious way of learning which underlies not only motor but also cognitive and social skills. This form of learning is based on both motor and perceptual information. Although many studies have investigated the perceptual and motor components of “online” skill learning, the effect of consolidation on perceptual and motor characteristics of skill learning has not been studied to our knowledge. In our research we used a sequence learning task to determine if consolidation had the same or different effect on the perceptual and the motor components of skill acquisition. We introduced a 12-h (including or not including sleep) and a 24-h (diurnal control) delay between the learning and the testing phase with AM-PM, PM-AM, AM-AM and PM-PM groups, in order to examine whether the offline period had differential effects on perceptual and motor learning. Although both perceptual and motor learning were significant in the testing phase, results showed that motor knowledge transfers more effectively than perceptual knowledge during the offline period, irrespective of whether sleep occurred or not and whether there was a 12- or 24-h delay period between the learning and the testing phase. These results have important implications for the debate concerning perceptual/motor learning and the role of sleep in skill acquisition.

Keywords: Consolidation, Implicit skill learning, Offline learning, Perceptual-motor learning, Sleep

[7] **Hallgató, E.,** Győri-Dani, D., Pekár, J., Janacsek, K., & Nemeth, D. (2013). The differential consolidation of perceptual and motor learning in skill acquisition. Cortex, 49(4), 1073–1081. https://doi.org/10.1016/j.cortex.2012.01.002

III/2. Introduction

Implicit skill learning occurs when information is acquired from an environment of complex stimuli without conscious access either to what was learned or to the fact that learning occurred (A. S. Reber, 1993). In everyday life, this learning mechanism is crucial for adapting to the environment and evaluating events. Implicit skill learning underlies not onlymotor but cognitive and social skills as well, it is therefore an important aspect of life from infancy to old age. Skill learning does not occur only during practice, in the so-called online periods, but also between practice periods, during the so-called offlineperiods. The process that occurs during the offline periods is referred to as consolidation which means stabilization of a memory trace after the initial acquisition. This process can result in increased resistance to interference or even improvement in performance following an offline period (Krakauer & Shadmehr, 2006; Nemeth, Janacsek, Londe, et al., 2010; Edwin M. Robertson, 2009; Song, 2009).

Most models of skill learning (Dennis & Cabeza, 2011; Doyon, Bellec, et al., 2009; Hikosaka et al., 1999, 2002; Keele, Ivry, Mayr, Hazeltine, & Heuer, 2003; Kincses et al., 2008) highlight the role of the basal ganglia and the cerebellum. One of the main debates in the field of skill learning is whether we rely on "our hands" or on "our eyes" (Deroost & Soetens, 2006; Keele et al., 2003; Mayr, 1996; Nemeth, Hallgató, Janacsek, Sándor, & Londe, 2009; Song et al., 2008; Ziessler & Nattkemper, 2001)? The goal of the present study is to determine if an offline period modifies the contribution of motor and perceptual components to implicit sequence learning. This issue is of particular interest because it deals with the question of whether sequence learning and consolidation are mediated by perceptual or by motor brain networks primarily (Deroost & Soetens, 2006; Goschke, 1998)

One of the most popular implicit learning tasks is the Serial Reaction Time (SRT) Task (Nissen & Bullemer, 1987) and its modification, the Alternating Serial Reaction Time (ASRT) Task (J. H. Howard & Howard, 1997; Nemeth, Janacsek, Londe, et al., 2010). In the original version a stimulus appears at one of four possible locations on the screen, and subjects have to press the button corresponding to that location. Unbeknownst to them, the sequence of subsequent locations (and correspondingly, the sequence of the responses) follows a predetermined order. Without becoming aware of the sequence, subjects learn the regularity - and as they learn, they produce faster and more accurate responses. When the sequence is changed to a random series of stimuli, subjects become slower and less accurate in responding. In this paradigm, however, it is not clear what exactly the subjects learn: they might learn the sequence of the stimuli (perceptual learning), the sequence of their own eye movements (oculomotor learning), the sequence of response locations (response-based learning) or the sequence of given fingers' movements (effector-based learning) (Cohen, Ivry, & Keele, 1990; Remillard, 2003; Willingham, 1999).

In a SRT study Willingham (1999) used two conditions to examine the perceptual and the motor factors of learning. In one condition the stimuluseresponse mapping was changed in the transfer (test) phase that followed the learning phase, so that half of the subjects had to press the same sequence of keys as in the learning phase but saw new stimuli, whereas the other half had to press a different sequence of keys as in the learning phase but saw the same stimuli as before. Willingham (1999) found that transfer was shown only when the motor sequence was kept constant, but not when the perceptual sequence was constant. In a previous study, Nemeth et al. (2009) compared the magnitude of perceptual and motor implicit sequence learning using a modification of the ASRT-task in a similar design. This task (ASRT-Race) contains second-order probabilistic sequences compared to classical SRT tasks that use deterministic sequences. ASRTRace allows measuring "pure" sequence learning separate from general skill improvements, where sequence learning is reflected in the difference between the reaction times to more predictable events as opposed to less predictable ones. In addition, this task eliminates the possibility of oculomotor learning as stimuli always appear in the same central position on the screen. In contrast to Willingham's findings, Nemeth et al. (2009) demonstrated that not only motor, but perceptual learning of second-order probabilistic sequences is possible. Furthermore, Nemeth et al. (2009) showed that the two types of learning do not differ significantly in magnitude. The weakness of the above mentioned perceptual-motor studies (Deroost & Soetens, 2006; Mayr, 1996; Nemeth et al., 2009; Remillard, 2003, 2009; Song et al., 2008; Willingham, 1999) is that experiments were conducted in one session. Using only one session for measuring skill learning relates to short-term performance changes in behavior and not to more permanent changes associated with learning. Consequently, it is important to address the question of the role of offline periods in perceptual and motor skill learning.

Recent reviews indicate that whether offline improvements occur at all, and whether they are sleep-dependent, varies with factors such as awareness, the formation of contextual associations and type of information to be learned (Debas et al., 2010; Doyon, Korman, et al., 2009; Nemeth, Janacsek, Londe, et al., 2010; Edwin M. Robertson, 2009; Edwin M. Robertson, Pascual-Leone, & Press, 2004; Siengsukon & Boyd, 2008; Song, 2009; Song et al., 2007b). For example, Robertson (2009) argues that the consolidation of explicit (goal-directed) and implicit (movement-based) learning is differentially affected by sleep and wakefulness. In implicit learning when there is no declarative knowledge about the task, consolidation may occur during both wakefulness and sleep. In line with the predictions of this theory, recent SRT studies found similar consolidation of implicit skills during both sleep and wakefulness (Nemeth, Janacsek, Londe, et al., 2010; Edwin M. Robertson et al., 2004; Song et al., 2007b).

Although many researches have investigated the perceptual and motor components of "online skill learning", to our knowledge, the effect of consolidation on perceptual and motor characteristics of skill acquisition has not been investigated so far (Deroost & Soetens, 2006; Mayr, 1996; Nemeth et al., 2009; Remillard, 2003, 2009; Song et al., 2008). In our study we used the ASRT-Race task (Nemeth et al., 2009) to examine the possible difference in the magnitude of motor and perceptual learning after a 12-h and a 24-h retention period. In addition, we also aimed at exploring the role of sleep in offline consolidation of these two factors of skill learning. Therefore a 12-h delay was administered between the Learning Phase and Transfer Phase of the experiment, during which participants either had a sleep (night group) or they were awake (day group). If both groups acquire the same level of skill in the Learning Phase, any difference between them in the Transfer Phase will answer the question whether the perceptual or the motor component stabilizes more effectively during the offline period. In order to avoid a time-of-day effect we also administered a 24-h delay condition.

III/3. Methods

III/3.1. Participants

There were 102 individuals (students attending the University of Szeged) in the experiment (mean age M = 22.34, standard deviation SD = 3.82; 44 males, 58 females). None of them suffered from any developmental, psychiatric or neurological disorders.

Participants were randomly assigned to the perceptual group or to the motor group. The perceptual and motor groups were further divided by the length of delay (12- or 24-h delay) and by the daytime (morning-first, AM-PM/ AM-AM and evening-first, PM-AM/PM-PM) (see **Table III/1**). The eight experimental groups did not differ in their sleep quality, $F(7,89) = 0.98$, $p = 0.45$, measured by the Pittsburgh Sleep Quality Index (Buysse, Reynolds, Monk, Berman, & Kupfer, 1989) (Due to data collection scheduling problems five out of 102 participants failed to administer this test). All individuals provided signed informed consent, and received no financial compensation for their participation.

Table III/1. General data of participants

Condition	Delay	Daytime	Mean age (SD)	N (Male/Female)
Perceptual	12-h	Morning-first (AM-PM)	20.82 (1.60)	11 (4/7)
		Evening-first (PM-AM)	22.75 (3.74)	11 (7/4)
	24-h	Morning-first (AM-PM)	23.72 (5.66)	14 (4/10)
		Evening-first (PM-AM)	21.63 (2.16)	14 (6/8)
Motor	12-h	Morning-first (AM-PM)	22.62 (3.98)	12 (8/4)
		Evening-first (PM-AM)	22.00 (1.84)	11 (4/7)
	24-h	Morning-first (AM-PM)	20.40 (2.01)	12 (3/9)
		Evening-first (PM-AM)	23.93 (5.48)	17 (8/9)

III/3.2. Procedure

All participants completed two sessions: a Learning Phase (Session 1) and a Transfer Phase (Session 2), separated by a 12- h or a 24-h delay (**Fig. III/1**). For the night groups, Session 1 was in the evening (between 7 pm and 9 pm), and Session 2 was in the morning (between 7 am and 9 am), with the opposite arrangement for the day groups. Thus, the offline period of the night group contained sleep, while the day group was awake during the offline period (**Fig. III/1**). Although previous studies with similar tasks and experimental designs showed no timeof-day effect either on general reaction times or on learning measures (Nemeth, Janacsek, Londe, et al., 2010; Press, Casement, Pascual-Leone, & Robertson, 2005; Edwin M. Robertson et al., 2004; Song et al., 2007b), we administered a 24-h delay condition. For the morning diurnal groups, both Session 1 and Session 2 were in the morning (between 7 am and 9 am) and for the evening diurnal groups, both Session 1 and Session 2 took place in the evening (between 7 pm and 9 pm).

III/3.3. Task

A modified version of the original ASRT (J. H. Howard & Howard, 1997) was used, the so-called ASRT-Race (Nemeth et al., 2009) in which the participants were instructed to drive an imaginary car on the road, as fast and as accurately as possible. The stimuli were the left, right, up and down arrows (5 cm long and 3 cm wide) appearing in the center of the screen, and representing the direction the car had to be steered. For example, when the subjects saw the right arrow, they had to press the right button on the keyboard to make a right turn with the car. All participants pressed the keys with their right hand. Session 1 consisted of 22 blocks, starting with a block containing 85 random presses (excluded from data analysis), after which the individuals were told that there was a car crash and the steering wheel failed. Due to the defective steering wheel they had to mentally rotate the arrows appearing on the screen by 90, and press the keyboard button designated to the rotated arrow, in order to maneuver the car in the right direction (**Fig. III/1/a**). For instance, if they saw the up arrow on the screen they had to press the right arrow on the keyboard, if they saw the right arrow they had to press the down arrow button, and so on (**Fig. III/1/c**). After the change in the instruction, there were 21 blocks, starting with one random block, in which participants could practice the new rules regarding the mental rotation, followed by 20 learning blocks (Learning phase). Each of the 20 learning blocks contained 85 key presses. The initial five stimuli were random (warm-up; excluded from data analysis), then an eight-element sequence alternated 10 times. Since the ASRT-task is based on a nonadjacent sequence, random and sequence elements alternate one after the other. For example 2-R-3-R-1-R-4-R, where R represents random trials and the numbers represent the sequence-specific elements, implicating the arrows' direction (1-up, 2-right, 3-down, 4-left). The stimulus remained on the screen until the participant pressed the correct button. The next arrow appeared following a 120-msec delay (response to stimulus interval) after the subject's correct response. These parameters are consistent with the original task presented by Howard and Howard (1997). During this delay, a fixation cross was displayed on the screen. Participants were told to respond as fast and as accurately as they could.

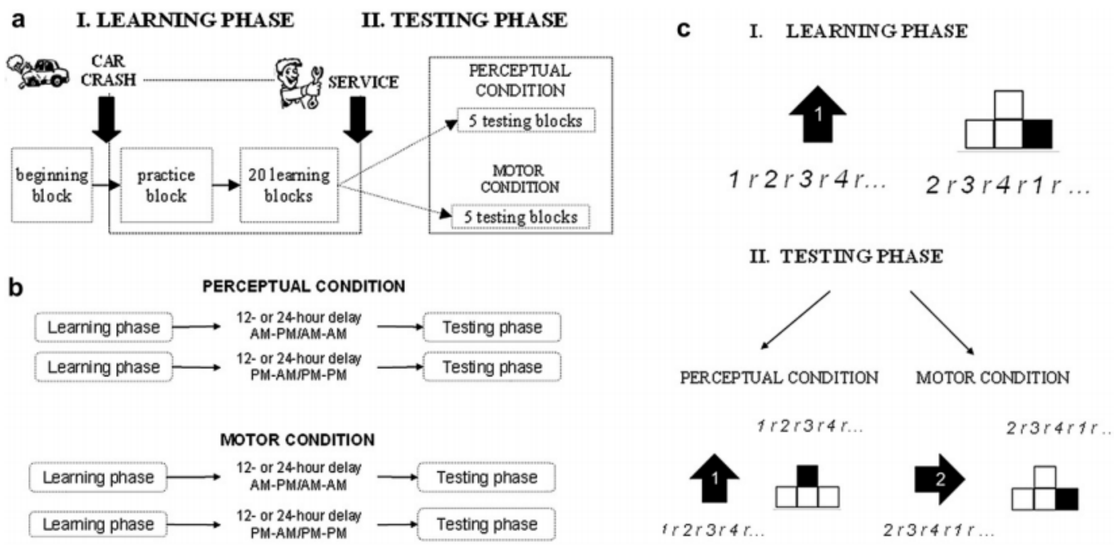


Figure III/1. Design of the experiment. (a) All participants completed the ASRT-Race sequence learning task in two sessions. There were 20 learning blocks in Session 1 and five testing blocks in Session 2. (b) The two sessions were separated by either a 12-h delay (in which participants had or had not slept) or a 24-h delay. (c) In Session 2, half of the subjects were administered in a new sequence which they had not seen before, but whose motor information corresponded to that of they had practiced in Session 1 (motor condition), while the other half of subjects were administered to the same perceptual information as in Session 1, but the pattern of motor responses changed due to the lack of mental rotation (perceptual condition).

Session 2 (Transfer Phase) took place either after a 12-h or a 24-h delay. The Transfer Phase consisted of five blocks. In this session participants were told that the car had been taken to a mechanic, and the steering wheel had been fixed, so they could use the answer keys corresponding to the arrows appearing on the screen (right button for right arrow, down button for down arrow, etc.). Half of the subjects participated in the motor condition, while the other half was assigned to the perceptual condition. Subjects in the motor condition were administered a new sequence which they had not seen before, but whose motor information corresponded to the one of they had practiced in Session 1, while subjects in the perceptual condition were administered to the same perceptual information as in Session 1, but the pattern of motor responses changed due to the lack of mental rotation (**Fig. III/1/c**). Thus, while in Session 1 all subjects performed the same task, in Session 2 they were divided into two groups (perceptual vs motor). The difference between the two groups allowed us to separate the motor and the perceptual information of the sequence previously learnt by the subjects. In this way we could determine whether the perceptual and the motor component had the same or different effects on learning. All the stimuli were displayed in the center of the screen in order to exclude the possible oculomotor aspect of learning. After Session 2, we administered a short questionnaire regarding the participants' possible explicit knowledge about the task (Song et al., 2007b). In keeping with other probabilistic SRT studies (Jiménez, Vaquero, & Lupiáñez, 2006; Nemeth, Janacsek, Londe, et al., 2010; Song et al., 2007b), none of them reported having noticed the sequences.

III/3.4. Data Analysis

Since the core structure of the tasks was the same as in the original ASRT, we followed the same procedures in our analysis (J. H. Howard & Howard, 1997; Nemeth, Janacsek, Londe, et al., 2010). As there is a fixed sequence in the ASRT-Race with alternating random elements (also known as non-adjacent sequence) (Remillard, 2008), for example 2-R-3-R-1-R-4-R, some triplets or runs of three events occur more frequently than others. For instance, following the illustration above, triplets such as 2_3, 3_1, 1_4, 4_2 (where "_" indicates the middle element of the triplet) can occur more often, because the third element (bold numbers) could be derived from the sequence, or could also be a random element. In contrast, triplets such as 4_1, 4_4 would occur infrequently, because in this case the third element could only be random. Following previous studies, we refer to the former as high-frequency triplets and the latter as low-frequency triplets. Because of this difference in frequencies of certain triplets, after observing two stimuli, a certain third stimulus can be expected with 62,5% probability (for example, 223 is five times more probable than 221 or 222 or 224). In our analysis, we determined for every stimulus if it was the more probable or the less probable continuation for the previous trials (see **Fig. III/2**). Participants are faster at the probable stimuli than at the less probable ones, revealing sequence learning in the ASRT paradigm (D. V. Howard et al., 2004; Song et al., 2007b).

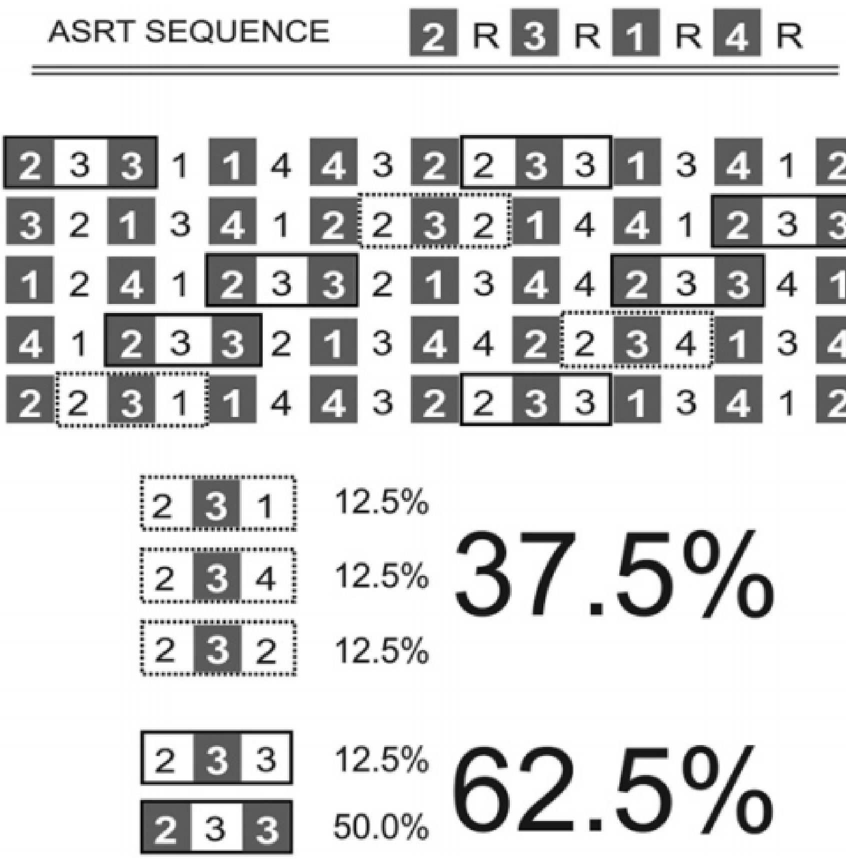


Figure III/2. In a typical ASRT sequence, there are more frequent (high frequency) triplets and less frequent (low-frequency) triplets. In other words, if we know what were the last two elements of the sequence (in this case 2-3-?), there is a 62.5% probability of a certain element as continuation, and only 12.5% probability of all of the other elements.

Similar to prior investigations, two kinds of lowfrequency triplets were excluded from the analysis; trills (e.g., 121, 434) and repetitions (e.g., 111, 222). These triplets are low frequency for all individuals, and people often show pre-existing response tendencies to them. By eliminating these triplets, we can assure that any high versus lowfrequency differences are due to learning, and not preexisting tendencies (D. V. Howard et al., 2004; Nemeth et al., 2009; Nemeth, Janacsek, Londe, et al., 2010).

Since the accuracy of the participants was very high (average over 94.92% in all groups, in all phases), our analysis focused on RT data. For statistical analysis, median RTs were calculated for correct responses only, for each subject for every five blocks, both for the low-frequency and highfrequency elements.

To define the index for Sequence Learning Effect (SLE)(Nemeth & Janacsek, 2011; Nemeth, Janacsek, Balogh, et al., 2010; Song et al., 2007b; Song, Marks, Howard, & Howard, 2009), we calculated the RT difference between the low and high-frequency triplets separately in the Learning Phase (Session 1) and in the Transfer Phase (Session 2) for every five blocks. As we subtracted mean RT of high frequency from low-frequency triplets, SLE was a positive number only if sequence learning occurred, a larger value indicating a stronger effect.

III/4. Results

III/4.1. Learning in Session 1

To be able to investigate the effect of transfer after 12- and 24-h delay, the learning in Session 1 must be similar in the groups. From this point of view, the end of Session 1 is crucial (Nemeth and Janacsek, 2011; Nemeth et al., 2010b; Press et al., 2005; Song et al., 2007). Therefore, we analyzed the SLE of the last five blocks of the Learning Phase for every group. Univariate analysis of variance (ANOVA) was conducted with CONDITION (perceptual vs motor), DAYTIME (morning-first vs evening-first groups) and DELAY (12- and 24-h) as betweensubject factors. ANOVA revealed significant sequence learning, $F(1,94) = 32.31$, $p < 0.001$, which is inferred from the test whether the overall mean is different from zero (Mean SLE = 11.16 msec). There were no other significant main effects or interactions involving CONDITION, DAYTIME and DELAY (all $p > 0.32$), thus these between-subject factors had no significant effect on sequence learning.

III/4.2. Transfer of SLE from Session 1 to Session 2

To determine whether the performance in Session 2 declined, improved, or was constant in relationship to the end of Session 1, we subtracted the SLE-score of the last five blocks of the Learning Phase from the SLE-score of the Transfer Phase (Transfer-SLE). As the groups were similar in SLE at the end of Session 1 (Learning Phase), any difference among groups in Transfer-SLE could be attributed to the differential effects of consolidation. We conducted a univariate ANOVA for this Transfer-SLE-score with CONDITION (perceptual vs motor), DAYTIME (morning-first vs eveningfirst groups) and DELAY (12- and 24-h) as between-subject factors. ANOVA revealed a main effect of CONDITION, $F(1,94) = 4.92$, $p = 0.029$], the motor group showing larger SLE than the perceptual group (**Fig. III/3**). ANOVA showed no significant main effect or interaction with DAYTIME (all $p > 0.45$), suggesting that the AM-PM, PM-AM, AM-AM and PM-PM groups did not differ in their SLE. In addition, main effect and interactions with DELAY were not significant either (all $p > 0.25$), suggesting that 12- and 24-h delay groups performed at a similar level.

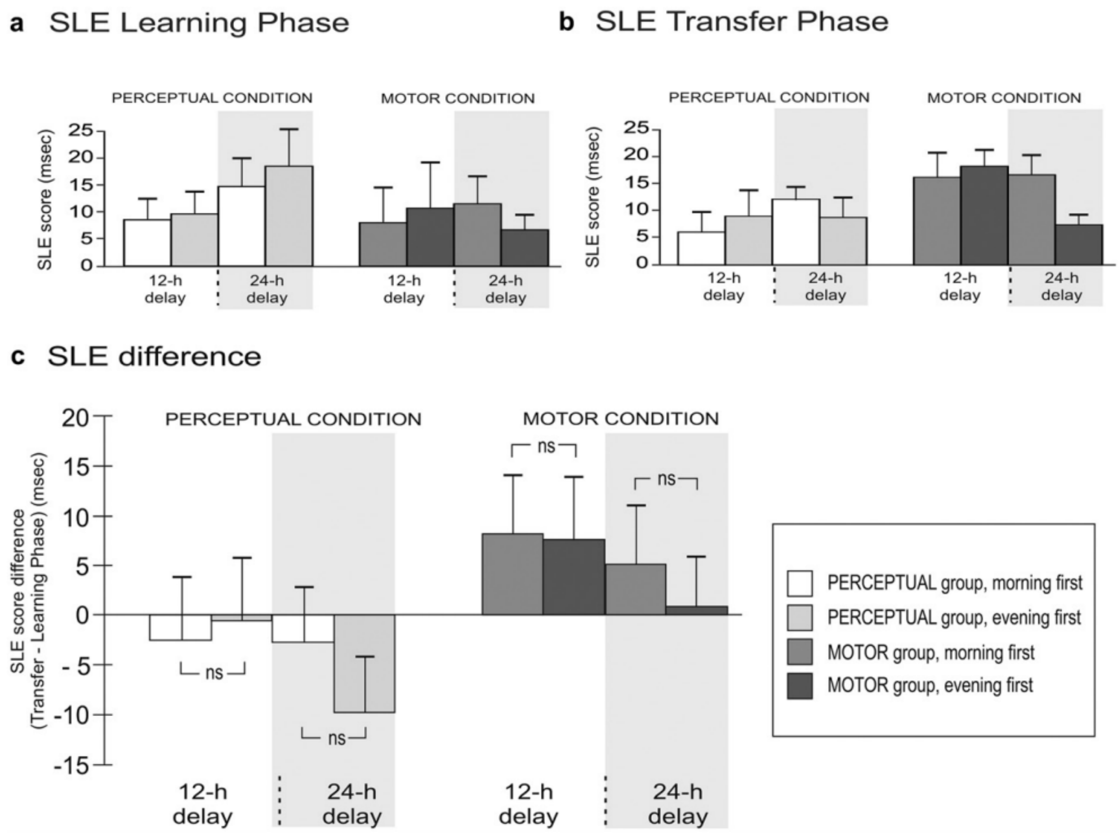


Figure III/3. (a) SLE-score of each experimental group in the last five blocks of the Learning Phase. (b) SLE-score of each experimental group in the Transfer Phase (Session 2). (c) Difference between SLE-scores of the five blocks of Transfer phase and the last five blocks of Learning phase (Transfer-SLE-score). The perceptual groups showed weaker transfer effect than the motor groups both after 12 and 24 h. Error bars indicate Standard Error of Mean.

Thus, the only significant effect in the ANOVA was the main effect of CONDITION, suggesting differential consolidation of perceptual and motor groups with better consolidation for the motor group, irrespective of the delay or daytime. Despite this difference in consolidation, SLE in Session 2 was significantly different from zero for both the perceptual and motor groups (one-sample t-tests for SLEscores:

$t(49) = 5.25$, $p < 0.001$ and $t(51) = 8.72$, $p < 0.001$ respectively). Thus, in spite of the weaker consolidation in the perceptual group, they still showed significant SLE in the Transfer Phase (Session 2). For detailed descriptive statistics see Supplementary Table **ST-III/1** in the Supplementary Materials.

III/4.3. Transfer or new motor learning in the Perceptual Group?

In order to find out whether the significant learning effect in Transfer Phase (Session 2) is due to new motor learning in the perceptual group we investigated the learning effect at the beginning of the Learning Phase (Session 1 - the first two sequence blocks) and learning effect in the Transfer Phase (Session 2 - Block 1-2) separately. We calculated SLE-scores for the first two blocks of Session 1 and Session 2. We submitted these scores to a one-sample t-test separately for Session 1 and Session 2. If we can show a significant learning effect in Session 1 - Block 1-2, the learning is very fast; and the results in Session 2 can be due to new motor learning. However, we found no significant learning effect in Session 1 - Block 1-2 in the perceptual group (one-sample t-test for SLE-score: $t(49) = 1.069$, $p = 0.291$, Mean SLE = 9.27). In contrast we found a significant learning effect in Session 2 - Block 1-2 (onesample t-test for SLE-score: $t(49) = 3.523$, $p = 0.001$, Mean SLE = 8.33). Hence it is likely that the learning effect in Session 2 (Transfer Phase) is attributable to preserved perceptual learning rather than to new motor learning. We found the same pattern in the motor condition (one-sample t-test for SLE-score in Session 1 - Block 1-2: $t(51) = 0.30$, $p = 0.765$, Mean SLE = 3.89; Session 2 - Block 1-2: $t(51) = 5.087$, $p < 0.001$, Mean SLE = 14.77). For detailed descriptive statistics see Supplementary Table **ST-III/2** in the Supplementary Materials.

III/5. Discussion

Our study investigated the role of 12-h and 24-h delay on perceptual and motor components of implicit skill learning, while eliminating oculomotor learning. In this way we connect two debates together: (1) one on the relative importance of perceptual and motor learning (2) the other on the effect of sleep on skill acquisition. We used the same method as Nemeth et al.'s study (Nemeth et al., 2009), except that in our research there was a 12-h (during which participants either had sleep or they were awake) or a 24-h (diurnal) offline period between the Learning and the Transfer Phase. We found significant sequence learning in the Learning Phase. After the 12-h and the 24-h offline period we found significant learning effect in both the perceptual and the motor conditions, however transfer in the motor condition was more effective compared to the perceptual condition. We did not find any sleep-effect on sequence learning in either condition.

The weaker consolidation of perceptual learning is in agreement with the results of Deroost and Soetens (2006) and Willingham (1999), who found no evidence of perceptual learning except for specific conditions. According to previous studies, perceptual learning only takes place when the structure of the sequence is simple, but in case of deterministic sequences with second-order dependencies and probabilistic sequences with first-order dependencies perceptual learning is not or only weakly present (Deroost & Soetens, 2006; Mayr, 1996; Remillard, 2003). Also, previous studies found perceptual learning in explicit conditions (Rüsseler & Rösler, 2000), and when a motor sequence was learnt concurrently (Mayr, 1996). In our study participants had no conscious awareness at all of the structure of the sequence, as the ASRTtask uses probabilistic sequences with second-order dependencies. The only condition that met Deroost and Soetens (2006) criteria is that in the Learning Phase participants learnt the perceptual and motor components concurrently. Compared to Nemeth et al. (2009) who found similar magnitudes of perceptual and motor learning immediately after the Learning Phase, we found a weaker perceptual learning effect in the Transfer phase both after a 12-h and a 24-h delay. Because the only difference was the delay duration, we can suppose that the differences between the results of the two studies can be related to the consolidation period. Thus, this one criterion (i.e., participants in the Learning Phase learnt the perceptual and motor components concurrently) can be enough for finding significant perceptual learning immediately after the Learning Phase (Meier & Cock, 2010; Nemeth et al., 2009; Weiermann, Cock, & Meier, 2010), however, it might result in weaker consolidation after the delay period. To put the puzzle together, based on the present study we can propose that the consolidation period has a differential effect on motor and perceptual components of learning, such that in the Transfer Phase the motor learning effect is larger than the perceptual one.

Song et al. (2008); Nemeth et al. (2009) and the present study are similar in the nature of the sequence structure and the implicitness of the task. Furthermore, the present study and the study of Nemeth et al. (2009) also eliminated the possibility of oculomotor learning. Because we focused only on the perceptual and motor learning while controlling for the oculomotor learning, the role of response-based learning and effector-based learning remained unclear (Cohen et al., 1990; Remillard, 2003; Willingham, 1999); therefore the exact nature of the underlying mechanism still needs to be investigated.

In addition to the question of perceptual and motor components of learning, our study has relevance for the sleep debate in skill consolidation (Debas et al., 2010; Doyon, Korman, et al., 2009; Gerván & Kovács, 2007; Karni, Tanne, Rubenstein, Askenasy, & Sagi, 1994; Edwin M. Robertson, 2009; Song, 2009; Stickgold & Walker, 2005; M. P. Walker, Brakefield, Morgan, Hobson, & Stickgold, 2002). As pointed out by Robertson (2009) and supported by Song et al. (2007b) and Nemeth et al. (2010), we found that sleep does not support sequence learning. In addition, sleep has no different role in the consolidation of motor and perceptual factors of implicit sequence learning. A plausible explanation can be that in the probabilistic sequence learning task used in this study, besides primary sensory and motor brain regions, sub-cortical structures and cerebellum are more involved (Doyon, 2008; Hikosaka et al., 1999, 2002), opposed to the more basic finger tapping tasks where sleep-dependent improvement was usually found (M. P. Walker et al., 2002).

To conclude, despite the 12-h or the 24-h offline period we found a significant perceptual and motor learning effect in the Transfer Phase, however the transfer of motor knowledge was more robust, irrespective of whether sleep occurred in the consolidation period or not. These results have important implications for the perceptual/motor and also for the sleep debate in skill learning in the following ways: (1) Previous experiments in this field included only one session which can reveal short-term performance changes in behavior. Consequently, it is important to use more sessions with many hours (even a day) delay between sessions for measuring permanent changes in neural plasticity. (2) Sleep has no contribution to this type of learning. However, further investigations need to explore more deeply conditions (including nature of sequence, awareness, perceptual/motor learning) in which sleep has a significant role in skill learning. (3) The retention period itself (regardless of sleep) has a modifying effect on the consolidation of perceptual/motor knowledge and the underlying brain networks.

III/6. Acknowledgements

Thanks to our mentors: Darlene V. Howard and James H. Howard, Jr. from Georgetown University. This research was supported by Bolyai Scholarship Program (D. N.) and OTKA K 82068. Thanks to Ágnes Szokolszky and Szabolcs Kéri helping us with the final version of the manuscript.

IV. EXPLICIT INSTRUCTIONS AND CONSOLIDATION PROMOTE REWIRING OF AUTOMATIC BEHAVIORS IN THE HUMAN MIND

(Study 3)[8]

IV/1. Abstract

One major challenge in human behavior and brain sciences is to understand how we can rewire already existing perceptual, motor, cognitive, and social skills or habits. Here we aimed to characterize one aspect of rewiring, namely, how we can update our knowledge of sequential/statistical regularities when they change. The dynamics of rewiring was explored from learning to consolidation using a unique experimental design which is suitable to capture the effect of implicit and explicit processing and the proactive and retroactive interference. Our results indicate that humans can rewire their knowledge of such regularities incidentally, and consolidation has a critical role in this process. Moreover, old and new knowledge can coexist, leading to effective adaptivity of the human mind in the changing environment, although the execution of the recently acquired knowledge may be more fluent than the execution of the previously learned one. These findings can contribute to a better understanding of the cognitive processes underlying behavior change, and can provide insights into how we can boost behavior change in various contexts, such as sports, educational settings or psychotherapy.

[8] **Szegedi-Hallgató, E.,** Janacsek, K., Vékony, T., Tasi, L. A., Kerepes, L., Hompoth, E. A., … Németh, D. (2017). Explicit instructions and consolidation promote rewiring of automatic behaviors in the human mind. *Scientific Reports*, *7*(1), 1–7. https://doi.org/10.1038/s41598-017-04500-3

IV/2. Introduction

French drivers face a real challenge when they have to drive in England for the first time. They might look in the wrong direction when checking the traffic, and incorrectly assume that there is no other vehicle, so they are free to go ahead at a crossroad. Their already well-developed perceptual-motor skill of driving becomes ineffectual or even harmful by leading them to false predictions in the new environment. Our study can contribute to a better understanding of how we can rewire our perceptual-motor skills in such situations. Can we adapt to the changed regularities of the environment without any external help, purely implicitly?

The aim of our study was to test, in a controlled experimental setting, how we can update – *rewire* – our knowledge of sequential/statistical regularities that thought to be an essential aspect of many everyday skills, such as playing a musical instrument or video games (Bergstrom, Howard, & Howard, 2012), learning/processing languages (Kaufman et al., 2010; Nemeth et al., 2011), and social skills (Heerey & Velani, 2010; Norman & Price, 2012). Modification of such knowledge can be empirically tested by teaching two differing sets of regularities (i.e., sequences) to participants. Characterizing the interference between the first-learned sequence and the newly encountered one is the key to understand the rewiring process. Retroactive interference can hinder our ability to activate old sequence knowledge once new reguralities have been learned (Dorfberger, Adi-Japha, & Karni, 2007; Goedert & Willingham, 2002; Yotsumoto, Watanabe, Chang, & Sasaki, 2013; M. P. Walker, Brakefield, Hobson, & Stickgold, 2003; Handa, Rhee, & Wright, 2016). Likewise, our previously developed automatisms can make it more difficult to learn new regularities, termed as proactive interference. Proactive interference effects were found even after limited practice when sequence A and sequence B were highly dissimilar (Goedert & Willingham, 2002); and it was also detected by contrasting performance on those chunks of movements that were common to both sequences with those that differed (Verneau, van der Kamp, Savelsbergh, & de Looze, 2015). While retroactive interference can be beneficial for overwriting old knowledge of sequential regularities, proactive interference works against successful rewiring.

To date no study has focused on the entire process of rewiring in humans that captured both retroactive and proactive interference effects, and at the same time assessed the impact of explicit processes in one experimental design. Here we present a study with such a design to tackle the question whether we can overcome the negative consequences of the interference without conscious effort (i.e., implicitly) or the awareness about the changed reguralities (i.e., explicitness) is essential for rewiring. Eighty-four healthy young adults performed a four-choice reaction time task (**Fig. IV/1/a**) on three consecutive days. The presentation order of the stimuli followed a probabilistic sequence on the first day (Sequence A in the Learning Phase), then this sequence changed to a different one on the second day (Sequence B in the Rewiring Phase). The two sequences shared some of their transitional probabilities (for details see Methods and **Fig. IV/1/b**), meaning that at some points in Sequence B the most probable upcoming stimulus was the same as in Sequence A (*unchanged* sequence parts). Other transitional probabilities changed: the most probable continuation of the previous trials was different from that on the previous day (*changed* sequence parts). This way we could compare learning with and without interference from the previous day (more details in Supplementary Methods **SM-IV/1.2** in the Supplementary Materials). In addition, to assess the impact of explicit information processing, sequences were learned either explicitly or implicitly: implicit learners were not aware of the sequence structures, while explicit learners were provided with cues that could be used on half of the trials (**Fig. IV/1/c**). We hypothesized that – if explicitness does help rewiring – it should be expected that explicit learners perform better even on those trials on which they did not have any explicit advantage over the implicit learners. Three groups of participants were compared: the Implicit-Implicit group learned both sequences implicitly (thus rewiring was also implicit); the Implicit-Explicit group learned the first sequence implicitly and the second sequence explicitly; the Explicit-Explicit group learned both sequences explicitly (**Fig. IV/1/d**).

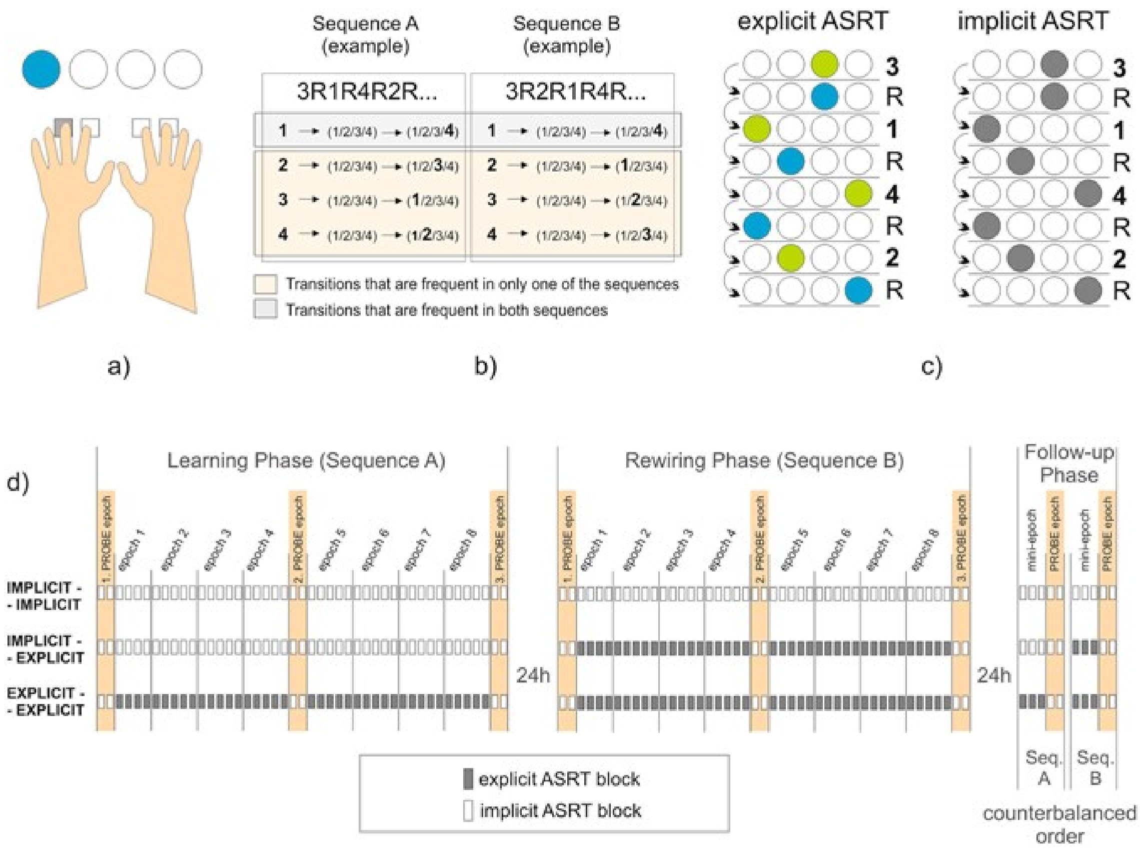


Figure IV/1

Methods and procedure. (**a**) ASRT task: participants were asked to respond to stimuli appearing on one of four locations, and press the corresponding key as fast and as accurately as they could. (**b**) Examples of sequences and their shared/differing transitions. Numbers indicate the locations of stimuli (1, 2, 3 or 4). The notation "R" indicates a random location out of the four possible ones. Knowing what stimulus appeared two trials before enables participants to anticipate the next stimulus as there is always a highly probable continuation (highlighted) and three less probable continuations. In Sequence A, for example, two trials after encountering a stimulus on the 2nd location, the most probable stimulus is one on the 3rd location (all the remaining possibilities being equally less likely). Sequence A and Sequence B share some of their transitional probabilities (e.g. encountering stimulus on the 4th location two trials after encountering one on the 1st location); and they differ on others (e.g. encountering stimulus on the 3rd location predicts a stimulus on the 2nd location two trials later in Sequence B, but not in Sequence A). Interference effects could be detected by contrasting the magnitude of learning of changed and unchanged sequence parts. (**c**) In the implicit version of the task, random and pattern trials appeared in the same color, thus they were indistinguishable to participants. In the explicit version of the task, random and pattern trials were of different colors, and participants were asked to keep track of the repeating pattern. (**d**) There were 85 trials in a block. There were 45 blocks in the Learning and Rewiring Phase, collapsed into bigger sections (epochs) for analysis. Probe epochs were implicit for all participants. Here we focus on experimental epochs (probe epochs led to similar results, for details see Supplementary Results **SR-IV/2.3** and **SR-IV/2.4** in the Supplementary Materials). In the Follow-up Phase, we aimed to test participants' performance on both sequences after equal amount of practice on each, without introducing much relearning. Therefore only 5 + 5 blocks of both sequences were presented on the third day. Two of these blocks were probe blocks.

IV/3. Methods

IV/3.1. Participants

Eighty-four healthy young adults took part in the experiment. Participants were recruited at University of Szeged and were randomly assigned to one of three groups: the Implicit-Implicit group ($N = 28$; 20 females; Age: $M = 20.46$ years, $SD = 2.10$), the Implicit-Explicit group ($N = 28$; 17 females; Age: $M = 22.14$ years, $SD = 1.96$), and the Explicit-Explicit group ($N = 27$; 18 females; Age: $M = 22.54$ years, $SD = 3.33$). One participant was excluded from the analysis because errorneously the same sequence was administered to him on each day of the study. Three groups did not differ in their scores on standard working memory and executive function tests (Digit Span: $p = 0.443$, $\eta_p^{\ 2} = 0.021$; Counting Span: $p = 0.440$, $\eta_p^{\ 2} = 0.022$; perseveration rates on the Wisconsin Card Sorting Task: $p = 0.710$, $\eta_p^{\ 2} = 0.010$; Stroop Test: $p = 0.578$, $\eta_p^{\ 2} = 0.015$). Participants did not suffer from any psychiatric or neurological disorders. None of the participants were aware of the purpose of the experiment. Prior to their inclusion in the study, participants provided informed consent to the procedure as approved by the research ethics committee of University of Szeged, Szeged, Hungary. The study was conducted in accordance with the Declaration of Helsinki and participants received course credits for taking part in the experiment.

IV/3.2. Task and Procedure

Participants performed a modified version of the Alternating Serial Reaction Time (ASRT) task (J. H. Howard & Howard, 1997). The program was coded in Psychopy (Peirce, 2007). In the *Implicit variant* of the task, four light grey circles were arranged horizontally on the screen. Four buttons on the keyboard corresponded to the four locations on the screen: Z, C, B, and M, respectively. Intervening buttons were removed to minimize false buttonpresses. Participants used their left and right middle and index fingers to respond to the targets. The stimulus stayed on the screen until a correct buttonpress was made (but it remained on the screen after an unapt button was pressed). Response to stimulus interval (RSI) was set to 120 ms. One block of trials consisted of five random (preparatory) trials followed by ten repetitions of a probabilistic sequence (10×8 trials). After each block, there was a pause of at least 10 seconds (terminated by participants) during which the average reaction time of correct buttonpresses and the percentage of erroneous buttonpresses were displayed.

Unbeknownst to the participants, the ASRT sequence consisted of a four-elements-long sequential pattern (e.g., 3–1–4–2) intersparsed with random elements (3–R–1–R–4–R–2–R). Because of the alternating pattern and random trials, the ASRT sequence can be regarded as a second order probabilistic sequence, meaning that the identity of the upcoming stimulus can be anticipated based on the *n-2* trial (one of the four possibilities is always more probable than the remaining three), regardless of the stimulus being a random or a pattern element (for more details see Supplementary Methods **SM-IV/1.1** in the Supplementary Materials).

In the *Explicit variants*, participants performed a cued version of the ASRT task (Nemeth, Janacsek, & Fiser, 2013; Song et al., 2007a, 2009). This version differed from the implicit task in three respects: firstly, random and pattern stimuli appeared in different colors (pattern elements appeared green while random elements appeared blue). Secondly, participants were instructed to pay attention primarily to the four-element long pattern (the green trials) to be able to report it after each block. Also they were told to constantly monitor this pattern and report if it changed during the course of a block (actually it never changed within a given block). Finally, the feedback after the blocks did not contain information about RTs and erroneous buttonpresses on random trials as the instruction highlighted performance on the pattern trials.

After learning an ASRT sequence (referred to as Sequence A) on the first day of the study (in the *Learning Phase*), participants were given a different ASRT sequence (referred to as Sequence B) on the second day (in the *Rewiring Phase*). Twelve different sequence combinations were used in the experiment in a counterbalanced order. All these combinations are presented in Supplementary Methods **SM-IV/1.3** in the Supplementary Materials. Three groups were compared based on the instructions they received each day: The *Implicit-Implicit* group performed the implicit version of the task in both the *Learning Phase* and in the *Rewiring Phase*; the *Implicit-Explicit*group performed the implicit version in the *Learning Phase* and the explicit version in the *Rewiring Phase;*finally, the *Explicit-Explicit* group performed the explicit version in both phases.

On the third day of the study – in the *Follow-up Phase* – the magnitude of statistical knowledge for both sequences was assessed to investigate possible retroactive interference effects after a 24-hour consolidation period. Then the *Triplet Sorting*

Task and the *Free Generation Task* were administered to assess the amount of explicit knowledge participants gained about both sequences (see descriptions of the tasks in Supplementary Methods **SM-IV/1.6** in the Supplementary Materials). The analysis of these tasks showed that participants indeed gained more explicit knowledge about the regularities when they performed the explicit version of the task, and that the explicit cues indeed can help differentiate between the two sequences and use the acquired knowledge more appropriately in the relevant context (for details see Supplementary Results **SR-IV/2.5** in the Supplementary Materials).

Finally, working memory and executive functions were also assessed on the third day. We wanted to ensure that the three groups had similar performance on these general cognitive functions, and thus the obtained results of rewiring could not be attributed to differences in these functions (see also the Participants section). The experiment took place in a quiet laboratory at University of Szeged (one participant a time). The whole procedure lasted approximately 60 minutes on the first and second day of the experiment, and an additional 40 minutes on the third day.

IV/4. Results

We measured statistical learning as (**a**) difference in reaction times (RTs) given to anticipated (probable) stimuli in contrast to unexpected (less probable) stimuli, termed as Statistical Learning Effect (SLE), and (**b**) by determining whether erroneous responses reflect anticipations of the most probable stimuli in cases when less probable trials came up (more details in Supplementary Methods **SM-IV/1.4** and **SM-IV/1.5** in the Supplementary Materials). Statistical learning measured by the SLE score was evident in both the Learning and Rewiring Phase (see 95% confidence intervals, CIs, on **Fig. IV/2/a**). In the Learning Phase, there could not possibly be any interference effects as only Sequence A had been introduced yet, we nevertheless contrasted the magnitude of learning of those transitional probabilities that were common during both Phases (unchanged sequence parts) and those that were about to change in the Rewiring Phase (changed sequence parts). As expected, there was no difference between the two ($p = 0.568$, Cohen's $d = 0.080$), indicating that they were equally easy to learn (**Fig. IV/2/a**, light vs. dark grey bars). In the Rewiring Phase, however, we found smaller statistical learning for the changed sequence parts ($p < 0.001, d = 0.829$) compared to the unchanged sequence parts (**Fig. IV/2/a**, light grey vs. blue bars). This was apparent in the Implicit-Implicit ($p < 0.001$, $d = 1.425$) and Explicit-Explicit groups ($p = 0.008$, $d = 0.737$), but not in the Implicit-Explicit group ($p = 0.128$, $d = 0.406$). These results suggest that the Implicit-Explicit group was the most successful in adapting to the new statistical regularities, even to the extent that their rewired knowledge was not much different than that of the unchanged sequence parts. Interference effects were most clearly shown by contrasting the magnitude of learning of the same transitional types (changed vs. unchanged) over the two Phases. As expected, learning of the unchanged sequence parts did not decline in the Rewiring Phase ($p = 0.151$, $d = 0.021$, **Fig. IV/2/a**, light grey bars in the Learning vs. Rewiring Phase), while learning of the changed sequence parts was significantly lower in Rewiring Phase than the original learning in the Learning Phase ($p < 0.001$, $d = 0.729$, **Fig. IV/2/a** dark grey vs. blue bars). This pattern was apparent in the Implicit-Implicit ($p < 0.001$, $d = 1.562$) and Explicit-Explicit ($p = 0.028$, $d = 0.850$) groups, but not in the Im*p*licit-Explicit group ($p = 0.561$, $d = 0.155$) – indicating, again, that the Implicit-Explicit group was the most successful in adapting to the changes in the sequence structure. The least successful group was the Implicit-Implicit group, evidenced by their statistical knowledge of the changed sequence parts in the Rewiring Phase (**Fig. IV/2/a**, blue bars) being significantly smaller than both the Explicit-Explicit ($p = 0.028$, $d = 0.743$) and Implicit-Explicit groups' ($p < 0.001$, $d = 1.198$). Additional analysis of the time course of rewiring revealed that the Implicit-Implicit group showed significant learning of the changed transitional probabilities only in the second half of the Rewiring Phase, suggesting slower updating of the previously acquired statistical knowledge (i.e., larger proactive interference) compared to the other two groups (for more details see Supplementary Results **SR-IV/2.1.1** in the Supplementary Materials).

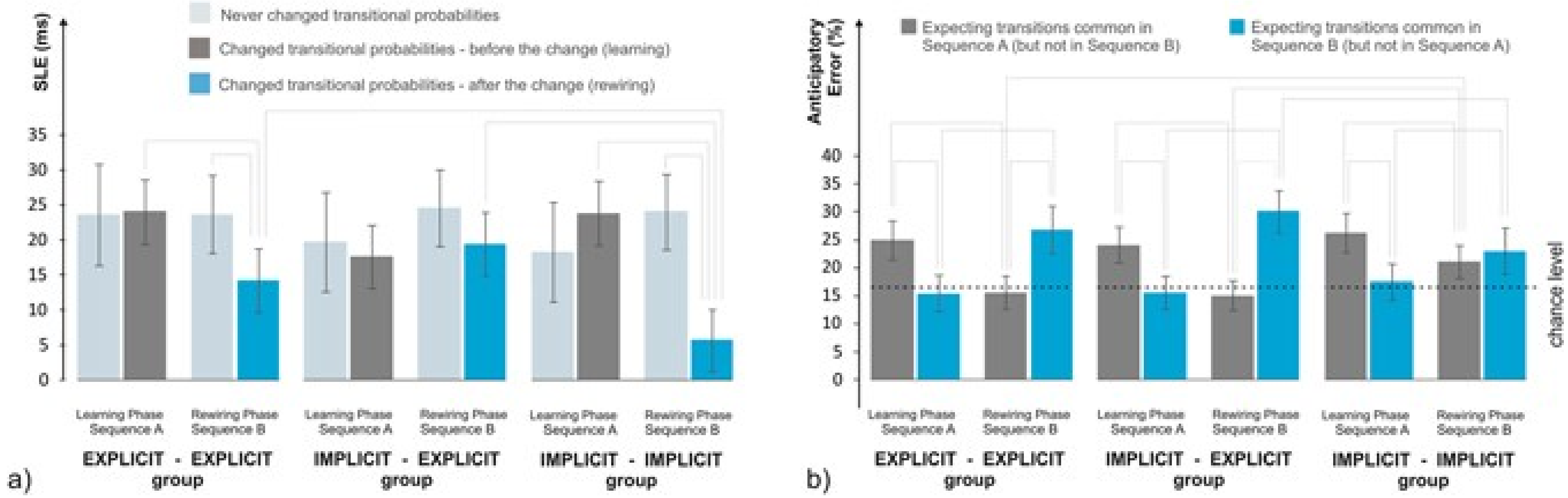


Figure IV/2. Learning and Rewiring. (**a**) Statistical learning effect (SLE) in the Learning and Rewiring Phase. The magnitude of SLE indicates the difference of reaction times (RTs) given to frequent transitions (more probable stimuli) in contrast to infrequent transitions (less probable stimuli). Some of the transitions had constant frequency in the Learning Phase and Rewiring Phase (unchanged transitions, light grey bars), while other transitions swapped their frequency – previously infrequent transitions became frequent in the Rewiring Phase, and vice versa (changed transitions, dark grey bars – before the change, blue bars – after the change occurred). Adapting to the changed statistical structure in the Rewiring Phase was shown to be more difficult than learning the contingencies in the first place in the Learning Phase. This was shown by SLEs being – on average – smaller for the changed transitions after the change in frequencies took place in the Rewiring Phase (blue bars) than before the change (dark grey bars). The Implicit-Explicit group did not show signs of such difficulty. (**b**) When a less probable stimulus came up, participants sometimes erroneously pressed the key corresponding to the most probable stimulus, termed as anticipatory errors. As two (partly) different sequences were taught, we differentiated between anticipations of Sequence A's most probable stimuli, and that of Sequence B's. Percentage of anticipatory errors of Sequence A (learned in the Learning Phase, grey bars) and Sequence B (learned in the Rewiring Phase, blue bars) over the two Phases, and chance level for anticipatory errors (dotted line) are shown. Each group showed adaptation to the current sequence, as anticipations for Sequence A were above chance level in the Learning Phase, while anticipations of Sequence B were above chance level in the Rewiring Phase. The Implicit-Implicit group additionaly showed above chance level anticipations of Sequence A during the Rewiring Phase, indicating the continuing influence of their knowledge gained in the Learning Phase. The solid lines connecting the bars indicate significant differences ($p < 0.05$). Error bars represent 95% confidence intervals (CIs).

On some of the trials, participants pressed a key that did not correspond to the stimulus. Some of these errors reflected anticipations of the most probable stimulus when the actual stimulus was a less probable one, termed as *anticipatory errors*. As two sequences were taught, we can measure anticipations of Sequence A's most probable transitions and those of Sequence B's (errors that could be regarded as anticipations of both sequences were not analysed). We compared the proportion of anticipatory errors to each other (anticipations of Sequence A vs. anticipations of Sequence B), and to a baseline proportion that could be expected by chance (16.67%, see **Fig. IV/2/b**). As expected, the Learning Phase was dominated by anticipations of Sequence A (dark grey bars on **Fig. IV/2/b**), while the Rewiring phase was dominated by anticipations of Sequence B (both $p < 0.001$, both $d > 1.061$, blue bars on **Fig. IV/2/b**). From another point of view, there were less anticipations of Sequence B in the Learning Phase than in the Rewiring Phase, and vice versa (both $p < 0.001$, both $d > 0.979$). This pattern of results was observed in all groups, although effect sizes were substantially smaller in the the Implicit-Implicit group (both $d < 0.672$) than in the other groups (all $d > 1.226$). Most importantly, anticipations of Sequence B in the Rewiring Phase (that indicate adaptation to the new sequence structure) were less pronounced in the Implicit-Implicit group than in the Implicit-Explicit group ($p = 0.047$, $d = 0.721$), while anticipations of Sequence A in the same Phase (indicating the continuing influence of the knowledge gained in the Learning Phase) were more pronounced in the Implicit-Implicit group than in the Implicit-Explicit and the Explicit-Explicit groups (both $p < 0.036$, both $d > 0.795$). The Implicit-Implicit group showed no significant difference in proportions of anticipating Sequence A and Sequence B during the Rewiring Phase ($p = 0.529$, $d = 0.225$), both being above chance level (see 95% CIs on **Fig. IV/2/b**). These results clearly point to the continuing influence of the no-longer valid statistical knowledge gained in the Learning Phase – that is, proactive interference – in the Implicit-Implicit group (see also Supplementary Results **SR-IV/2.1.2** in the Supplementary Materials). The remaining two groups who performed the Rewiring Phase explicitly showed no signs of such interference, indicating a beneficial effect of awareness about the sequence structures.

Participants were retested on the third day (in the Follow-up Phase, **Fig. IV/1/d**) for both sequences to test whether the first one became overwritten by the second one. Participants showed better performance on the transitions that were frequent in both the

Learning and Rewiring Phases (**Fig. IV/3/a**, light grey bars) than on those that were frequent in only one of the Phases (**Fig. IV/3/a**, dark grey and blue bars; $p = 0.003$, $d = 0.506$), which is not surprising given that the former ones were practiced almost twice as much. More importantly, better performance was expressed for Sequence B than for Sequence A ($p = 0.015$, $d = 0.404$). This pattern of results may indicate that statistical knowledge for Sequence A became partly overwritten by knowledge of Sequence B, showing retroactive interference which is beneficial for the rewiring process. No group differences were observed (ps > 0.303; see also Supplementary Results **SR-IV/2.2.1** in the Supplementary Materials).

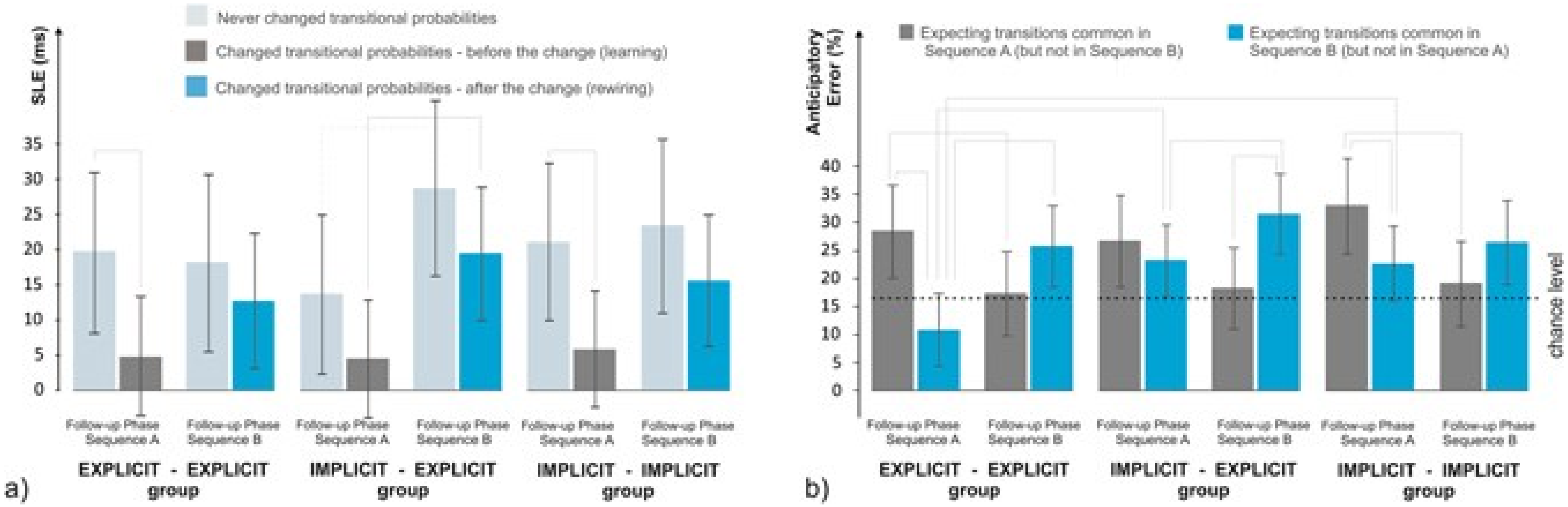


Figure IV/3. Testing the efficiency of the rewiring process after a 24-hour consolidation period, in the Follow-up Phase. (**a**) Overall, SLEs for the changed sequence parts (dark grey and blue bars) were smaller than that of the unchanged sequence parts (light grey bars). In the case of the changed sequence parts, statistical structure that corresponded to Sequence B (that was learned in the Rewiring Phase) was retained better than statistical structure that corresponded to Sequence A (that was learned in the Learning Phase) – the latter not reliably differing from zero (as shown by 95% CIs). This pattern indicates adaptation to the changed statistical regularities of Sequence B taking place in the Rewiring Phase, and no observable proactive interference of Sequence A, thus, overall successful rewiring. No group differences were observed. (**b**) Chance level for anticipatory errors are shown by the dotted line. Each group showed adaptation to the current sequence as anticipations for Sequence A were above chance level when performing Sequence A in the Follow-up Phase; while anticipations for Sequence B were above chance level when performing Sequence B. This pattern suggests that the old and new knowledge, acquired on the first and second day of the experiment, coexisted after a 24-hour delay period, and were accessible when required. The solid lines connecting the bars indicate significant differences ($p < 0.05$). Dotted lines indicate trend level differences ($p < 0.10$). Error bars represent 95% CIs.

When performing Sequence A on the third day of the study, anticipations of Sequence A were more common than anticipations of Sequence B ($p = 0.004$, $d = 0.533$), and than what might have been expected by chance (**Fig. IV/3/b**, dark grey bars). When performing Sequence B, on the other hand, anticipations of Sequence B outnumbered anticipations of Sequence A ($p = 0.009$, $d = 0.503$), and were more numerous than expected by chance (**Fig. IV/3/b**, blue bars). From another point of view, anticipations of Sequence A were significantly more pronounced when performing Sequence A than when performing Sequence B, and vice versa (both $p < 0.003$, $d > 0.494$). This pattern of results indicate no proactive or retroactive interference effects, as participants were able to quickly adapt to changes in the statistical structure, and suggests that knowledge about the two statistical structures coexist and can be adaptively used in the appropriate situation. No group differences were observed ($p = 0.745$; see also Supplementary Results **SR-IV/2.2.2** in the Supplementary Materials).

In our study, we compared performance of groups that learned/rewired their knowledge with vs. without explicit cues. The Free Generation and the Triplet Sorting Tasks were used to explore to what extent their knowledge remained implicit or became explicit (see Supplementary Methods **SM-IV/1.6.1** and **SM-IV/1.6.2** in the Supplementary Materials). In the Free Generation Task, they were asked to generate alternating sequences similar/dissimilar to the ones they encountered during the experiment. The results revealed clear group differences with participants in the explicit conditions exhibiting better performance, as a result of the explicit cues and instructions. Participants who learned/rewired their knowledge without explicit cues (i.e., implicitly) might have also gained some explicit knowledge about the regularities as well, although we cannot rule out the possibility that they used different strategies during task compared to the explicit group, and those strategies resulted in a somewhat similar performance in the end (for details see Supplementary Results **SR-IV/2.5.1** in Supplemetary Materials). The Triplet Sorting Task more directly tested their knowledge about the same statistical structures (triplets) that also provided the basis of the RT and anticipatory error analyses. The results of this task support the interpretation that knowledge of these statistical structures remained implicit for the implicit learners (Supplementary Results **SR-IV/2.5.2** in the Supplementary Materials). This interpretation is also in line with previous ASRT studies showing that participants remain unaware of the stimulus structure if it is not explicitly cued (Song et al., 2007b), and even after extended practice (e.g., ten days, D. V. Howard et al., 2004). Nevertheless, we can never totally exclude the possibility that some degree of explicit knowledge developed even if the sequence structure was not cued.

IV/5. Discussion

In summary, we found successful rewiring of the acquired knowledge in all three experimental groups. In the Rewiring Phase the group that learned implicitly and rewired with the help of explicit cues (i.e., the Implicit-Explicit group) showed better performance than the other groups. In other words, explicit cues during the rewiring process led to faster adaptation to the changed reguralities, evidenced both in the cued part of the task (experimental epochs) as well as in the uncued part (probe epochs, see Supplementary Results **SR-IV/2.3** in the Supplementary Materials). By the end of the rewiring period, all groups showed similar performance suggesting an efficient but slower rewiring in the Implicit-Implicit group as well. We also found evidence that the first learned sequence was accessible when needed, shown by sequence specific anticipatory errors in the Follow-up Phase, although the motor execution of it was not as fluent as the execution of the secondly learned sequence.

The aim of our study was to test, in a controlled experimental setting, how we can update – *rewire* – our knowledge of sequential/statistical regularities that thought to be an essential aspect of many everyday skills, such as playing a musical instrument or video games (Bergstrom et al., 2012), learning/processing languages (Kaufman et al., 2010; Nemeth et al., 2011), and social skills (Heerey & Velani, 2010; Norman & Price, 2012). Nevertheless, these skills are far more complex than the task employed in the current study, and they may encompass other types of sequential/statistical knowledge, have longer/more variable "practice schedule", and may combine different regularities coming from different domains (such perceptual and motor). Although in our study participants were presented with regularities both in the perceptual and motor domains (correlated visual and motor stimulus streams), here we did not aim to separate the knowledge acquired in these two domains. Previous ASRT studies have showed that participants acquire both the perceptual and motor knowledge and they also retain their knowledge after a delay period (Nemeth et al., 2009; Hallgató, Győri-Dani, Pekár, Janacsek, & Nemeth, 2013; Song et al., 2008). Future studies need to directly test how various types and complexity of sequential/statistical knowledge can be updated when the underlying regularities change, whether and how sequence complexity interacts with the explicit advantage observed in our study, and whether and how these processes differ across domains.

Based on our findings, rewiring of a relatively simple statistical knowledge – that we tested in the current study – show a complex picture: proactive interference, which works against the adaptation to the changed regularities, is stronger when learning and rewiring is implicit, while explicit cues about these changed regularities can help speed up the rewiring process. After a 24-hour delay period, proactive interference is volumed down, while retroactive interference is volumed up, suggesting that consolidation of the updated knowledge about the changed regularities has a critical role in successful rewiring. The fact that both the old and new, updated knowledge seems to remain accessible highlights the adaptive nature of the human mind, making it possible to dynamically use the appropriate procedures corresponding to various environments. Our findings can contribute to a better understanding of the cognitive processes underlying behavior change.

IV/6. Author Contributions

E.Sz.-H. anayzed data and wrote the manusript; K.J. designed the study, supervised data acquisition, analyzed data, and wrote the manuscript; D.N. designed the study, supervised data acquisition, wrote the manuscript; T.V. participated in data acquisition, data management and analysis; L.A.T. participated in data acquisition, data management and analysis, L.K. participated in data acquisition, data management and analysis; E.A.H. participated in data acquisition, data management and analysis; A.B. participated in data acquisition, data management and analysis.

IV/7. Competing Interests

The authors declare that they have no competing interests.

V. DIFFERENT LEVELS OF STATISTICAL LEARNING – HIDDEN POTENTIALS OF SEQUENCE LEARNING TASKS

(Study 4)[9]

V/1. Abstract

In this paper, we reexamined the typical analysis methods of a visuomotor sequence learning task, namely the ASRT task (J. H. Howard & Howard, 1997). We pointed out that the current analysis of data could be improved by paying more attention to pre-existing biases (i.e. by eliminating artifacts by using new filters) and by introducing a new data grouping that is more in line with the task's inherent statistical structure. These suggestions result in more types of learning scores that can be quantified and also in purer measures. Importantly, the filtering method proposed in this paper also results in higher individual variability, possibly indicating that it had been masked previously with the usual methods. The implications of our findings relate to other sequence learning tasks as well, and opens up opportunities to study different types of implicit learning phenomena.

Keywords: ASRT task, sequence learning, statistical learning, analysis methods, data filtering, individual differences, types of learning.

---

9 **Szegedi-Hallgató, E.**, Janacsek, K., & Nemeth, D. (2019). Different levels of statistical learning—Hidden potentials of sequence learning tasks. *PloS One*, *14*(9), e0221966. https://doi.org/10.1371/journal.pone.0221966

V/2. Introduction

When previous experiences facilitate performance even though the current task does not require conscious or intentional recollection of those experiences, implicit memory is revealed (Schacter, 1987). The Serial Reaction Time (SRT) task (Nissen & Bullemer, 1987) is a commonly used task measuring implicit learning and memory in the visuomotor domain; people are instructed to respond to a sequence of stimuli by pressing a corresponding button (usually having a 1:1 stimulus-response mapping), and even though they are not aware that the same pattern of successive trials is repeated over and over again, they nevertheless show improvement compared to their reactions to random (or pseudorandom) streams of stimuli. A drawback of the design is that learning can only be assessed at certain points (via inserting blocks of random stimuli), and that, due to the simplicity of the SRT sequences, people may become aware of them after all, in which case explicit memory is being measured instead of or in addition to implicit learning(D. V. Howard & Howard, 1992).

A modified version of the task, namely the Alternating Serial Reaction Time task (ASRT), has been introduced twenty years ago as a possible solution to the aforementioned problems (J. H. Howard & Howard, 1997); and it turned out that even the test-retest reliability is better using this variant (Stark-Inbar, Raza, Taylor, & Ivry, 2016). At the time of writing these lines, that article introducing the ASRT task had been cited three hundred times (317, to be precise), and *Google Scholar* has about 200 results for the expression „alternating serial reaction time", out of which 87 has been published since 2015. Clearly, the task gained popularity as a research tool recently, which is not surprising given its advantages over the classical SRT. Having a lot of experience with it ourselves, we began to feel there is even more to it than currently recognized. In the present paper, we aim to discuss the potential challenges of its currently used analysis methods and to provide some ideas about how to overcome these flaws. Importantly, most of the concerns (and solutions) discussed in this paper are directly applicable to other sequence learning tasks as well.

V/2.1. About the ASRT Task

In the ASRT – to make the predetermined sequence less apparent - a four element long pattern (e.g. 1-4-2-3) is intervened by random elements (i.e. 1-R-4-R-2-R-

3-R). Participants generally don't recognize the pattern (or the fact that there is a pattern), and still react to pattern trials faster and more accurately than to random trials (referred to as *pattern learning*). Moreover, the relative advantage of pattern trials can be assessed at any point of learning, or continuously throughout learning.

At first, the typical result may seem like evidence that people are capable of somehow detecting the pattern – and thus being able to respond to these elements more efficiently – even though pattern trials are hidden between random elements), but this is not necessarily the case. As a consequence of the alternation of pattern and random trials, some stimulus combinations are more frequent than others and some trials are more predictable than others, possibly leading to faster and more accurate responses to them (referred to as statistical learning). When assessing stimulus combinations of at least three consecutive trials, the variability of such combinations depends on the number of random elements they contain. For example, random-ending triplets (three consecutive trials, i.e. R-P-R) are four times as variable as pattern-ending triplets (i.e. P-R-P), since they contain two random elements instead of one. Accordingly, particular R-P-R combinations occur with a much lower frequency than particular P-R-P combinations. Moreover, some of the R-P-R triplets mimic P-R-P triplets (e.g. 1-2-2 can occur as both an R-P-R and a P-R-P triplet) further increasing the frequency of those instances, and further increasing the difference between the so-called „high-frequency triplets" and „low-frequency triplets". What's important is that most of the high-frequency combinations end on pattern trials, while all of the low-frequency combinations end on random trials, and this way trial type (pattern vs. random) and statistical features of stimuli are heavily confounded. Not surprisingly then, learning can be detected by contrasting trial types or by contrasting triplet types (irrespective of which of the two information types drive learning). Both methods can be found in the literature: some researchers treat the ASRT as a pattern-learning task, which is revealed by making comparisons solely on the basis of trial type (pattern vs. random) (e.g. Barnes et al., 2008; D. V. Howard & Howard, 2001; Japikse, Negash, Howard, & Howard, 2003; Negash, Howard, Japikse, & Howard, 2003). Others treat the task as a statistical learning task, since their comparisons are being made solely on the basis of triplet type (frequent vs. infrequent) while ignoring trial type (pattern vs. random) (e.g. Nemeth et al., 2009; Nemeth, Janacsek, Londe, et al., 2010; Nemeth et al., 2011; Janacsek, Fiser, & Nemeth, 2012; Hallgató et al., 2013; Nemeth, Janacsek, Polner, et al., 2013; Stark-Inbar et al., 2016). Finally, in the minority of cases, both factors (trial type and triplet type) are considered simultaneously, (e.g. Horvath, Torok, Pesthy, Nemeth, & Janacsek, 2018; Janacsek et al., 2012; Kóbor et al., 2018; Nemeth, Janacsek, & Fiser, 2013; Schwartz et al., 2003; Simor et al., 2019), making it possible to assess the relative contribution of the two learning types.

Howard and Howard (1997), for example, compared high-frequency triplets that end on random trials (hereinafter RH), high-frequency triplets that end on pattern trials (hereinafter PH) and low-frequency triplets always ending on random trials (hereinafter RL). They found that RH trials were responded to faster and more accurately than RL trials - thus triplet frequency learning did occur. This result couldn't be attributed to pattern learning since only responses to random trials were compared. At the same time, they also found that PH trials were reacted to faster and more accurately than RH trials, possibly indicating pattern (rule) learning. As a reminder, these trials are the ending trials of the same triplets (e.g. 1-2-**2**), but one of the triplets is an R-P-R triplet (thus the critical, final trial is a random trial) while the other is a P-R-P triplet (thus the final trial being a pattern trial). But here is the catch: although triplet level statistical information couldn't act as a confound in this measure, higher order statistical information could (i.e. although the trials being compared are the ending trials of identical triplets, they differ on the N-3th trial). The authors, recognizing this, used the term *higher order learning* when referring to the measure derived from contrasting RH and PH trials.

From Howard & Howard's (1997) work we do know now that triplet level statistical learning occurs in the task, but we still don't know whether it's pattern learning and/or higher order statistical learning that explains improvement of performance that cannot be attributed to triplet level statistical learning (i.e. *higher order learning*). We only know that there is a little *extra* to triplet learning. And albeit being little, this extra is not marginal; this measure differentiates between age groups (J. H. Howard & Howard, 1997). From modified versions of the ASRT task we also know that higher order *statistical* learning is possible: it has been shown that even third-order statistical regularities can be learned by humans, and also that such learning is reduced in the old compared to the young (D. V. Howard et al., 2004; Bennett et al., 2007), just as the *higher order learning* measure is reduced in elderly. But, of course, this does not exclude the possibility that pattern learning also occurs in the ASRT.

Despite the uncertainty that remains about this measure, it is still surprising that only a handful of studies quantified it at all, as it costs nothing to do so and it opens new opportunities for data interpretation. First, if overall differences exist between groups, it can be determined whether differences arise from triplet level learning, higher order learning or both; and second, when no overall differences are detected with the simpler methods, it may be due to decreased sensitivity to detect higher order (subtler) learning. Indeed, only a few studies reported group differences in the ASRT literature using the less elaborate analysis methods, (Barnes et al., 2010; Bergstrom et al., 2012; Hedenius et al., 2011; J. H. Howard et al., 2006; Janacsek, Ambrus, Paulus, Antal, & Nemeth, 2015; Janacsek, Borbély-Ipkovich, Nemeth, & Gonda, 2018; Japikse et al., 2003; Marvel, Schwartz, Howard, & Howard, 2005; Negash et al., 2007; Nemeth, Janacsek, Király, et al., 2013; Schwartz et al., 2003; Takács et al., 2018). We need to increase the sensitivity of the employed analysis methods in order to make the ASRT a truly effective tool measuring implicit learning capabilities. One way of doing so is differentiating between different kinds and levels of learning that can be detected, and that are confounded in the typical analyses. Not only would this result in more diverse information about a particular participant's learning ability, but also in purer measures. The ASRT task might be a goldmine, we should stop digging coal.

V/2.2. Statistical properties and analysis methods of the task

We have talked about how pattern learning is confounded by statistical learning, and how unclear it is what constitutes „higher order learning". The story is however even more complex. For example, in the typical analyses of ASRT data no distinction is being made between joint probability learning (how frequent a particular combination is, e.g. 1-2-2) and conditional probability learning (how often does 1-2-… end with 2), and although the terminology points to the former (e.g. the terms „low-frequency triplet" and „high-frequency triplet"), the way we typically analyze data is more in line with the latter (since we analyze reaction times given to the final elements of triplets; i.e. we measure whether a particular response is faster following a specific set of trials contrasted with different sets of trials). Humans are capable of both kinds of statistical learning (J. H. Howard et al., 2008), and as we will show, the ASRT task has the potential to distinguish between the two. Furthermore, pattern learning and higher than second-order statistical learning can also be separated (even if not perfectly, but at least to a higher degree than we used to).

We summarized trial probabilities and combination frequencies on **Fig. V/1** using two color scales (shades of gray representing combinations frequencies and shades of blue representing the predictability of a given trial; darker shades represent higher frequencies/probabilities) on four levels (0-3 preceding trials taken into consideration). Each bar represents the total number of trials/combinations on a given level (e.g. one-third of a bar represents one-third of the combinations on that level). The upper half of the bars represent combinations that end on a random trial, while the lower half represents combinations that end on a pattern trial. The points of the bars that are at the same height represent the same trials (considering 0-3 antecedent trials when moving from left two right); four examples are shown in boxes connected via red lines.

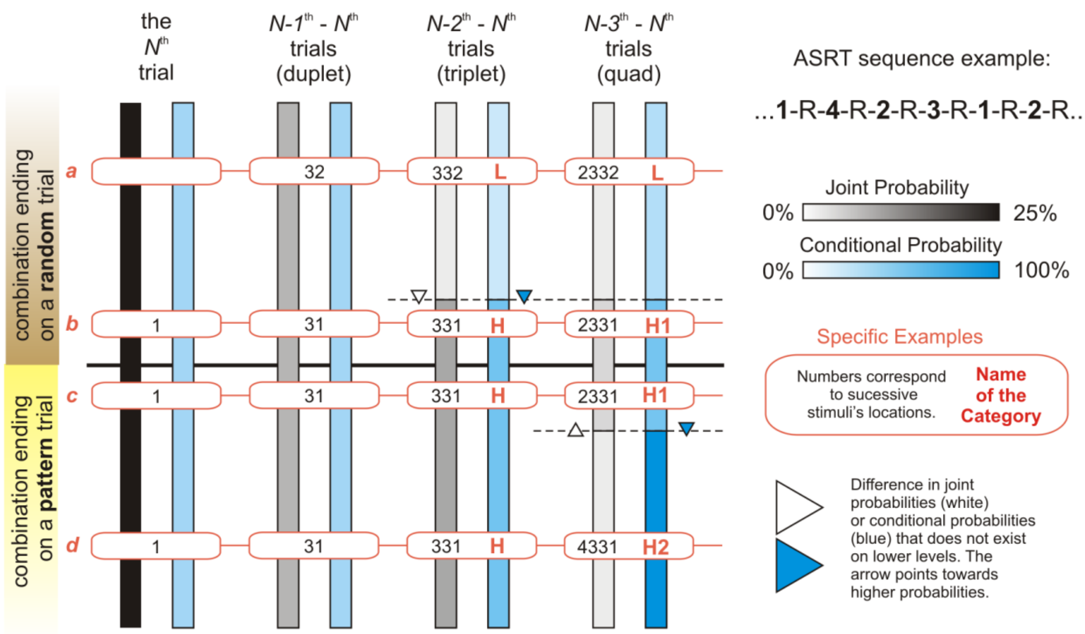


Figure V/1. Statistical properties of the ASRT trials and trial combinations. Shades of gray represent combination frequencies. Shades of blue represent the predictability of a given trial. Darker shades represent higher frequencies/probabilities. Zero to three preceding trials are taken into consideration – see clusters of bars from left to right). Each bar represents the total number of trials/combinations on a given level (e.g. one-third of a bar represents one-third of the combinations on that level). The upper half of the bars represent combinations that end on a random trial, while the lower half represents combinations that end on a pattern trial. The points of the bars that are at the same height represent the same trials (considering 0-3 antecedent trials when moving from left two right); connected boxes show specific examples of the categories.

Several things may be noticed by looking at **Fig. V/1**:

1) Single trials (1, 2, 3 or 4) or duplets (e.g. 12, 13, 14, etc.) are of uniform statistical properties throughout the sequence since the first two groups of bars are of uniform color. In 50% of the cases, these trials/duplets end on pattern trials (bottom halves of the bars), in the remaining cases the same combinations end on random trials (top halves of the bars; contrast the examples *b* and *c*, for example, showing that the combination 31 occurs both ways)
2) When at least two preceding trials are taken into consideration (triplets, quads, etc.) some trials are more predictable than others (blue bars are not uniformly colored) and some combinations are more frequent than others (gray bars are also not uniformly colored either). Moreover, these categories do not overlap perfectly, as, for example, when considering quads, there are only two different shades of gray but three shades of blue (meaning two categories on the basis of combination frequencies, and three categories on the basis of conditional probabilities). Higher joint probabilities sometimes correspond to higher conditional probabilities, e.g. when considering triplets; other times they go in different directions, e.g. when considering quads.
3) On the level of triplets, two categories can be distinguished based on joint probabilities, and the same category boundaries separate trials with different conditional probabilities. E.g. the combination 332 is less frequent than the combination 331 (light gray vs. darker gray part of the first bar), and simultaneously, after the preceding trials 33 it is more probable that a stimulus 1 will follow and not the stimulus 2 (light blue vs. dark blue part of the second bar). The category with the higher probabilities (both joint and conditional) is denoted as H, while the category with the lower probabilities is denoted as L, see the examples *a* vs. *[b and c and d]* on **Fig. V/1**.

Members of the H category can further be divided into two subcategories when the N-3th trial is considered (the new categories being H1 and H2 quads, respectively). With the usual analysis methods, there was no distinction being made between these two quad types. As it can be read from the figure, H1 quads are more frequent than H2 quads (see the dark vs. light gray colors of the first bar), but the final trial of H1 combinations is less probable given its antecedents than the final trial of H2 combinations (see the light blue vs. darker blue colors of the second bar). So, for example, 2331 is a more frequent combination than 4331, but while combinations starting with 433 consistently end with 1, combinations starting with 233 can end with 1, 2, 3 or 4.

As noted earlier, some researchers analyze the data gathered with the ASRT task by contrasting trials of different *Trial Type*, i.e. pattern and random trials, (e.g. Barnes et al., 2008; D. V. Howard & Howard, 2001; Japikse et al., 2003; Negash et al., 2003), resulting in a learning measure called *Pattern Learning* or *Trial Type Effect*. On **Fig. V/1** this corresponds to contrasting the upper half of the trials/combinations with the lower half, i.e. contrasting the exemplars *a*, *b* with *c, d*. This kind of analysis bears on the implicit assumption that the ASRT is primarily a rule-learning (pattern-learning) task, and does not take statistical properties into consideration, albeit being heavily confounded by them; e.g. when considering triplets, members of the H category (e.g. 331, examples *c, d*) are contrasted with a mix of H and L category members (331 and 332, examples *a, b*). We will refer to this analysis method as Model 1.

Other times the assumption is that the ASRT is primarily a triplet learning task (thus a statistical learning task). The learning measure is derived from contrasting performance on H vs. L category members resulting in a measure called *sequence-specific learning, sequence learning effect* or *triplet type effect* (Barnes et al., 2008; Hallgató et al., 2013; D. V. Howard & Howard, 1992, 2001; J. H. Howard & Howard, 1997; Japikse et al., 2003; Negash et al., 2003; Song et al., 2007b). This model does not explicitly deal with the possibility of higher-order learning (e.g. quad level and higher), and thus it does not differentiate between combination frequency learning and trial probability learning (since the correlation between the two is 100% up to the level of triplets). It also doesn't take Trial Type (pattern vs. random) into consideration, albeit these factors are confounded; e.g. 332 only occurs as a combination ending on a random trial (example *a* on **Fig. V/1**), but 331 occurs both ways (examples *b*, *c*, *d* on **Fig. V/1**). Hereinafter we will refer to this analysis method as Model 2.

A third analysis tradition considers both triplet level statistical information,(i.e. Triplet Type; H and L categories) and Trial Type (pattern vs. random trials) (Janacsek et al., 2012; Kóbor et al., 2018; Nemeth, Janacsek, & Fiser, 2013; Schwartz et al., 2003; Simor et al., 2019). This model distinguishes three categories and three learning measures: the difference between performance on random-ending L (LR)and random-ending H (HR) trials is usually called *pure statistical learning* (examples *a* vs. *b* on **Fig. V/1**) , the difference between random-ending H (HR) and pattern-ending H (HP) trials is called *higher order sequence learning* (examples *b* vs. *[c* and *d]* on **Fig. V/1**), while the difference between LR and HP trials is called *maximized learning* (Nemeth, Janacsek, & Fiser, 2013) (examples *a* vs. *d* on **Fig. V/1**). Hereinafter we will refer to this analysis method as Model 3. Importantly, this method treats pattern trials as a uniform category, while in reality pattern trials can be divided into the subcategories H1 and H2 considering quad level statistical information (e.g. high frequency triplets such as 331 may be part of quads 2331 – H1 category – or 1331 / 3331 / 4331 – H2 category, see **Fig. V/1** examples *b, c* and *d*). This is particularly important as the *Higher Order Learning* measure was the one to differentiate between age groups (J. H. Howard & Howard, 1997), and the authors raised the possibility themselves that the measure might include higher level statistical learning in addition to or instead of pattern learning. The problem is that the Higher Order Sequence Learning measure contrasts quads from the H1 statistical category with quads from both H1 and H2 categories. If the driving force of learning is indeed statistical information, it is plausible to assume that this measure is underestimated, as the difference between H1 and (H1+H2) quads must be smaller than the difference between pure groups of H1 and H2 quads. Thus, we suggest that „higher order statistical learning" (i.e. quad learning) could be detected more efficiently if H1 quads would be contrasted with H2 quads instead of contrasting HP and HR trials (i.e. we suggest contrasting *a* vs. *[b and c]* vs. *d* instead of contrasting *a* vs. *b* vs. *[c and d]* on **Fig. V/1**).

For this reason, we introduce Model 4 and Model 5 in this paper as possibly better analysis methods. In Model 4, quad level statistical information is considered, but Trial Type (random vs. pattern) is not. Thus, it treats the ASRT as a solely statistical learning task with no rule-learning (pattern-learning) component. The categories being compared are L, H1 and H2 (*a* vs. *[b and c]* vs. *d* on **Fig. V/1**, quad columns). Importantly, trial predictability (i.e. conditional probabilities) and combination frequencies dissociate clearly in this case: H2 combinations are less frequent than H1 combinations but their final trial can be anticipated with much higher probability given the first three trials of the combination The difference between L and H1 trials could be called *triplet learning* (+ *pattern learning*), the difference between H1 and H2 trials *quad learning* (+ *pattern learning*), while the difference between L and H2 trials could be called *maximized learning*. In this Model, the categories differ more clearly in the statistical aspect and less clearly in *Trial Type*, while the opposite was true in Model 3 (hence the parentheses in the suggested names for the learning measures; trial type learning is secondary in Model 4). Thus, if the primary driving force of learning is sensitivity to statistical information (rather then sensitivity to the hidden pattern), Model 4 should fare better than Model 3, and vice versa.

Lastly, Model 5 would consider both *Trial Type* (pattern or random) and *Quad Type* (L, H1 and H2 categories), resulting in four categories (examples *a* vs. *b* vs. *c* vs. *d* on **Fig. V/1**, quad columns, considering whether the combination ends on a pattern or random trial as well). Again, on the level of quads, trial predictability and combination frequency dissociate (they point in the opposite direction), thus their relative impact might be assessed just as with Model 4. As a bonus, random-ending H1 trials (H1R) can be contrasted with pattern-ending H1 trials (H1P), leading to a pattern-learning measure that is less confounded by statistical information than the pattern-learning measures of the previous models. We propose the following names for the resulting learning measures: *triplet learning* (LR vs. H1R; examples *a* vs. *b* on **Fig. V/1**), *pattern learning* (H1R vs. H1P; examples *b* vs. *c* on **Fig. V/1**), *quad learning* (H1P vs H2P; examples *c* vs. *d* on **Fig. V/1**) and *maximized learning* (LR vs. H2P; examples *a* vs. *d* on **Fig. V/1**).

An overview of Model 1 to Model 5 is illustrated on **Fig. V/2**. One of the aims of the current study was to compare these models in terms of goodness of fit, thus to decide whether it pays off to use a more elaborate model when analyzing ASRT data.

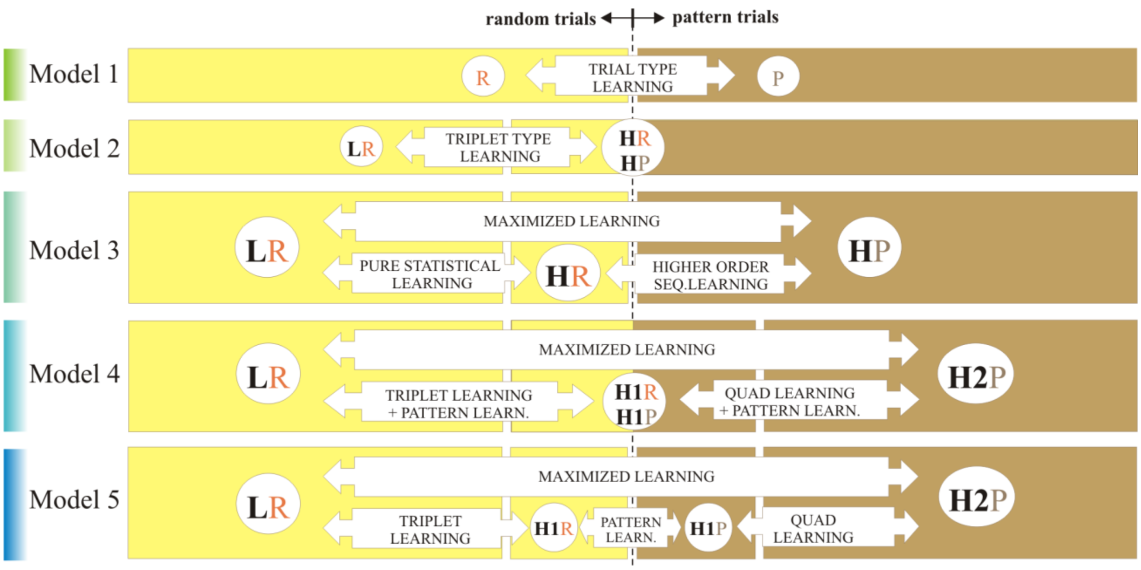


Figure V/2. Different models of the ASRT task as a basis of extracting different learning scores. P – pattern trials, R – random trials, L – low probability trials, H – high probability trials (H1 and H2 being subcategories of the latter; H2 trials are more probable than H1 trials, but at the same time triplets that end on a H2 trial are less frequent than triplets that end on a H1 trial). Models 1-3 has been typically used as a basis of data analysis; Model 4-5 are introduced in this paper.

V/2.3. Confounding variables in the ASRT task

The ASRT task is a reaction time task, and reaction times vary as a function of many factors. Fatigue (its effects could be approached as in (Török, Janacsek, Nagy, Orbán, & Nemeth, 2017)), boredom, stimulus timing, the number of response locations etc. may all affect the magnitude and variability of reaction times (E. Hick, 1952; Grosjean, Rosenbaum, & Elsinger, 2001; Woods, Wyma, Yund, Herron, & Reed, 2015), and thus our ability to detect learning on the task, or even learning itself. But these factors at least have a similar impact on the groups of trials that are to be contrasted in the ASRT task - either because they are constant (e.g. response-stimulus interval), or because the different trial types are evenly distributed throughout the task, thus even time-dependent factors such as fatigue have similar effects on the different trial types in a given time window.

There is, however, at least one factor which may act as a confounding variable: for some stimulus combinations, e.g. serial repetitions of the same stimuli, response facilitation is observed when contrasted with other combinations, e.g. an inconsistent pattern of alternations and repetitions. These so-called *sequential effects* (Remington, 1969) are evincible in random streams of stimuli, but also from reaction time tasks in which the conditional probabilities of stimuli vary, see (Kornblum, 1973). We have no solid idea of exactly which combinations should be relatively „easier" (facilitated) compared to others because this phenomenon has mostly been studied in binary-choice reaction time tasks (Kirby, 1976; Soetens, Boer, & Hueting, 1985; Vervaeck & Boer, 1980; but see Lee, Beesley, & Livesey, 2016), those combinations being less numerous and less complex than the combinations in the ASRT task. Also, the type and direction of these effects depend strongly on the response to stimulus interval (RSI). The automatic facilitation effect (that is of interest to us) typically occur with RSIs of 100 ms or less, although the exact values tested vary from experiment to experiment, as summarized by (Gao, Wong-Lin, Holmes, Simen, & Cohen, 2009), and these results are mostly derived from two-choice reaction time tasks and thus might not apply for the ASRT. In the absence of concrete expectations of how and to what extent sequential effects occur in the ASRT (and bearing in mind that ASRTs with different RSIs may differ in this regard), the wisest thing we can do is to ensure that the groups of trials that are to be contrasted in the ASRT (e.g. pattern vs. random trials or highly predictable vs. moderately/slightly predictable trials, etc.) belong to the same types of combinations with respect to local sequential effects ("easy" or "hard").

The most influential proposal that aimed at reducing unwanted sequential effects was of Howard et al. (2004) who eliminated spans (also called *trills*, e.g. *a-b-a*) and repetitions (e.g. *a-a-a*) from the analysis since these types of triplets always occur as random trials for each participant (irrespective of the particular sequence being taught). As Song and her coworkers put it, "perfomance on trills and repetitions could reflect preexisting biases, rather than sequence learning" (Song et al., 2007a, p. 168.). Each of the remaining 48 triplets can be described in the abstract form as *cba, bba or baa* (where different letters represent different stimuli), and the proportion of these types of triplets is similar in the groups being compared on the basis of their statistical properties (i.e. high- vs. low-frequency triplets). Even if one type of triplets, e.g. *baa* is easier than the other types of triplets (since it ends on a repetition), this shouldn't pose a problem because the proportion of *baa* triplets is similar across high- and low-frequency triplets. Moreover, each of the 48 individual triplets have an equal chance of being a high-frequency triplet (they are high-frequency triplets in some of the sequences, and low-frequency triplets for the remaining sequences), thus, on the group level, pre-existing biases shouldn't prevail. Since in this case filtering is applied at triplet level, hereinafter we refer to this method as *triplet filtering.*

Interestingly, Song et al. (2007a) found that preexisting biases can be found even on the level of quads. At the beginning of their ASRT, RH triplets were temporarily faster and more accurately reacted to than PH triplets – for a similar finiding see (Nemeth, Janacsek, & Fiser, 2013). In order to eliminate possible preexisting biases that could cause this effect, they categorized quads into seven categories: "those that contain two repeated pairs (i.e., 1122); a repeated pair in the first (1124), second (1224), or last (1244) position; a run of three in the first position (1112); a trill in the first position (1213); or no repeated elements (1243)" (p. 170.). After removing all the unequally represented quad types of this sort, the unexpected difference between RH and PH trials in the first session disappeared, whereas the difference between low- and high-frequency triplets remained. Although the paradoxical RH-PH difference only manifested in the first session (after 150 repetitions of the pattern), and reversed afterward, it is still quite surprising that this method of eliminating pre-existing biases on the level of quads did not become commonly used. One reason might be that the description of this method was limited to a few words in a footnote.

In this work, we propose the elimination of quad level preexisting biases by using a similar method. Our notations were derived the following way: whatever the current stimulus was (position 1, 2, 3 or 4), it was denoted as *„a".* If the previous stimulus was identical to the current one, it was also denoted as *„a"*, thus the combination of the two was denoted as *„aa".* Otherwise, if the previous stimulus was different, the combination was denoted as *„ba".* Going further, if the N-2th trial was identical to the N-1th or Nth trial, it was denoted with the same letter as the one that it was identical to (e.g. „aba" or „bba"); otherwise, it got the following letter from the alphabet (e.g. *„c"*). This way a quad that consistsed of four different stimuli was always denoted as *„dcba"* (irrespective of whether it was derived from 1-2-3-4, 3-1-4-2 or else). Importantly, we assigned these letters to stimuli starting with the Nth trial and going backward in order to be able to match combinations of different lengths. For example, a triplet consisting of three different stimuli was „**cba**”, and the same triplet could be part of a quad „a**cba**”, „b**cba**”, „c**cba**” or „d**cba**”.

As a difference to Song et al. (2007a), our quad categories rely on the abstract structure of the quads, thus differentiating between 1-2-2-1 and 3-2-2-1 quads (*abba* vs. *cbba*), and between 1-2-1-1 and 3-2-1-1 quads (*abaa* vs. *cbaa)*, too; moreover we observed a category that was not mentioned by Song et al. (2007a), namely *acba* quads (e.g. 1-2-3-1). Only three out of 13 categories are counterbalanced across the groups of trials being compared within subjects (e.g. P vs. R in Model 1, L vs. H in Model 2, etc., see **Fig. V/2**) and across participants (i.e. any particular quad having an equal chance of belonging to either statistical category). These quad types are *dcba, cbba* and *acba.* Hereinafter, we will refer to this filtering method as *Quad Filtering*. As a specific example of the possible benefit of using the Quad Filtering is the elimination of *bbaa* quads (e.g. 1122, 1133, 2233, etc.), which seem to be the easiest (fastest) combinations of all. These combinations only occur as members of the H1 category, moreover, they constitute approximately 25% of that category. Different repetition-ending combinations (e.g. *abaa, cbaa*) do occur in other statistical categories as well (e.g. L, H2), and they are also reacted to relatively fast (compared to nonrepetition-ending combinations), but they only make up 8-16% of a particular category. In other words, H1 category is, on average, easier than the L or H2 categories, which might manifest in overestimated Triplet Learning measures and either underestimated Quad Learning measures (if dominantly conditional probabilities are being learned) or overestimated Quad Learning measures (if dominantly joint probabilities are being learned). Using the previous models of ASRT, this bias could have manifested as an overestimation of the Pure Statistical Learning measure of Model 3, and either in an underestimation of the Higher Order Learning measure - given that learning is driven by conditional probabilities (it could also cause the paradoxical negative difference between the HR and HP categories), or an overestimation of the same measure (given that learning is driven by joint frequencies).

We would like to highlight, however, that these filtering methods do not necessarily eliminate pre-existing biases on the individual level. Even if the percentage of, say, *dcba* quads is counterbalanced across statistical categories (i.e. within participants) and across participants, it is still reasonable to assume that some of these quads are easier than others. For example, Lee, Beesley, & Livesey (2016) observed that sequences of trials with many changes in directions (of movements) are harder to react to than sequences of trials where the direction of movement does not change (e.g. 1-4-2-3 is harder than 1-2-3-4). Since quads are not counterbalanced in this respect within participants, learning on some or all of the sequences might be confounded by such biases despite being controlled for on the group level. This is particularly important if we aim is to measure individual differences in ASRT learning: any correlation with other measures (or the lack of correlation) might be due to such confounds.

Also, while the elimination of quad level pre-existing biases on the group level should make the interpretation of the results more straightforward, one has to keep in mind that higher level sequential effects could still act as a confound. Lee et al. (2016) observed that lower level sequential effects had a higher effect size than higher order sequential effects (e.g. $\eta_p^2 = 0.881$ on the triplet level vs. $\eta_p^2 = 0.275$ on the quad level and $\eta_p^2 = 0.314$ on the quint level), but the latter nevertheless influenced reaction times. This indicates that such biases should be controlled for at least on the level of quints, or on even higher levels, to minimize their impact when assessing statistical learning. Unfortunately, there are no quints that are counterbalanced across all the relevant statistical categories across participants, so there is no easy way of assessing the impact of preexisting biases on this level. Further studies should investigate the magnitude of such biases, for example by using random series of stimuli in a 4-choice (ASRT-like) reaction time tasks and assessing the preexisting tendencies when reacting to quints that also occur in the ASRT sequences.

V/2.4. The aim of the study

In the previous section, we described how different types of information might be the basis of learning or might influence learning in the ASRT task, such as trial type, trial probability, combination frequency and preexisting biases to certain stimulus combinations. We also noted that there are at least three types of analysis utilized by different research groups (we will refer to these as Model 1, Model 2 and Model 3, respectively), and we made suggestions on how to improve these methods (we will elaborate on these when introducing Model 4 and Model 5 later in this text).

In the following sections, we will first review the different analysis methods (Model 1 to Model 5) with respect to possible confounds in their learning measures. We considered such a detailed description to be helpful not only for deciding which analysis method to use for one's purposes in the future but also for the evaluation (or re-evaluation) of previous findings. Along with the description of the learning measures of these models, we will present data confirming and illustrating the theoretical considerations reviewed so far.

Second, we will compare the models (Model 1 to Model 5) in terms of goodness of fit, which might corroborate our proposals from a different perspective (i.e. how efficient a Model is in capturing the different aspects of learning) on the ASRT task. Although we will present and compare findings got with different filtering methods (No Filter, Triplet Filter, and Quad Filter), the main focus of this section will be on differences between Models (Model1-5) within Filtering Methods, and not vice versa. The reason for this is that the superiority of one filtering method over another cannot easily be justified via statistics (i.e. the purer, bias-free effect size might be smaller than the biased; it should nevertheless be preferred. Statistics can only tell us about the magnitude of effects, but not their purity).

Third, we will elaborate on the specific learning effects in each Model (Model1-Model5). These measures cannot be directly compared (remember, the reason for introducing Model4 and Model5 was the fact that specific learning measures in Model 1-3 might reflect the mixed learning of different types of information), and so the main focus of this section will be the comparison of different Filtering Methods within the Models. We will discuss the effect of these filters on the magnitude of learning and individual variability that can be detected in the task.

Last but not least, in the fourth section, we will examine the learning scores calculated with the methods that we propose. The main focus here will be on whether participants (as a group) showed learning of the different statistical properties of the sequence, and also the percentage of participants who showed learning of these properties (individually). We will also consider the time-course of learning of these aspects (e.g. does quad learning occur later in time than triplet learning?).

V/3. Methods

We based our analysis on data originally collected by (and published in) Török, Janacsek, Nagy, Orbán, & Németh (2017) with the authors' permission. Because of this, we copied the most relevant parts of their Methods section (not necessarily in the same order as originally provided); a more detailed description can be found in the aforementioned paper.

V/3.1. Participants

One hundred and eighty healthy young adults participated in the study, mean age $M = 24.64$ ($SD = 4.11$), $Min_{age} = 18$, $Max_{age} = 48$; 28 male/152 female. All participants had normal or corrected-to-normal vision and none of them reported a history of any neurological and/or psychiatric condition. All participants provided written informed consent before enrollment and received course credits for taking part in the experiment. The study was approved by the United Ethical Review Committee for Research in Psychology (EPKEB) in Hungary (Approval number: 30/2012) and by the research ethics committee of Eötvös Loránd University, Budapest, Hungary. The study was conducted in accordance with the Declaration of Helsinki.

V/3.2. Equipment

The Alternating Serial Reaction Time (ASRT) task was used to measure statistical learning capabilities of individuals (J. H. Howard & Howard, 1997).

V/3.3. Procedure

Participants were instructed to press a corresponding key (Z, C, B, or M on a QWERTY keyboard) as quickly and accurately as they could after the stimulus was presented. The target remained on the screen until the participant pressed the correct button. The response to stimulus interval (RSI) was 120 msec. The ASRT task consisted of 45 presentation blocks in total, with 85 stimulus presentations per block. After each of these training blocks, participants received feedback about their overall RT and accuracy for 5 seconds, and then they were given a 10-s rest before starting a new block. Each of the three sets of 15 training blocks constitutes a training session. Between training sessions, a longer (3–5 min) break was introduced.

Each participant was given a randomly chosen ASRT sequence (out of the six possible sequences). This way 32 of the participants got the sequence *1-r-2-r-3-r-4-r*, 29 participants got *1-r-2-r-4-r-3-r,* 31 participants got *1-r-3-r-2-r-4-r,* 33 participants got *1-r-3-r-4-r-2-r,* 29 participants got *1-r-4-r-2-r-3-r* and 26 participants got *1-r-4-r-3-r-2-r*. EPRIME 2.0 was used as a stimulus presentation software (Schneider, Eschman, & Zuccolotto, 2012).

V/3.4. Statistical Analyses

Probability distributions of continuous variables (subsets *a* vs. *b*) were compared using the *Kolmogorov-Smirnov* test and the *Mann-Whitney test*. Effect sizes for such differences were computed in the form of *Probability of Superiority*, i.e. the probability that a randomly chosen value from subset *b* is higher than a randomly chosen value from subset *a*. Distributions of nominal variables were compared using the Chi-Squared test, and Cramer's V was computed as the corresponding effect size.

Models' goodness of fit was computed in the form of *adjusted R-squared* values (in the case of reaction times) and *Cramer's V* values (in the case of error data). Variability was computed in the form of *standard deviations (SD)* and *coefficients of variation (CV)*. Specific learning scores were quantified as *Cohen's d* effect sizes (reaction times) and *Cramer's V* values (error data). For the comparison of these values (within Models or within Filtering Methods) we used *ANOVA*s, and we reported *partial eta squared* effect sizes along with *p* values.

To assess whether the variability of two data sets is different we used the *Levene-test*. The reliability of the measures was assessed via the *split-half method*.

V/4. Results

V/4.1. Variables that contribute to the learning scores of different Models using different filtering methods

In this section we aimed to statistically confirm the considerations we discussed so far on a theoretical basis, which is not only important in order to strengthen our message, but also because the ASRT sequence is not fully pre-determined (as half of the trials are randomly determined), thus the actual sequence varies from participant to participant. While it is always true that there is 25% chance for a random stimulus to be

1, 2, 3 or 4, it is not guaranteed that in a particular sequence these outcomes will have frequencies that match their probabilities (e.g. stimulus 2 might come 32% of the time for a particular participant). According to the law of large numbers, the more trials we have, the more the actual frequencies will approach theoretical probabilities. In this study, participants performed 45 blocks of ASRT which corresponds to approximately 1750 random trials that shape the overall statistical properties of the sequence. We aimed to assess to what extent do previously described considerations apply on the individual level with this amount of random trials. Is it possible, for example, that for some participants there is a difference in actual statistical properties between two categories that should not differ based on theory (e.g. H1 and H2 trials differing in triplet level conditional probabilities)? Or is it possible that for some participants quad filtering is not effective in balancing out combination types across statistical categories? How often do these kinds of anomalies occur with the three different filtering methods?

V/4.1.1. Trial Type Proportions

As it can be read from **Fig. V/2**, most of the statistical categories in the different models contain only P trials (Model 1 P; Model 3 HP; Model 4 H2; Model 5 H1P and H2P) or only R trials (Model 1 R; Model 2 L; Model 3 LR and HR; Model 4 L; Model 5 LR and H1R). The only exceptions are Model 2's H category and Model 4's H1 category which contain both P and R trials; the H1 category is made up of 50% R and 50% P trials (regardless of the filter being used); while the H category of Model 2 consists of 20% R and 80% P trials when No Filter or Triplet Filtering is applied; and it contains 33% R and 67% P when the Quad Filter is applied (see Supplementary Table **ST-V/1** in the Supplementary Materials for corresponding statistics).

Ideally only those categories should differ in the P/R proportions that are used to compute *Pattern Learning* scores, i.e. the learning that (possibly) occurs if participants are sensitive to the trial type (P or R) in addition to statistical information, such as the P vs. R category in Model 1 and the H1P vs. H1R categories of Model 5. In other cases, the differences in P/R proportions are not of a concern because the contrasts admittedly assess mixed effects of different learning types (e.g. the HP vs. HR categories of Model 3 or the Maximized Learning scores of Models 3-5). The only problematic contrasts are the L vs. H categories in Model 2 and the L vs. H1 and H1 vs. H2 categories of Model 4 (see Supplementary Table **ST-V/1** in the Supplementary Materials). Remember, these models treat the ASRT as a primarily statistical learning task; nevertheless, the Triplet Learning and Quad Learning scores might be confounded by pattern learning resulting in an overestimation of statistical learning.

V/4.1.2. Combination Frequencies (Joint Frequencies)

As described earlier, some combinations of consecutive trials are more frequent than others – and the longer combinations we assess, the more clusters we find, see **Fig. V/3** for illustration. On this figure, histograms of combination frequencies are shown for each Model's every category. It can be seen that from Model 1 to Model 5 the distributions are getting „narrower" (indicating a better categorization based on statistical properties).

Since learning scores are based on contrasting different categories within models, it is crucial that those categories should differ in combination frequencies that explicitly try to capture this aspect of learning (e.g. the H vs. L categories in Model 2; the LR vs. HR category of Model 3, the L vs. H1 categories of Model 4, and the LR vs. H1R categories of Model 5 when assessing triplet level learning; and the H1 vs. H2 categories of Model 4; and H1P vs. H2P categories of Model 5 when assessing quad learning). Some contrasts admittedly assess mixed effects, such as the HP vs. HR contrast of Model 3 and the Maximized Learning scores of Model 3-5.

As it can be seen on **Fig. V/3** (and read from Supplementary Table **ST-V/2** in the Supplementary Materials), these criteria are mostly met. However, for a small subset of participants, triplet level frequencies also differed between the HR vs. HP categories of Model 3 (~3% of participants); between the H1 and H2 categories of Model 4 (~ 4-8% of participants); between H1R and H1P categories of Model 5 (~2-4% of participants) and between H1P and H2P categories of Model 5 (~5-7% of participants). On a positive note, many of the previously discussed (predicted) effects were also confirmed; e.g. that Model 2 is better in capturing the triplet level statistical properties of the sequence than Model 1 (since effect sizes are higher for the former); and that the H1 vs. H2 distinction of Model 4 (and the H1P vs. H2P distinction of Model 5) also leads to higher quad level differences than the distinction HR vs. HP in Model 3 (for these statistics, see SupplementaryTable **ST-V/2** in the Supplementary Materials).

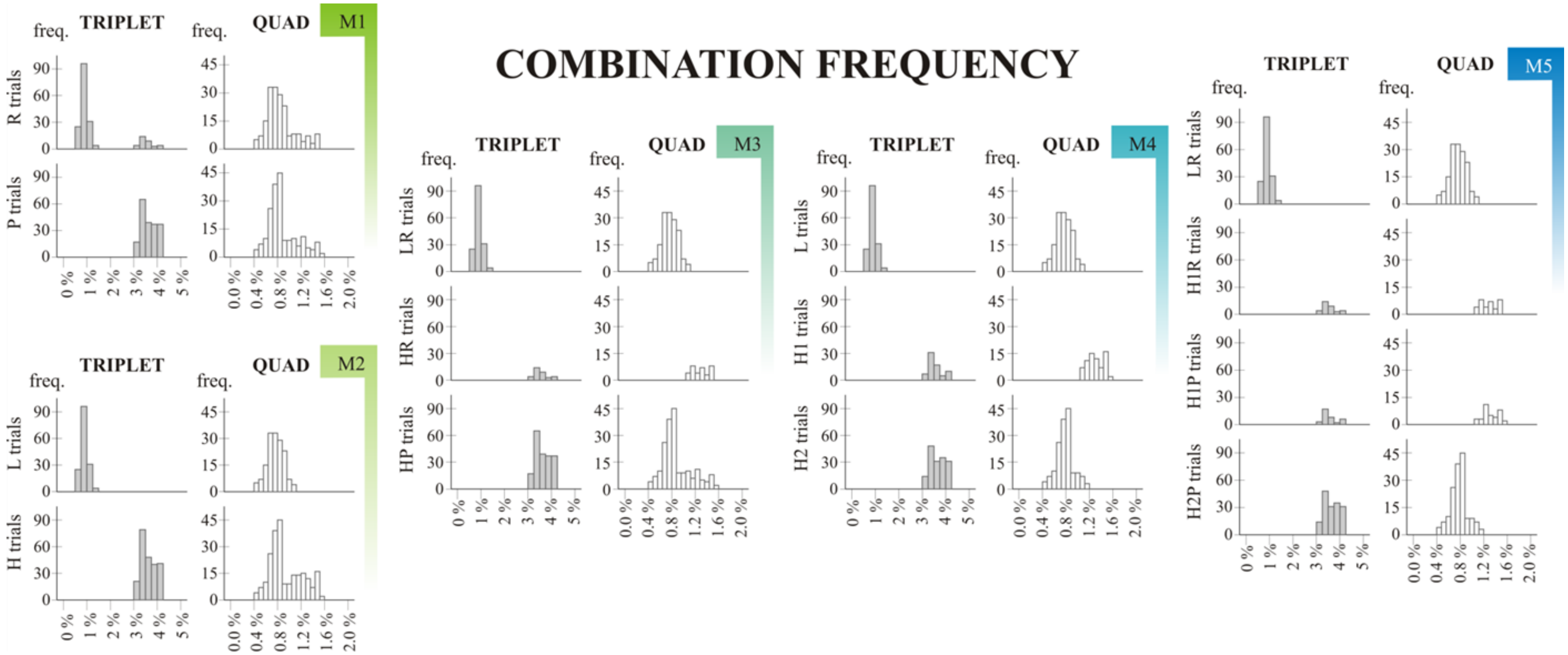


Figure V/3. Combination Frequency. M1 – Model 1; M2 – Model 2; M3 – Model 3; M4 – Model 4; M5 – Model 5. Combination frequency histograms are based on the ninth epoch (final ~400 trials) of a randomly chosen subject (subject number 111). The X axis shows the combination frequencies that occured in the given epoch of the ASRT task; the Y axis represent the frequency with which these occured. Two (triplet level) or three (quad level) preceding trials were taken into consideration when calculating joint probabilities (represented in different columns). Different rows represent different statistical categories within Models.

V/4.1.3. Trial Probabilities (Conditional Probabilities)

Here the same conditions apply as described in the *Combination Frequencies* subsection since category boundaries are the same for joint frequencies and conditional probabilities (even though the direction of differences are not always the same). E.g. based on joint probabilities it could be expected that participants perform better on H1 than on H2 trials (since H1 combinations are more frequent than H2 combinations); based on conditional probabilities, however, better performance could be expected on H2 trials (since conditional probabilities are higher than for H1 trials).

The results are illustrated in **Fig. V/4** and the corresponding statistics can be found in Supplementary Table **ST-V/3** in the Supplementary Materials; these are all very similar to those discussed earlier at *Combination Frequencies.* As a plus, it was shown that quad level statistical information has a higher impact on Model 1 and Model 2 learning scores when assessing trial probabilities then when assessing combination frequencies (in line with the theoretical predictions; see Supplementary Tables **ST-V/2** vs. **ST-V/3** in the Supplementary Materials).

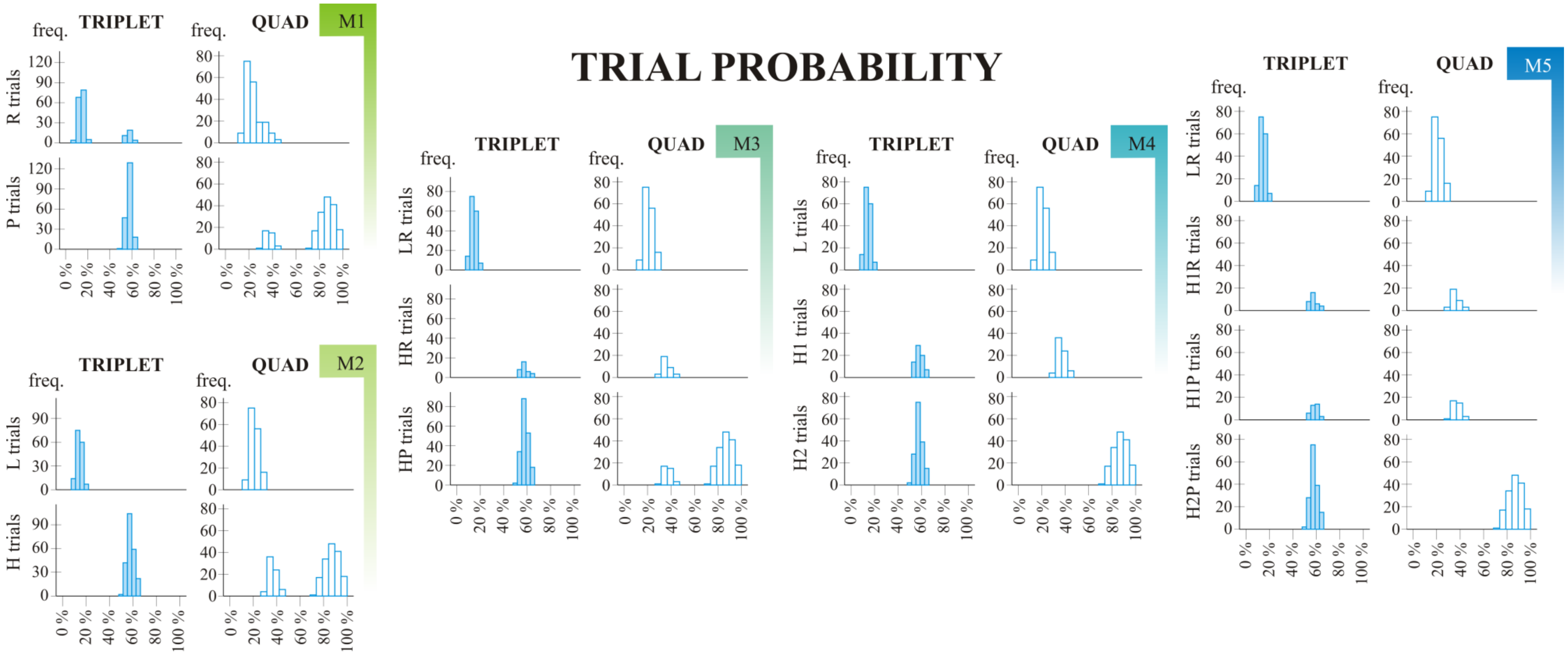


Figure V/4. Trial Probability. M1 – Model 1; M2 – Model 2; M3 – Model 3; M4 – Model 4; M5 – Model 5. Trial probability histograms are based on the ninth epoch (final ~400 trials) of a randomly chosen subject (subject number 111). The X axis shows trial probabilities that occured in the given epoch of the ASRT task; the Y axis represent the frequency with which these occured. Two (triplet level) or three (quad level) preceding trials were taken into consideration when calculating joint probabilities (represented in different columns). Different rows represent different statistical categories within Models.

V/4.1.4. Abstract Structure of the Combinations

Ideally, all of the categories (within models) should consist of similar combination types with regard to the combinations' abstract structure, since pre-existing biases are not something we aim to measure in this task. The nevertheless existing differences can be reduced by applying filters (the usual Triplet Filter or the now-proposed, stricter, Quad Filter), see **Fig. V/5**. However, as can be read from Supplementary Table **ST-V/4** in the Supplementary Materials, some differences remain. When the triplet filter is applied, and only three consecutive trials are considered, the most affected learning scores are those got by contrasting the P vs. R categories in Model 1 (affecting the scores of ~11% of participants), the HR vs. HP categories of Model 3 (~11% of participants); and the H1R vs. H1P categories of Model 5 (~15% of participants). By applying the stricter quad filter, the percentage of affected participants was reduced to 2%, 6%, and 5%, respectively. When considering four consecutive trials (i.e. quads), almost all of the learning scores are affected with Triplet Filtering (100% of participants showing such differences between the contrasted categories, with the exception of H1R vs. H1P learning in Model 5, which was only affected in 26% of participants). Evidently, by applying quad filtering, these numbers are greatly reduced. The most affected learning scores are the HR vs. HP learning in Model 3, and the H1R vs. H1P learning of Model 5 (affecting 14% and 19% of participants, respectively).

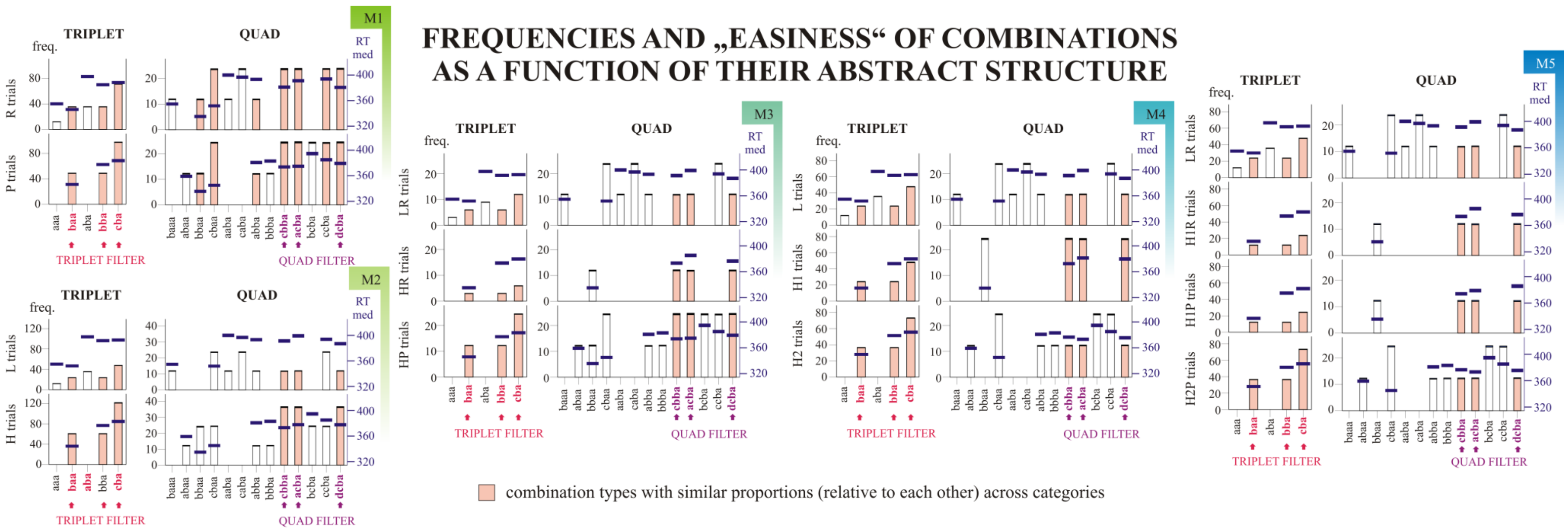


Figure V/5. Abstract Structure of the Combinations. M1 – Model 1; M2 – Model 2; M3 – Model 3; M4 – Model 4; M5 – Model 5. The abstract structure of the combinations were defined the following way: the final trial of a combination was always denoted as a; the preceding trial as either a (if it was the same as the final trial) or b (in all other cases). If the N-2th trial was identical to the Nth or N-1th trial, the same notation was used as before (e.g. a or b), in all other cases a new notation was introduced (eg. c), etc. Bars indicate the mean number of category members in an epoch (~400 trials) calculated for each epoch of each participant. The black boxes at the top of the bars indicate the 95% confidence intervals of these means. The relative proportion of categories colored rose is identical in the Model's subcategories. Dark blue boxes indicate the 95% confidence intervals of means of median RT values corresponding to the different categories (again, computed separately for each epoch of each participant). Red and purple arrows point to categories that are analyzed with Triplet Filter and Quad Filter, respectively.

The fact that around 20% of participants show differences in combinations' distribution (based on their abstract structure) for these Pattern Learning scores is important, and it should be taken as a warning to interpret these learning scores with caution.

In sum, only some of the learning scores fare well in measuring purely that factor that they aim to. The best model in this aspect is Model 5 which results in 4 (relatively) pure measures given that all factors contribute to learning. If, however, some of the information types are not picked up by participants, then less complex models could fare just as well as Model 5. For example, if pattern learning does not occur in the task (meaning that people do not differentiate pattern and random trials in addition to differences in statistical properties), then Model 4 should fare as good as Model 5. The question of which model results in highest explanatory power is assessed in the following section.

V/4.2. Comparison of the Models' goodness of fit

As described in the previous section, Models 1-5 distinguish between an increasing number of categories (based on Trial Type and/or Statistical Information), and the main question is whether it's worth to use the more elaborate models or there is no difference in how well they capture the essense of learning on the task. E.g. differentiating betweeen H1 and H2 trials (in Model 4 and Model 5) only makes sense if participants (or people, in general) are able to differentiate betweeen these categories when performing the task, which, on the other hand, should be reflected in differences in mean reaction times and/or accuracy. A way of assessing the goodness of fit of a model is by computing Adjusted R^2 values (for reaction times data) and by computing Cramer's Vs (for error data), and this is what we did for each model. If introducing or changing sub-categories explains the variability of data to a higher degree (i.e. there *is* a difference between H1 and H2 trials), the fit of the model will be higher, thus the goodness of models can be directly compared.

In this section the focus is on the goodness of models and not the effect of filtering; nevertheless we calculated the aforementioned effect sizes separately for each filtering method (no filter, triplet filter, quad filter), mainly to examine whether the change in effect sizes shows a similar pattern irrespective of the filtering being used.

V/4.2.1. Reaction Times data

Each block consisted of 85 trials, five warm-up (random) trials and 80 ASRT trials (the alternation of pattern and random trials). Warm-up trials were not analysed, neither were trials 6-8 in a block (since it is only from trial 9 that the first full ASRT-quad is reached). This way 9.4% of trials were excluded. Out of the remaining trials, additional 18 percent was excluded due to errorneous responses on any of the quads trials (in other words, only those reaction time data points were analysed which corresponded to correct answers preceeded by another three correct answers in a row). Reaction times higher than 1000ms or lower than 150ms were also excluded from analysis (0.1 percent of the remaining data). Additionally, reaction times having a Z score higher than 2 or lower than -2 were removed from each epoch from each statistical category of the most sophisticated Model (i.e. Model 5, LR, H1R, H1P and H2P) for each participant to minimize the effect of outliers. This way 4.5 % of the remaining data was removed when using no filter; 16% when using the triplet filter and 64.6% when using the quad filter (the high percentage of excluded trials using the triplet filter and quad filter results from the filters themselves, not from so many Z-scores having a high absolute value). At the end, an average of 2710 trials were analysed per participant when using no filter; an average of 2384 trials when using triplet filter and an average of 1006 trials when using the quad filter.

We computed individual adjusted R^2-s for each epoch of each participant as a way of assessing the goodness of fit of each Model; since there were nine epochs, this resulted in nine values per participant. These were than averaged to yield a single value for everyone. The effect of different filtering methods was also taken into account by computing these effect sizes for each filtering type separately (No Filter, Triplet Filter and Quad Filter). The goodness of fits were then compared by a FILTER TYPE (3 levels: No Filter, Triplet Filter, Quad Filter) x MODEL (5 levels: Model 1 - Model 5) Repeated Measures ANOVA. Sphericity was assessed with Mauchly's Test, and if this precondition was not met, degrees of freedom were adjusted with the Greenhouse-Geisser method. Bonferroni-corrected post hoc tests were performed whenever the omnibus ANOVA showed significant main effects or interactions. Partial eta squared effect sizes are reported in line with significant main effects or interactions in the ANOVA.

The main effect of FILTER TYPE was significant, $F(1.553, 278.066) = 25.562$, $MSE < 0.001$, $p < 0.001$, $\eta_p^2 = 0.125$, indicating that, on average, the goodness of fits differed as a function of the filter used. These differences were better captured with a quadratic model than with a linear one ($p < 0.001$ vs. $p = 0.190$ and $\eta_p^2 = 0.271$ vs. $\eta_p^2 = 0.010$ respectively). Bonferroni corrected post hoc tests revealed that means of adjusted R squared values were highest with the Quad Filter and lowest with the Triplet Filter, all contrasts being significant ($p < 0.001$) except for the contrast No Filter vs. Quad Filter ($p = 0.569$). The main effect of MODEL was also significant, $F(1.384, 247.759) = 408.371$, $MSE < 0.001$, $p < 0.001$, $\eta_p^2 = 0.695$, indicating that model goodness of fits differed as a function of the Model used in the analysis. These differences was best explained with a linear model (as compared to quadratic, cubic or higher order models; $\eta_p^2 = 0.778$ for the linear model, and $\eta_p^2 < 0.555$ for the other models), as values grew monotonicaly from Model 1 to Model 5. Bonferroni corrected post hoc tests revealed that all paired comparisons were significant (all $p < 0.001$). Finally, the interaction of FILTER TYPE x MODEL was also significant, $F(2.492, 446.058) = 11.122$, $MSE < 0.001$, $p < 0.001$, $\eta_p^2 = 0.058$, indicating that the monotonic growth of adjusted R squared values as a funtion of MODEL were not equivalent with the three filtering methods used. Bonferroni corrected post hoc tests revealed that each Model differed from all the others within each filtering method (all $p < 0.012$). The effect of the differing filters was also quite consistent with each Model, showing that both the No Filter condition and the Quad filter condition yielded higher fits than the Triplet Filter condition (all $p < 0.001$), the Quad Filter and No Filter condition not differing from each other in 4 out of 5 cases (all $p > 0.437$, except for Model 2 where $p = 0.006$). The results are shown on **Fig. V/6/a**). A more fine grained, epoch-by-epoch analysis of adjusted R^2 values is shown on Supplementary Figure **SF-V/1** in the Supplementary Materials.

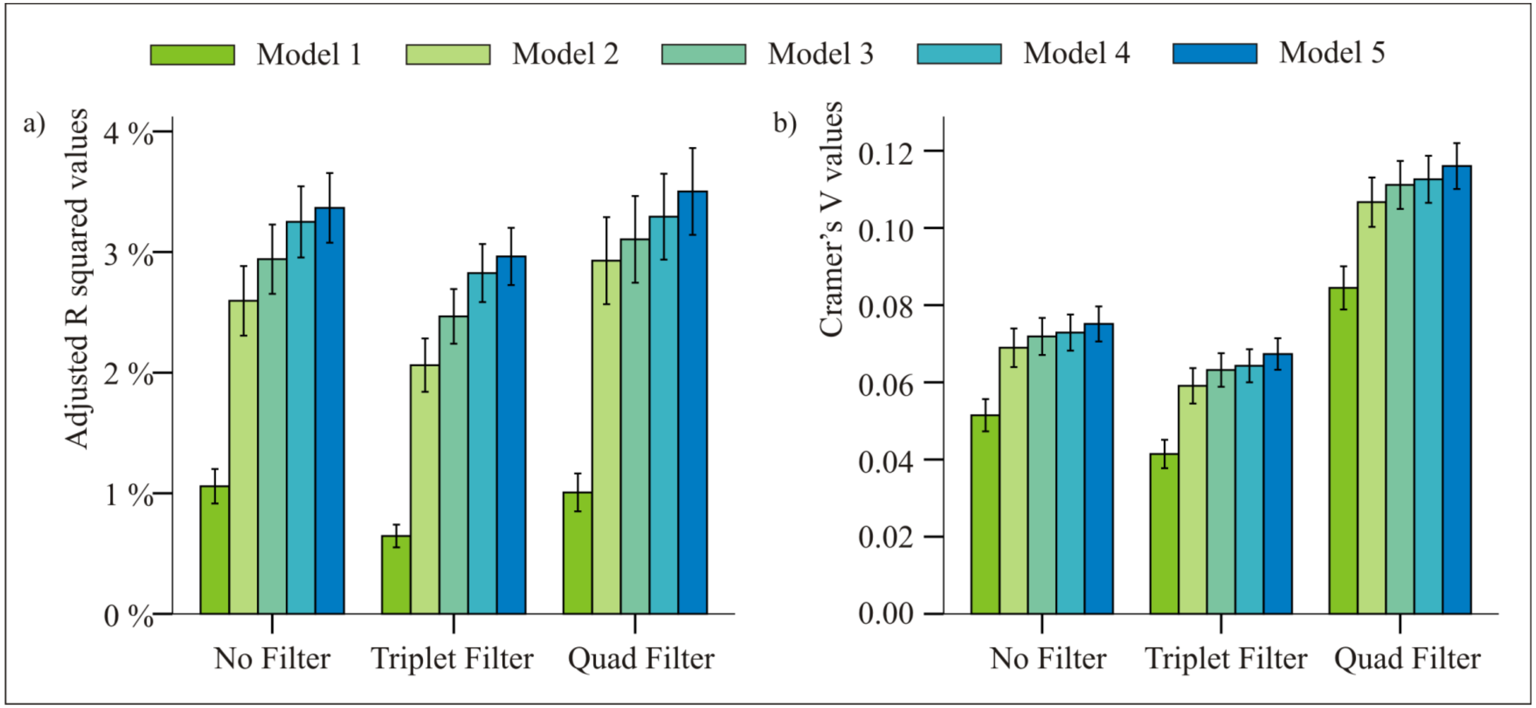


Figure V/6. Goodness of fit of the different models within each filtering method. a) Individual Adjusted R2 values based on reaction times. Each Model differed from all the other Models within each filtering method (all $p < 0.012$). b) Individual Cramer's V values based on error data. Each Model differed from all the others within each filtering method, except for the differences Model3 vs. Model4 (no filter $p = 0.166$, triplet filter $p = 0.359$, quad filter $p = 0.261$). Error bars are 95% confidence interval.

V/4.2.2. Errors

Warm-up trials were not analysed, neither were trials 6-8 in a block (since it is only from trial 9 that the first full ASRT-quad is reached). This way 9.4% of trials were excluded. Additionaly, only those data points were included that were preceeded by at least three correct responses in a row (this way it could be ensured that the sequence of buttonpresses corresponded to the intended combinations before a critical trial); this resulted in the removal of additional 13.9% of the remaining trials. When no filtering was applyied, an average of 2981 trials were analysed per participant; triplet filtering resulted in an average of 2610 trials, while quad filtering in an average of 1113 trials per parcitipant. The goodness of fit of the different models were than calculated in the form of Cramer's V values (data from the nine epochs were collapsed into a single category due to the small number of errors) separately for each filtering method. To compare the obtained Cramer V values, we run a FILTER TYPE (3 levels: No Filter, Triplet Filter, Quad Filter) x MODEL (5 levels: Model 1 - Model 5) Repeated Measures ANOVA. Sphericity was assessed with Mauchly's Test, and if this precondition was not met, degrees of freedom were adjusted with the Greenhouse-Geisser method. Bonferroni-corrected post hoc tests were performed whenever the omnibus ANOVA showed significant main effects or interactions. Partial eta squared effect sizes are reported in line with significant main effects or interactions in the ANOVA.

The main effect of FILTER TYPE was significant, $F(1.472, 263.495) = 489.885$, MSE = 0.002, $p < 0.001$, $\eta_p^2 = 0.732$, indicating that, on average, the goodness of fits differed as a function of the filter used. These differences were better captured with a quadratic model than with a linear one (both $p < 0.001$, but the effect size for the quadratic model is $\eta_p^2 = 0.828$, while it is $\eta_p^2 = 0.676$ for the linear model). Bonferroni corrected post hoc tests revealed that means of Cramer V values were highest with the Quad Filter and lowest with the Triplet Filter, all contrasts being significant ($p < 0.001$). The main effect of MODEL was also significant, $F(1.598, 286.124) = 281.264$, $MSE = 0.001$, $p < 0.001$, $\eta_p^2 = 0.611$, indicating that model goodness of fits differed as a function of the Model used in the analysis. These differences were best explained with a linear model (as compared to quadratic, cubic or higher order models, $\eta_p^2 = 0.688$ for the linear model, and 0.503, 0.470 and 0.047 for the higher order models, respecitvely); values grew monotonicaly from Model 1 to Model 5. Bonferroni corrected post hoc tests revealed that all paired comparisons were significant (all $p < 0.001$) except for the difference between Model3 and Model4 (p = 0.231). Finally, the interaction of FILTER TYPE x MODEL was also significant, $F(1.747, 312.721) = 40.517$, $MSE < 0.001$, $p < 0.001$, $\eta_p^2 = 0.185$, indicating that the monotonic growing of adjusted R squared values as a funtion of MODEL were not equivalent with the three filtering methods used. Bonferroni corrected post hoc tests revealed that each Model differed from all the others within each filtering method, except for the differences Model3 vs. Model4 (no filter p = 0.166, triplet filter p = 0.359, quad filter p = 0.261). The effect of the differing filters was also quite consistent with each Model, showing all filtering methods differed from the rest (all $p < 0.001$). The results are shown on **Fig. V/6/b**).

V/4.3. Comparison of the Filters

V/4.3.1. Mean Reaction Times and Error Percentages belonging to the Models' categories

To get a more sophisticated picture, we calculated the mean reaction times and error percentages broken down by the categories specified by the Models separately for each filtering method. This was done for each of the nine epochs for each participant, and then these nine values were averaged to yield a single value for each cell for each participant. We summarized the means of these mean reaction times and mean error percentages, standard deviations (SD) of these means and the coefficients of variations of these means (CV = SD/mean in %) (Supplementary Table **ST-V/5** in the Supplementary Materials).

Does filtering alter mean reaction times corresponding to the categories specified by a certain model? To answer this question, we ran Repeated Measures ANOVAs with FILTERING (no filter, triplet filter, quad filter) as an independent variable (and categories' means as dependent variables). Our results showed that all category means differed as a function of FILTERING (all $p < 0.001$, all $\eta_p^2 > 0.255$) except for the H2 and H2P categories (in Model4 and Model5, respectively; $p = 0.998$, $\eta_p^2 < 0.001$). In cases of significant omnibus ANOVAs, Bonferroni corrected post hoc tests were run. Triplet filtering, in contrast to no filtering, altered the mean reaction times in the *random* (R) category in Model 1 ($p < 0.001$), and low-frequency triplets' reaction times in Model 2-5 on a trend level ($p = 0.095$). In all these cases, means got lower. Quad filtering, on the other hand, increased means in all of the categories in each Model (both relative to no filtering and relative to triplet filtering, all $p < 0.001$), which indicates that, predominantly, „easy" combinations had been eliminated with this filter (see also **Fig. V/5** for a similar conclusion).

A very similar pattern emerged with Repeated Measures ANOVAs performed on the mean percentages of errors: a significant main effect of FILTERING was observed for each category of each Model (all $p < 0.001$, all $\eta_p^2 > 0.113$) except for the H2 and H2P categories of Model4 and Model5, respectively ($p = 0.412$, $\eta_p^2 = 0.005$). In cases of significant omnibus ANOVAs, Bonferroni corrected post hoc tests were run. Triplet filtering, in contrast to no filtering, altered (lowered) the mean error percentages in the *random* (R) category in Model 1 and low-frequency triplets' reaction times in Model 2-5 (all $p < 0.001$). Quad filtering, on the other hand, increased mean percentages of errors (both in contrast to no filtering and triplet filtering), and this increase was significant in all but one of the cases (all $p < 0.001$; except for the R category of Model1 where $p > 0.999$).

V/4.3.2. Learning Effects

Solely the fact that mean reaction times and error percentages are subject to change when not all data is included is not surprising, and, in itself, not very meaningful. The real question is whether *learning effects* are subject to change when we apply different filters (e.g. whether category means change in parallel or some are affected more than others, or in other directions than others, resulting in changed learning scores as well). To answer this question, we calculated all the possible learning effects in the form of Cohen's d-s (RT data) and Cramer's V-s (error data) for each Model and each filtering method individually. In the case of reaction times, these effect sizes were calculated separately for the nine epochs and then averaged for each participant; in the case of errors, the data from the nine epochs were pooled for each participant (due to very low numbers of errors), and thus only one Cramer's V effect size was calculated per cell. SupplementaryTable **ST-V/6** (in the Supplementary Materials) summarizes the means of the individual effect sizes, the SD of these means and the CV of these means. Positive values indicate that the difference between categories showed the expected pattern, while negative values indicate the opposite (e.g. the easier/more predictable trials being responded to slower or less accurately), usually an unexpected result. As an exception, the contrasts H1 vs. H2 in Model 4 and H1P vs. H2P in Model 5 might result in negative values if joint probabilty learning is higher/more dominant than conditional probability learning, thus they couldn't automatically be considered *false negatives*.

V/4.3.2.1. Specific learning effects based on reaction times

To assess whether the filtering method had an effect on individual effect sizes, we first run Repeated Measures ANOVA-s on the Cohen's d values obtained for all the possible learning measures of the five Models with FILTER (no filter, triplet filter, quad filter) as an independent variable. Filter had an effect in all cases (all $p < 0.001$, all $\eta_p^2 > 0.164$), except for the *pattern learning* measure of Model5 (H1P vs. H1R), which remained unchanged ($p = 0.626$, $\eta_p^2 = 0.003$). In cases of significant omnibus ANOVAs, Bonferroni corrected post hoc tests were run. Triplet filtering (in contrast to no filtering) left some of the learning measures unaffected (those that are based solely on high-frequency triplets; i.e. HR vs. HP in Model 3; H1 vs. H2 in Model4; H1P vs. H2P and H1R vs. H1P in Model5). In all the remaining cases individual effect sizes decreased as a result of triplet filtering (all $p < 0.048$). Quad filtering, on the other hand, resulted in mixed effects. It increased effect sizes obtained in the simple models Model1 and Model2 (P vs. R; H vs. L – both of which could be considered quite mixed effects), and in the more elaborated Models (3-5) it increased those effects that depicted higher order statistical learning measures (HR vs. HP in Model3, H1 vs. H2 in Model4; and H1P vs. H2P in Model5; all $p < 0.001$). It is worth noting that some of these values not only increased but reversed their direction when applying the quad filter, leading to qualitatively different conclusions about learning. At the same time when higher order statistical learning effects increased, effects that depict the relatively pure measure of triplet level statistical learning (LR vs. HR in Model3; L vs. H1 in Model4; and LR vs. H1R in Model5) decreased in contrast to other filtering options when applying the quad filter (all $p < 0.001$). Learning measures depicting *maximum learning* (HP vs. LR in Model3, H2 vs. L in Model4 and H2P vs. LR in Model5) showed a significant increase (all $p < 0.001$), but its worth remembering that this is admittedly a mixed effect showing the summarized changes in different measures of each Model.

V/4.3.2.2. Specific learning effects based on errors

Error data showed a similar (although not identical) pattern. In this case, the dependent variables were the Cramer's V values, the independent variable was again the way of filtering data (FILTER: no filter, triplet filter, quad filter). The main effect of FILTER was significant in most of the cases (all p < 0.006, all $\eta_p^2 > 0.037$) except for the *triplet learning effects* HR vs. LR in Model3 and its equivalent H1R vs. LR in Model5 ($p = 0.584$, $\eta_p^2 = 0.002$) and the *pattern learning* effect of Model5 (H1P vs. H1R, $p = 0.335$, $\eta_p^2 = 0.006$). In cases of significant omnibus ANOVAs, Bonferroni corrected post hoc tests were run. Triplet filtering, in contrast to no filtering, decreased effect sizes in all cases except for those that are by definition unaffected by triplet filtering (all $p < 0.002$). Quad filtering, on the other hand, had a differential effect on learning measures depending on what kind of learning they depicted. Again, learning scores associated with higher order statistical learning (HR vs. HP in Model3; H1 vs. H2 in Model4; H1P vs. H2P in Model5) showed an increase when the quad filter was applied (all $p < 0.001$) – and, again, these effects reversed their directions from being on average negative to being on average positive. At the same time, the learning measure depicting pure triplet level learning (L vs H1 in Model4) decreased (all $p < 0.004$). Effect sizes of the relatively simple models Model1 and Model2 (showing quite mixed effects) also decreased when the quad filter was applied, but this decrease was only significant relative to the *no filter* condition (both $p < 0.002$) but not relative to the triplet filtering condition (both $p > 0.999$). *Maximum Learning* effects of Model3, Model4, and Model5 also showed somewhat mixed effects. When the quad filter was applied, there was an increase in effect sizes relative to triplet filtering in the case of Model4 and Model5 (both $p < 0.001$). None of the remaining contrasts approached significance (all $p > 0.177$). The mixed effects of *Maximized Learning* measures are not of a surprise since these reflect the sum of the positive and negative changes of individual learning measures of each Model.

V/4.3.3. Variability

It is crucial to be able to detect individual differences (i.e. between-subjects variability) with any task. One way of doing so is to assess the variability of individual learning scores (e.g. standard deviation, SD). If variability decreases with a stricter filtering, it might indicate that some of the differences seen in previous studies are attributable to differences in pre-existing biases to certain movement combinations. If, however, variability is increased when using a stricter filtering, it might mean that some of the variability in implicit learning capabilities were previously masked due to the systematic noise attributable to pre-existing tendencies. Importantly, this kind of variability may also increase as a result of increased noise (i.e. less precise estimates) on the individual level when a stricter filtering comes with a smaller number of trials being analysed.

V/4.3.3.1. How does filtering affect the variability of the learning scores?

To test the homogeneity of variances, Levene-test was applied on individual learning scores with FILTERING (No Filter, Triplet Filter, Quad Filter) as an independent variable (for this particular analysis treated as a between-subjects variable). According to the test, filtering had a significant effect on variances in most of the cases ($p < 0.032$); the exceptions were Higher Order Learning score of Model 3; $F(2, 537) = 1.683$, $p = 0.187$; and the Quad Learning score in Model 4; $F(2, 537) = 1.977$, $p = 0.140$. To unpack the observed differences, we also computed the Levene test in pairs (No Filtering vs. Triplet Filtering; No Filtering vs. Quad Filtering and Triplet Filtering vs. Quad Filtering). The results are shown in **Table V/1.**

Table V/1. Between-subjects variability of individual learning scores as a function of filtering

		Using the Triplet Filter (as opposed to No Filter)			Using the Quad Filter (as opposed to No Filter)			Using the Quad Filter (as opposed to Triplet Filter)		
		Variance change ↓ decrease ↑ increase	Levene test F value	Levene test p value	Variance change ↓ decrease ↑ increase	Levene test F value	Levene test p value	Variance change ↓ decrease ↑ increase	Levene test F value	Levene test p value
M1	Trial Type Effect	↓	**11.019**	**0.001***	↓	1.064	0.303	↑	**4.913**	**0.027***
M2	Sequence Specific L.	↓	**4.892**	**0.028***	↑	**9.386**	**0.002***	↑	**26.676**	**<0.001***
M3	Pure Statistical Learning	↓	0.919	0.338	↑	**7.412**	**0.007***	↑	**13.095**	**<0.001***
	Higher Order Seq. Learn.	=	0.000	> 0.999	↑	2.451	0.118	↑	2.451	0.118
	Maximized Learning	↓	**5.433**	**0.020***	↑	**7.258**	**0.007***	↑	**23.779**	**<0.001***
M4	Triplet Learn. (+ Pattern L.)	↓	1.131	0.288	↑	**7.587**	**0.006***	↑	**14.805**	**<0.001***
	Quad Learn (+ Pattern L.)	=	0.000	> 0.999	↑	**4.872**	**0.028***	↑	**4.872**	**0.028***
	Maximized Learning	↓	**6.465**	**0.011***	↑	**7.443**	**0.007***	↑	**24.574**	**<0.001***
M5	Triplet Learning	↓	0.919	0.338	↑	**7.412**	**0.007***	↑	**13.095**	**<0.001***
	Pattern Learning	=	0.000	> 0.999	↑	**9.822**	**0.002***	↑	**9.822**	**0.002***
	Quad Learning	=	0.000	> 0.999	↑	***2.734***	***0.099***+	↑	***2.734***	***0.099***+
	Maximized Learning	↓	**6.465**	**0.011***	↑	**7.443**	**0.007***	↑	**24.574**	**<0.001***

M1-M5: Model 1 – Model 5

* significant difference, p < .05

+ tendency towars significance, p < .10

V/4.3.3.2. Is higher variability caused by less precise estimates?

Standard deviations (SD) of the learning scores and the coefficients of variations (CV) are shown in Supplementary Table **ST-V/6** in the Supplementary Materials. The fact that CVs show a similar pattern to SDs indicates that higher standard deviations are not only a straightforward consequence of higher means. Moreover, in the case of triplet learning measures, a typical result is that decreased means are associated with increased standard deviations (e.g. Model 2 L-H, Model 3 LR-HR, Model 4 L-H1, Model 5 LR-H1R). This pattern of results was not only consistent across reaction times and error data (see Supplementary Table **ST-V/6** in the Supplementary Materials), but also across the six ASRT sequences when assessed separately (SD and CV in Supplementary Tables **ST-V/8** and **ST-V/9** in the Supplementary Materials; and **ST-V/11** and **ST-V/12** in the Supplementary Materials; while Supplementary Tables **ST-V/7** and **ST-V/10** shows the means of individual effect sized broken down by the six possible ASRT sequences. Note that CV values could be inflated in cases when means approach zero).

Why did these differences arise? In an optimistic scenario, they are the result of quad filtering making it possible for us to detect previously undetectable (masked) individual differences in learning capabilities. In a pessimistic case, however, higher variability stems from other sources. For example, it may be a consequence of noisier estimates on the individual level since the number of analyzed trials is smallest with quad filtering. In order to check up on this possibility, we calculated the within-subject standard deviations (SD) and coefficients of variations (CV) of reaction times for each epoch and each statistical category for each participant. We averaged the values obtained for the nine epochs in order to get a single value for each category for each participant (see Supplementary Table **ST-V/13** in the Supplementary Materials). Note that we could not compute standard deviations (and CVs) for error data within individuals since accuracy is a binary data type (a particular press is either correct or incorrect).

We ran Repeated Measures ANOVAs on these SD values with FILTER (no filter, triplet filtering, quad filtering) as a within-subject factor. The main effect was significant every time (all $p < 0.001$, all $\eta_p^2 > 0.124$). To disentangle these omnibus effects, Bonferroni corrected post hoc tests were run. Triplet filtering (contrasted with no filtering) left some of the SD-s unaffected (since, by definition, triplet filtering does not affect high-frequency triplets). In all the remaining categories, SDs decreased as a result of triplet filtering (all $p < 0.012$). Quad filtering resulted in further decreases in all categories (even those unaffected by triplet filtering), all $p < 0.001$. This effect was quite consistent, as participants have shown the previously described pattern in (on average) 5-6 epochs out of 9 (see **Table V/2**).

Table V/2. Within-subject variability of the estimates (that the learning scores are based on) as a function of filtering.

Model	Category	**Mean Number of Epochs** in which the SD decreased with the use of a particular Filter for a given individual *(SD of the Mean)*			**Mean Number of Epochs** in which the CV decreased with the use of a particular Filter for a given individual *(SD of the Mean)*		
		TF compared to NF	**QF** compared to NF	**QF** compared to TF	**TF** compared to NF	**QF** compared to NF	**QF** compared to TF
M1	R	**4.66** *(2.03)*	**5.88** *(1.97)*	**5.61** *(2.03)*	**4.23** *(1.97)*	**6.27** *(2.08)*	**6.17** *(2.22)*
	P	*No Diff.*	**5.69** *(1.97)*	**5.69** *(1.97)*	*No Diff.*	**5.82** *(2.11)*	**5.82** *(2.11)*
M2	L	**4.50** *(1.97)*	**5.12** *(2.09)*	**5.03** *(2.11)*	**4.53** *(1.92)*	**5.74** *(2.18)*	**5.63** *(2.16)*
	H	*No Diff.*	**5.83** *(2.02)*	**5.83** *(2.02)*	*No Diff.*	**6.18** *(2.10)*	**6.18** *(2.10)*
M3	LR	**4.50** *(1.97)*	**5.12** *(2.09)*	**5.03** *(2.11)*	**4.53** *(1.92)*	**5.74** *(2.18)*	**5.63** *(2.16)*
	HR	*No Diff.*	**5.53** *(2.16)*	**5.53** *(2.16)*	*No Diff.*	**6.47** *(2.14)*	**6.47** *(2.14)*
	HP	*No Diff.*	**5.69** *(1.97)*	**5.69** *(1.97)*	*No Diff.*	**5.82** *(2.11)*	**5.82** *(2.11)*
M4	L	**4.50** *(1.97)*	**5.12** *(2.09)*	**5.03** *(2.11)*	**4.53** *(1.92)*	**5.74** *(2.18)*	**5.63** *(2.16)*
	H1	*No Diff.*	**5.88** *(2.37)*	**5.88** *(2.37)*	*No Diff.*	**6.72** *(2.24)*	**6.72** *(2.24)*
	H2	*No Diff.*	**5.38** *(1.83)*	**5.38** *(1.83)*	*No Diff.*	**5.32** *(2.03)*	**5.32** *(2.03)*
M5	LR	**4.50** *(1.97)*	**5.12** *(2.09)*	**5.03** *(2.11)*	**4.53** (1.92)	**5.74** *(2.18)*	**5.63** *(2.16)*
	H1R	*No Diff.*	**5.53** *(2.16)*	**5.53** *(2.16)*	*No Diff.*	**6.47** *(2.14)*	**6.47** *(2.14)*
	H1P	*No Diff.*	**5.54** *(2.20)*	**5.54** *(2.20)*	*No Diff.*	**6.37** *(2.13)*	**6.37** *(2.13)*
	H2P	*No Diff.*	**5.38** *(1.83)*	**5.38** *(1.83)*	*No Diff.*	**5.32** *(2.03)*	**5.32** *(2.03)*

M1-M5: Model 1 – Model 5. NF: No Filter, TF: Triplet Filter, QF: Quad Filter. No Diff: Filtering did not affect the category, thus no difference could be observed

The decrease of standard deviations should not be attributed to decreased means exclusively since some of the means actually increased with the use of filters (see Supplementary Table **ST-V/5** in the Supplementary Materials). To corroborate this thought, we conducted the previously described ANOVAs on CVs, too. The main effect of FILTERING was, again, significant in all cases, all $p < 0.001$, all $\eta_p^2 > 0.189$. Bonferroni corrected post hoc tests showed that triplet filtering has left these values largely unaffected (except for the L and LR categories of Model2, Model3, Model4, and Model5; a slight decrease on a trend level, $p = 0.090$). Quad filtering, on the other hand, decreased CV-s in all cases (all $p < 0.001$), indicating that, in this case, standard deviations decreased more than the means. This result clearly indicates that higher individual differences are not caused by noisier estimates of mean reaction times on the individual level. Importantly, this pattern was not specific to a subset of the ASRT sequences, as it was observed in all of them (see Supplementary Tables **ST-V/14** and **ST-V/15** in the Supplementary Materials).

V/4.3.3.3. Does higher variability go in hand with lower reliability?

Reliability assesses whether the outcome of a test would be similar when repeated. Unsystematic noise decreases reliability, while systematic patterns in variation increase it. Thus, the higher the systematic variation in our data (relative to unsystematic noise), the higher the reliability indices will be. The problem here is that pre-existing biases – that we aim to reduce with more strict filtering methods – may actually introduce systematic variability rather than unsystematic noise. That being said, it is entirely possible that by reducing the variability that is attributable to pre-existing biases, reliability indices drop; and this is exactly what we have found.

We calculated split-half reliability of all of the measures by randomly assigning each keypress to one of two categories; the individual effect sizes were then computed for both sets, and the correlation of the two values was computed. In the case of reaction times, individual Cohen's d-s were calculated for each Session (Epoch 1-3, Epoch 4-6 and Epoch 7-9) rather than for each Epoch separately in order to compensate for the low number of trials per epoch when the data is split in two, and the three values were then averaged to yield a single effect size for both sets. In the case of accuracy, a single Cramer's V was calculated for Epochs 1-9 for both sets (data from the nine epochs collapsed due to low overall error rates). Split-half reliabilities for reaction times are shown in **Table V/3**.

Table V/3. Split-half reliability of each of the possible learning scores (Models 1-5, all filtering types) based on reaction times.

We assigned each trial one of two possible codes, and the resulting two sets were analysed separately (thus learning, fatigue, etc. affected each set similarly). In the case of reaction times, learning scores were computed for each Session (epochs 1-3, epochs 4-6 and epochs 7-9), and then averaged. In the case of accuracy, a single Cramer's V was calculated for Epochs 1-9 for both sets (data from the nine epochs collapsed due to low overall error rates). The correlation between the two subsets is shown in the table (Pearson correlation coefficients).

		Reaction Times			Accuracy		
		No Filter	Triplet Filter	Quad Filter	No Filter	Triplet Filter	Quad Filter
M1	**R-P** *TrialType effect*	.770**	.545**	.365**	.514**	.354**	.267**
M2	**L-H** *Seq. Spec. L.*	.843**	.713**	.614**	.576**	.459**	.416**
M3	**LR-HR** *Pure Stat. L.*	.691**	.630**	.556**	.356**	.336**	.270**
	HR-HP *Higher Ord. L.*	.366**	.366**	.236**	.137	.137	.025
	LR-HP *Max. Learning*	.835**	.687**	.581**	.573**	.446**	.400**
M4	**L-H1** *Triplet L.+ P. L.*	.788**	.707**	.603**	.489**	.414**	.347**
	H1-H2 *Quad L.+ P. L.*	.595**	.595**	.374**	.176*	.176*	.008
	L-H2 *Max.Learning*	.828**	.680**	.538**	.535**	.423**	.337**
M5	**LR-H1R** *Triplet Learning*	.691**	.630**	.556**	.356**	.336**	.270**
	H1R-H1P *Pattern Learning*	.090	.090	.226**	.025	.025	.041
	H1P-H2P *Quad Learning*	.477**	.477**	.363**	.078	.078	.022
	LR-H2P *Max. Learning*	.828**	.680**	.538**	.535**	.423**	.337**

* p < .05, ** p < .01

As can be seen, reliability indices dropped substantially when using the quad filter. This may be attributable to the possibility that pre-existing biases are a form of a systematic artifact (rather than noise), as noted earlier. Conversely, it is also possible that the drop is attributable to increased levels of noise (since fewer trials are analyzed with stricter filtering). However, even these lower indices are not surprising (at least for reaction times data), as the reliability of implicit learning measures is often low (Lebel & Paunonen, 2011), and difference scores might also have lower reliabilities than the components they are derived from, proportionally to the correlation between the original components (Edwards, 2001). As a comparison, Kaufman et al. used a probabilistic SRT task that is in many aspects similar to the ASRT used in our study, and the split-half reliability of the RT difference score was 0.33 (Kaufman et al., 2010). It is worth emphasizing that reliability is a different concept from validity, i.e. whether we measure what we aim to measure. For methodological reasons, we argue that quad filtering makes the ASRT a less reliable but more valid task for assessing implicit learning.

V/4.4. New insights - What is being learned ASRT task?

After arguing for the use of the newly proposed analysis methods (Model 5 and Quad Filtering), we would like to briefly review its results. The focus here is not whether results differ from those gained using the typical analysis (they do, see above), but purely descriptively: is there evidence for quad learning, triplet learning and pattern learning on the group level? What percentage of participants show learning on these measures?

In the case of reaction times, the first question was assessed with a Repeated Measures ANOVA, with Epochs (1-9) as an independent variable and individual Cohen's d values (in each epoch) as the dependent variable. This way both the overall learning and its time course were assessed. In the case of error data, time course couldn't be taken into account due to the low overall number of errors. Thus, in this case, a single Cramer's V value was calculated for each participant (data of Epoch 1-9 collapsed into a single category), and these values were compared to zero using a one-sample t-test.

To assess the second question, we identified participants with individual Cohen's d values and Cramer's V values exceeding the limit of 0.2 and 0.05, respectively. We also quantified the percentage of participants whose effect sizes exceeded these limits but the difference was in the unexpected direction (presumably false negatives). If the ratio of false positives is similar to the ratio of false negatives, then the percentage of true positives could be gauged by subtracting the (false) negatives from the positives. In the case of reaction times, we also calculated the percentage of *reliable learners* by identifying participants who had shown at least a small learning (Cohen's d > 0.2) in at least 5 (out of 9) epochs. Such a detailed analysis was not possible in the case of accuracy data due to the low overall error rates.

For error data, Individual effect sizes (Cramer's V-s) were calculated for the collapsed data of the nine epochs because of the low number of errors (consequently, the time-course of learning could not be examined). Since these effect sizes represent some kind of an individual average, maximum values are expected to be lower than RT-related effect sizes. For this reason, we decided to use Cramer's V > 0.05 as an indication of small learning (instead of the conventional Cramer's V > 0.10). Our results are presented in **Table V/4**.

Table V/4. Average Learning Scores and the percentage of participants who learned particular types of information.

		Triplet Learn.	Quad Learn.	Pattern Learn.	Max Learn.
RT data	Overall learning (descriptive statistics, and the ANOVA's intercept)	M = 0.350, SEM = 0.014; F(1,179) = 597.78, MSE = 0.331, p < 0.001, η_p^2 = 0.770	M = 0.056 , SEM = 0.010; F(1,179) = 30.79, MSE = 0.163, p < 0.001, η_p^2 = 0.147	M = -0.028 , SEM = 0.009; F(1, 179) = 10.72, MSE = 0.122, p = 0.001, η_p^2 = 0.057	M = 0.376, SEM = 0.014; F(1, 179) = 767.92, MSE = 0.298, p < 0.001, η_p^2 = 0.811
	ANOVA's main effect of EPOCH	F(8, 1432) = 23.03, MSE = 0.108, p < 0.001, η_p^2 = 0.114	F(8, 1432) = 2.09, MSE = 0.104, p = 0.033, η_p^2 = 0.012	F(8, 1432) = 1.27, MSE = 0.097, p = 0.251, η_p^2 = 0.007	F(7, 1333) = 32.77, MSE = 0.097, p < 0.001, η_p^2 = 0.155
	% of positive learners (% of reliably positive learners)	76.77% (74.44%)	12.22% (13.33%)	5.00% (5.56%)	87.22% (83.89%)
	% of negative learners (% of reliably negative learners)	0.00% (0.00%)	4.44% (5.00%)	4.44% (10.56%)	0.00% (0.00%)
	% of true learners (% of reliably true learners)	76.77% (74.44%)	7.78% (8.33%)	0.56% (-5.00%)	87.22% (83.89%)
Accuracy data	Overall learning (descriptive statistics, and the results of the t-test)	M = 0.057, SEM = 0.004, t(179) = 13.766, p < 0.001, *Cohen's d* = 1.026	M = 0.012, SEM = 0.003, t(178) = 3.500, p = 0.001, *Cohen's d* = 0.262	M = 0.001, SEM = 0.03, t(178) = 0.233, p = 0.816, *Cohen's d* = 0.018	M = 0.069, SEM = 0.004, t(179) = 16.136, p < 0.001, *Cohen's d* = 1.202
	% of positive learners	53.33%	22.35%	14.53%	63.89%
	% of negative learners	3.33%	9.50%	12.29%	1.11%
	% of true learners	50.00%	12.85%	2.24%	62.78%

Lastly, we wanted to assess how these results relate to those obtained with the typical analysis methods (Model 1, No Filter; and Model 2 and Model 3, Triplet Filter). Supplementary Figure **SF-V/2** in the Supplementary Materials shows mean effect sizes of various learning scores (obtained for reaction times); **SF-V/3** in the Supplementary Materials shows the percentage of positive and negative learners (broken down by epochs) obtained for reaction times data; and **SF-V/4** in the Supplementary Materials shows the percentage of positive and negative learners obtained for error data. We also identified reliably positive learners (based on reaction times) for earlier Models, and calculated Phi coefficients between pairs of learning scores of different Models to assess whether there is a correspondence of who is considered a reliable learner with the different analysis methods. The results showed that the correspondence is low (the highest Phi value being around 0.4; see Supplementary Table **SF-V/16** in the Supplementary Materials.

V/5. Discussion

The ASRT (J. H. Howard & Howard, 1997) task is a visuomotor sequence learning task designed to measure implicit learning and memory. In this paper, we discussed in detail the many possible information types that could be learned (such as pattern learning and different levels of joint frequency learning and conditional probability learning), and our concerns that these types of learning are not sufficiently differentiated by the currently used analysis methods. Moreover, as we have shown, the learning measures that are typically extracted from data might be biased by pre-existing tendencies to certain stimulus combinations, indicating that the ASRT does not measure (only) what it supposed to. We provided a presentation of how different analysis methods and filtering methods result in different levels of artifacts and biases, a hopefully practical aid for the (re)interpretation of the results obtained with the task. We also proposed new analysis methods (with a somewhat new terminology) and a filtering method that eliminates at least some of the biases discussed so far and thus can be used in future studies (or for reanalyzing already existing datasets).

V/5.1. Are the new analysis methods better?

In the second section of the paper, we compared the goodness of fit of models that are the basis of the different analysis methods (models already in use, and those that we proposed in the current paper). Our results showed that more elaborate models have a better fit indicating that there is a benefit of using these (newly proposed) methods instead of the typically used ones. We also looked at the specific effects that could be extracted from data (more elaborate models having more and purer measures), and how different filtering methods alter the effect sizes of these measures on the individual level. Our result indicated that different filtering sometimes leads to both quantitatively and qualitatively different conclusions. Importantly, quad filter did not only affect the specific effect sizes but also increased the individual variability that could be detected with the task. And variability is crucial – if everyone seems to show the same performance, there is no need to measure it.

V/5.2. What did the new analyses reveal?

With the usual triplet filter, we replicated the findings of Song et al. (2007a) and Németh et al. (Nemeth, Janacsek, & Fiser, 2013) who demonstrated that RH trials were (at least in the first few epochs of the task) reacted to faster and more accurately than PH trials. This paradoxical pattern of results still remained when H1 vs. H2 trials were compared instead of PH and RH trials – in the latter case, this could have been interpreted as a reflection of higher *joint probability learning* (contrasted with *conditional probability* learning) on the group level. However, using the quad filter, the paradoxical result disappeared, and even turned into its contrary (for a similar result see (Song et al., 2007a), indicating that it was *conditional probability learning* that dominated. We here thus argue that the paradoxical pattern of results (RH performance > PH performance) is an artifact attributable to pre-existing biases on the quad level. This reasoning might be at odds with the fact that the paradoxical result was temporary in Song et. al (Song et al., 2007a). However, Soetens, Boer, & Hueting (1985) found that practice actually reduces sequential effects (i.e. pre-existing tendencies), while it is reasonable to assume that statistical learning would either stagnate or increase with practice.

This is an important finding since the terminology frequently found in ASRT studies implies joint frequency learning (e.g. the terms „low-frequency triplet", „high-frequency triplet", etc.) and not conditional probability learning. Such terminology might mislead researchers to have the wrong focus when complying stimulus material – even with a conservative viewpoint, both factors should be considered. Beside practical considerations, these results also add to the theoretical debate whether implicit learning is based on learning chunks (which in this case corresponds to joint probability learning, e.g. the relative frequency of different combinations) or statistical computations (conditional probability learning) (Perruchet & Pacton, 2006).

Our new analysis methods revealed a dissociation between pattern learning and quad learning (which were confounded in previously quantified measures, i.e. *higher order learning*). On the group level, there was a significant quad learning effect (quantified as the difference between H1P and H2P trials), although smaller in magnitude than triplet level learning (quantified as the difference between H1R and LR trials); at the same time no pattern learning was observed (H1P vs. H1R) following this amount of practice. Moreover, H1P trials were slightly slower than H1R trials, which was unexpected. This new paradox might reflect pre-existing biases on the level of quints which were not controlled for in this study.

Besides these results observed on a group level, we also quantified the percentage of participants who showed at least a „small" learning effect (Cohen d > 0.2 in the cases of reaction times). From these percentages, we subtracted the percentage of participants showing similar effect sizes in the opposite directions (e.g. low probability trials being reacted to faster than high probability trials). The resulting number could be thought of as reflecting true learners (i.e. true positives without false positives). Strikingly, the percentage of participants showing at least a small triplet learning effect grew from ~25% to ~ 70% as learning progressed from epoch 1 to epoch 9 (interestingly, Parshina, Obeid, Che, Ricker, & Brooks (2018), with a somewhat different methodology and using Model 2 as a basis of their analysis, found that 64.9% of their participants showed triplet learning on the ASRT task, corroborating the fact that approximately 1/3 of people fail to exhibit such learning). At the same time, the percentage of participants showing true pattern learning remained around zero throughout. Quad learning was observed for ~5% to ~20% of participants.There was fluctuation in these numbers but no gradual increase as learning progressed(a gradual increase could be observed in two out of three Sessions, though; see the Methods section. Although highly speculative, it is possible that at the beginning of Sessions participants are less tired, and their reaction times are faster and less variable, making it hard to detect such subtle effects). It is also possible that both joint frequency learning and trial probability learning occur, but since they have opposite effects on reaction times, it manifests as very low (near-zero) quad learning effect. The latter possibility should be assessed by a sequence learning task in which the two kinds of probabilities are systematically varied.

In the case of accuracy, small effect sizes were defined as a Cramer V value > 0.05. In this case, much smaller percentages of true learners were observed - the percentage of true pattern learners was ~2%, the percentage of true quad learners ~15% and the percentage of true triplet learners ~50%. Importantly, accuracy data has not provided information regarding the time course of learning.

It was interesting to see that quad learning was significant on the group level but the number of quad learners did not increase over the nine epochs. This result might indicate that the observed learning is some kind of artifact (e.g. preexisting biases that were not controlled for, e.g. on the level of quints), or that a few participants were able to extract quad level information from the very beginning of learning, while, for others, 45 blocks of ASRT was not enough for this. The latter interpretation is supported by data showing that higher order learning occurs slowly over many sessions, as in (J. H. Howard & Howard, 1997; Song et al., 2007a) although the differences seen in these studies are the mix of quad learning effects, pattern learning effects and pre-existing biases; while the support for the former interpretation comes from the fact that *pattern learning* was *negative* throughout the task which can only be attributed to pre-existing biases (since pattern trials have, on average, higher conditional probabilities/joint frequencies than random trials, thus they should be faster anyway). If such pre-existing biases affect the *pattern learning* measure this way, they could also be affecting the quad learning measure (but remember that quad learning is higher than pattern learning, irrespective of the sign, possibly suggesting true learning in addition to such biases). Future studies need to clarify the issue.

The finding that no detectable pattern learning occurred during these 45 blocks of ASRT is an important one, too. It suggests that the ASRT task is a primarily statistical learning task (and not a pattern learning one), at least after this amount of practice. If future research corroborates this finding even after extended practice with the task, then this measure could be used as an examination whether the hidden sequence remained undetected by participants, complementing the sorting tasks, e.g. (A. Destrebecqz & Cleeremans, 2001), and production tasks, e.g. (Song et al., 2008), usually utilized to assess explicit knowlegde. For explicit learners, i.e. who become aware of the hidden sequence and thus are able to consciously anticipate pattern trials but not random trials, this measure should have a positive value - as observed in the explicit variants of the ASRT task when assessing the *higher order learning* measure (Nemeth, Janacsek, & Fiser, 2013; Song et al., 2007a). For implicit learners, on the other hand, the value of the measure should remain around zero.

Finally, we would like to draw attention to our way of quantifying individual learning as standardized effect sizes (Cohen's d and Cramer's V values) instead of just raw differences. This way the comparison of groups having different overall reaction times could be fairer; it's easier to identify people who truly learned or at least the percentage of true learners (e.g. by selecting participants showing at least a small effect size throughout the task, e.g. Cohen's $d > 0.2$). Due to increased individual variability, we could have missed the overall benefits of the newly proposed method looking only at group level effect sizes (e.g. when computing Cohen's d, a higher difference score increases the effect size but higher variability decreases it), so it might have looked as if our method have no benefit over the typically used methods).

In sum, we believe that the analysis and filtering methods that we proposed have several advantages over the typically used methods. First, more types and more levels of learning can be detected, and thus it becomes possible to differentiate between people with similar overall learning scores but different learning profiles. Also, if two groups differ in overall scores, it becomes possible to disentangle which learning type causes the overall difference, and, in the long run, may help us build better models of implicit learning. Second, more strict filtering results in purer measures of learning and more expressed individual differences. The fact that individual variability is higher using the stricter filtering shows that some of the variability had been masked with typical filtering methods, which in turn questions the results showing weak or no correlation between statistical learning on the ASRT task and other measures, e.g. (Stark-Inbar et al., 2016; Parshina et al., 2018).

V/5.3. Limitations and future directions

As a limitation for the newly proposed filtering method (but also for the typically used triplet filtering), it is important to note that these filters work on the group level when the six ASRT sequences are given to participants in the same proportions, but they do not necessarily work on the individual level (or when sequences are not counterbalanced across participants). The quads that are being compared within participants are similar in their abstract structure (consisting of *dcba, cbba* and *acba* quads), but some of the, say, *dcba* quads might still be easier than other exemplars of the same type. Also, as our results clearly indicated in the case of the pattern learning measure, the amount of biases may be modified even by the N-4th trials (i.e. quint level pre-existing biases). Furthermore, the direction and magnitude of such biases may vary as a function of the ASRT sequence administered. Bearing this in mind, if the main focus of a study is measuring individual differences (rather than group differences and relatively bias-free average learning scores) than the same ASRT sequence should be given to all participants. Further studies should address which of the six sequences should be preferred (bearing the lowest amount of pre-existing biases), and they should clarify the relationship between sensitivity to such biases and learning capacity in the ASRT task (if there is any).

We only moved one step forward from the usual filtering – quad level information being considered on the group level instead of triplet level information. We decided to stop here, and not to go deeper, for two reasons. First, there are no quints that are part of each of the six ASRT sequences, thus full counterbalancing isn't possible on the group level. Second, although some measures could be quantified in a more bias-free manner on the individual level if quint level information was considered, this would involve massive reduction in the number of trials being analyzed, and we speculated that the costs (increased noise) would be higher than the benefits (even smaller magnitude of pre-existing biases) of doing so.

Similarly, quint level statistical information could also have been considered but we decided not to go for it. The reason for this was that pre-existing biases could not have been ruled out even if we quantified this measure, and based on the small number of participants who actually learned quad level statistical information (~10-20%, see above), we speculated that quint level learning is probably close to zero after this amount of practice (thus this additional measure wouldn't be of much use). On the other hand, such comparisons could only be made based on a smaller number of trials being analyzed, which could, in turn, result in increased noise.

The ASRT data analyzed in this paper consists of 45 blocks of 80 trials for each participant. Both shorter, e.g. (Nemeth et al., 2009), and longer, e.g. (J. H. Howard & Howard, 1997; Schwartz et al., 2003), ASRT-s have been used in the past, but it is reasonable to assume that *higher order learning*, i.e. quad learning and/or pattern learning occurs over longer periods of time; see the results of (J. H. Howard & Howard, 1997; Song et al., 2007a). The analysis method proposed in this paper should be applied to data gained from extended practice to verify this hypothesis. This could reveal individual differences even in the dynamics of joint frequency vs. conditional probability learning, which would be an important step in understanding the nature of implicit statistical learning.

Future studies using longer ASRTs should clarify what types of learning are typical to humans and in what proportions. It should also be assessed whether these different types of learning correlate with each other and whether they rely on different brain structures (or on different kinds of connectivity/brain activity) (Kóbor et al., 2018).

V/5.4. General conclusion

We believe that the ASRT task is a great tool for measuring implicit sequence learning and memory – it might even be more promising than we ever thought. However, in order to get more out of it, we need to improve our analysis methods and take possible confounding factors more seriously. In this paper, we provided a possible solution to these problems. Our results point to the ASRT being primarily a statistical learning task (at least in the short term), where triplet learning occurs for most of the participants but quad learning is the privilege of fewer. We have also shown that these results depend strongly on the filter being used, and for methodological reasons, we suggest the usage of the Quad Filter in the future.

Importantly, the concerns outlined in this work are not specific to the ASRT task. It is reasonable to assume that pre-existing biases are present in other sequence learning tasks as well, such as the serial reaction time task (Nissen & Bullemer, 1987) or the finger tapping task (C. M. Walker, Walker, & Sunderland, 2003); and also that at least some people are sensitive to higher order statistics (e.g. third order statistical information), which should also be taken into consideration when complying stimulus material for all kinds of sequence learning tasks.

V/6. Data Availability Statement

All data is available without restriction in at **osf.io/q6u8k.**

V/7. Financial Disclosure

This research was supported by the National Brain Research Program (project 2017-1.2.1-NKP-2017-00002), the Hungarian Scientific Research Fund (NKFIH-OTKA K 128016, to D.N., NKFIH-OTKA PD 124148 to K.J.), IDEXLYON Fellowship (to D.N.), and a Janos Bolyai Research Fellowship of the Hungarian Academy of Sciences (to K. J.). This research was supported by the EU-funded Hungarian grant EFOP-3.6.1-16-2016-00008 (to E. Sz-H.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

V/8. Author Contributions

Emese Szegedi-Hallgató: Conceptualization, Data curation, Formal analysis, Funding acquisition, Resources, Software, Visualization, Writing – original draft, Writing – review & editing. Karolina Janacsek: Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Writing – review & editing. Dezso Nemeth: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

VI. GENERAL DISCUSSION

VI/1. Summary of findings: Study 1 – Study 4

The goal of the presented experiments was to learn about the nature of implicit statistical learning and the resulting representation; besides theoretical questions (such as whether statistical learning is domain-general; what factors influence it; etc) we were also interested in methodological questions (how to improve the tasks that measure it and/or how to make data analysis more efficient in identifying different aspects of learning).

VI/1.1. Study 1

In Study 1 the main question was whether the implicit knowledge in a visuomotor sequence learning task stems from response-related information (what button should be pressed next, which finger should be moved next) or also perceptual information (at which location should we expect the next stimulus). To answer the question, we modified the ASRT task in such a way that, although motor and perceptual information remained correlated, they were not identical anymore. By changing either the perceptual or the motor sequence, and by assessing its possible negative impact on the performance of participants, we could indirectly assess the role of particular information types in performing a sequence learning task. In this experiment, we were able to show statistically detectable implicit knowledge both when the motor or the perceptual information was changed, implying that both factors contribute to implicit learning in such paradigms (and that they do so to a similar degree).

VI/1.2. Study 2

In Study 2 we extended our findings of Study 1 by introducing an offline period before changing one of the information streams. We found that the transfer of perceptual information was weaker (but nevertheless significant) than the transfer of motor information both after 12 hours and 24 hours. Sleep had no impact on the consolidation of learning (as no differences were observed between groups of participants who slept and those who didn't sleep during the offline period).

VI/1.3. Study 3

In Study 3 we assessed the impact of interfering information from previous learning episodes (i.e. proactive interference) by analyzing performance on chunks of sequences that were either similarly frequent across the subsequent sessions or changed their frequency. Three groups were compared: the implicit-implicit group performed the ASRT implicitly in each session; the implicit-explicit group learned the first sequence implicitly but proceeded with the second explicitly; while the explicit-explicit group performed the explicit version of the ASRT in every session. Importantly, the explicitness of the ASRT meant that pattern trials were shown with a different color, and so the four-element long pattern was easy to spot, and pattern trials were easy to anticipate. However, since we were interested in statistical learning rather than pattern learning, only the intervening random elements were analyzed for all groups as a function of their statistical properties (probable trial or not so probable given the previous trials). Thus, we were interested in whether explicit knowledge about the sequence's structure has an impact on the implicit learning of the statistical structure (in other words, whether the implicit and explicit memory systems somehow cooperate; or whether the different kinds of operations under the different conditions affect implicit learning differently).

The most important result of the experiment was that „rewiring" (i.e. learning information that interfered with previous learning) was most successful in the Implicit-Explicit group and least successful in the Implicit-Implicit group. In the Implicit-Explicit group, learning of the new (interfering) information was as successful as learning the first information (where no interference could have been in play), while participants in the Implicit-Implicit group showed less rewiring (and slower update of no-longer-valid knowledge) than the other two groups. The analysis of errors (rather than reaction times) also corroborated these findings.

On the third day of the study, knowledge for both information sets (both sequences) was assessed. By evaluating performance on those chunks that were subject to interference, it was found that participants do better when the later-learned sequence was assessed, possibly indicating forgetting of the previously-valid skill (or it might be a retroactive interference). No group differences were observed.

VI/1.4. Study 4

One of the goals of Study 4 was to improve the analysis methods of the ASRT task in order to separate different levels of implicit statistical learning (e.g. triplet-level learning from quad-level learning) and the learning of the embedded pattern (i.e. pattern learning). We also proposed the use of a stricter data filtering to minimize the effect of pre-existing biases when calculating different learning scores.

The most important finding was that the proposed analysis method resulted in higher goodness of fit (quantified as adjusted R^2 values) than the previous methods, and the pattern of results differed qualitatively with the use of the stricter filter. For example, a learning score which was typically (paradoxically) negative with the typical analysis methods, became positive with the new filter. These results point out the effect of pre-existing biases in this paradigm, and the need to reanalyze previous findings to minimize the artifacts.

As a separate goal, we also assessed the psychometric properties of the ASRT task. Paradoxically (but not inexplicably) reliability of the learning scores decreased with the new analysis method. We attribute these changes to the fact that pre-existing biases introduced a systematic artifact to the previous analysis methods, and when this systematic pattern was eliminated, individual differences in statistical learning became unmasked, and higher variability resulted in poorer reliability scores. However, we argue that with the newly proposed analysis method, the ASRT also became a better tool in measuring implicit statistical learning (even if the scores are less reliable, they are more valid, i.e. less affected by artifacts).

VI/2. Discussion of findings

Our studies aimed to gain a better understanding of implicit statistical learning phenomena, and about the diversity of findings of seemingly similar processes. We also emphasized the need to assess the psychometric properties of the measures of implicit statistical learning, since many findings – or null findings – could be a result of some methodological rather than theoretical issues. In this Dissertation, we addressed many of these topics, such as modality specificity in implicit learning; the (in)dependence from other cognitive domains, types of statistics that could be learned, and methodological considerations in the measurement.

VI/2.1. Implicit learning – One or Many?

VI/2.1.1. Modality specificity

In Study 1 and Study 2 we found that both the visual and the motor sequence had been learned by participants, and although similar magnitudes of transfer have been observed for the two types of information when performance was assessed in a single session (without delay), the consolidation of the perceptual information seemed to be weaker than the consolidation of the motor information after a delay of 12 or 24 hours.

Importantly, the existence of perceptual statistical learning was inferred from the fact that – in spite of the interference caused by the change of the motor sequence – learning scores in the second phase was greater than zero. However, two further possibilities could have led to the observed results.

First, it is possible that no perceptual transfer occurred, only motor learning of the new (interfering) sequence. In Study 2, we quantified learning scores in the first two blocks of Session 1 (before the change in either the perceptual or the motor information had taken place), and in the first two blocks of Session 2 (i.e. immediately after the changes occurred). We found significant learning effects in the latter but not in the former case – therefore these results might indicate that the learning scores of Session 2 are transferred from Session 1, and not learned anew, since learning of the motor information was not this fast even without interfering information in Session 1.

Second, it is also possible that the transfer is not perceptual in nature, but motor transfer, since the two interfering motor sequences had a few similar chunks that were frequent throughout (a possibility we did not address at the time of publication). However, we analyzed the data and found similar results even when controlling for this possibility (*unpublished results*).

In sum, it has been shown that both kinds of learning occur, and consolidation is different for the two types of learning. These results support the notion of multiple implicit statistical learning submodules rather than a single system that is responsible for all kinds of implicit learning phenomena (in line with Emberson et al., 2011; Li et al., 2018; Walk & Conway, 2016).

VI/2.1.2. Independency from other cognitive abilities

In Study 3 we compared the implicit statistical learning abilities of participants who were either given information about the embedded sequence in the ASRT or not. Importantly, even participants performing the explicit version of the task could only anticipate pattern trials explicitly – any statistical learning effects detected on random trials were implicit nevertheless.

We found that the implicit statistical learning was similar across groups when learning the first sequence, thus knowledge about pattern trials did not help participants learn the statistical properties of the task when assessed on random trials. However, when a new sequence was introduced, and proactive interference from the first learning episode had to be overcome, the explicit groups outperformed the implicit group (or more precisely, the implicit group needed more time to overcome the interference than the explicit groups).

One possibility is that those receiving explicit information about every second trial had a different attitude, or were less bored, than the implicit groups, and they might have concentrated more even on random trials, leading to better performance. As an alternative, it is also possible that telling participants about the embedded structure somehow emphasized the statistical structure of the sequence, too. It has been found, for example, that the exaggeration of some features of speech acts as a perceptual catalyst whereby it helps infants discriminate between similar inputs (Karzon, 1985), and that infant-directed speech (e.g. motherese, which also exaggerates important aspects of speech) is a more effective signal for learning phonetic categories than adult-directed speech (de Boer & Kuhl, 2003). Our results could have been similarly caused by the emphasis that explicitly shown pattern trials created.

The fact that knowledge about the sequence in the ASRT somehow affected the implicit processes indicates that implicit and explicit processes are not independent (for a similar conclusion, see Boyd & Winstein, 2003; Arnaud Destrebecqz et al., 2005; Lagarde et al., 2002). Our results are in line with Boyd & Winstein (2003) who also found facilitation of implicit learning following explicit information in healthy adults – however, they used a deterministic 10 elements long sequence, and hence from the point when the sequence was explicitly stated, each element of the sequence could have been consciously anticipated. In our case, however, those trials where explicit anticipations could have been in play were eliminated. For this reason, we think that our results are stronger support for the notion of interacting systems, although the exact mechanism has to be identified.

VI/2.1.3. Type of statistics – Does it matter?

It has been known that humans are capable of both conditional probability learning and joint frequency learning (J. H. Howard et al., 2008; Thiessen et al., 2013; Thiessen, 2017) and it has been suggested that they are results of independent processes (Thiessen, 2017). In Study 4 we have shown that the ASRT task is adequate to distinguish between the two types of learning (or at least in showing which one dominates for a given individual), and that results point towards a greater dominance of conditional probability learning (13% vs. 5% of participants showing reliable pattern for conditional probability learning and joint frequency learning, respectively).

Furthermore, it has also been shown that humans are able to learn higher-order statistical structure (e.g. four consecutive trials – quads – or even higher levels) (Remillard, 2008, 2010, 2011), moreover, that learning of higher-order information can be selectively impaired (in dyslexia: W. Du & Kelly, 2013; J. H. Howard et al., 2006; in Parkinson's disease: Smith & McDowall, 2004; in Schizophrenia: Schwartz et al., 2003; with age: J. H. Howard et al., 2007; D. V. Howard et al., 2004; Feeney et al., 2002; J. H. Howard & Howard, 1997; Urry et al., 2018). In Study 4 we have shown that – although traditionally only the level of triplets is assessed – quad-level learning could also be quantified without any modification to the task, simply by applying a different analysis method (and this is independent from the question of joint versus conditional probabilities). By reanalyzing the huge amount of already existing ASRT datasets, we could get closer to understand the nature of quad learning, and whether triplet level learning and quad level learning are degrees of the same capability, or they dissociate within subjects. This could be assessed by verifying the correlation between the two in large samples; or by examining the learning curves, e.g. whether quad learning follows triplet level learning or it happens in parallel (for those who are sensitive to these statistics).

VI/2.2. The psychometric properties of the ASRT task

VI/2.2.1. Low reliability

Although the relatively low reliability of implicit compared to explicit measures has been acknowledged (Lebel & Paunonen, 2011), reliability indices are rarely included in implicit learning research (but see Siegelman & Frost, 2015). In Study 4 we have shown that the split-half reliability indices of ASRT learning scores vary from 0.02 to 0.84 depending on the type of the analysis (e.g. data grouping and filtering) and on type of learning that is being assessed (e.g. triplet-level learning or quad-level learning). In general, triplet learning scores are more reliable than quad learning scores (~0.6 vs. ~0.4), and the reliability of pattern learning is the worst – downright unreliable (~0.15). By using a stricter filter to eliminate the effects of pre-existing biases, individual variability got higher and reliability indices typically got lower, underscoring that without the strict filtering, the performance of different participants is more similar than their true learning abilities – differences are just being masked. Thus, there is a trade-off between validity and reliability, and this needs to be considered for analysis.

VI/2.2.2. Low individual variability

It has been assumed that implicit learning is robust and shows small inter- and intraindividual variability (A. S. Reber, 1993; A. S. Reber & Allen, 2000). Accordingly, individual differences in implicit cognition remain largely unexplored (A. S. Reber & Allen, 2000; but see Kaufman et al., 2010; and Kalra et al., 2019).

We contribute to this field of research by providing inter-individual variability indices for different learning scores in Study 4. We quantified the spread in data both in absolute (standard deviation) and in relative (coefficients of variation) terms; unfortunately, we are not aware of any standards by which we could tell whether the values that we found refute the hypothesis of „small individual variability". We would need similar descriptions of individual variability in other fields of cognitive psychology (e.g. explicit learning capacity) to draw any conclusions.

Another important message of Study 4 is that the variability of performance is sometimes shaped by factors we do not intend to measure (artifacts; e.g. pre-existing biases); in the ASRT such biases make the detectable differences smaller – an effect that has implications from the interpretation of result to the theories of implicit learning.

VI/2.2.3. Issues related to reaction-time based measures

It has been suggested that accuracy is a better measure than difference scores derived from reaction times (Urry et al., 2015, 2018). We did not directly test this hypothesis, but nevertheless contributed to this debate by showing that accuracy measures are less reliable than reaction time measures (Study 4, reliability indices). Also, generally speaking, we found reaction time based and error ratio based measures to point in the same direction in our studies, although sometimes complementing each other. For example in Study 3, by looking at reaction time based measures, we found that the implicit-implicit group showed reduced rewiring (compared to the other two groups). By looking at anticipatory errors, we found that this effect was due to the implicit-implicit group still expecting stimuli that were no longer probable, while the other two groups stopped expecting them shortly after the change in sequences. Thus, we were able to find a possible explanation for an effect that we also detected with reaction-time based measures.

In sum, we did not find unequivocal evidence that accuracy-based measures fare better than reaction-time based measures, and we even found that their reliability is substantially smaller. It is possible that accuracy-based measures in ASRT are actually prone to result in floor effects (contrary to the claims of Urry et al., 2015) because of the very high accuracy rates that are expected from participants in this task.

We also assessed the impact of pre-existing biases in serial reaction time measures (requiring serial motor responding). Such biases were acknowledged by Song et al. (2007a) in the context of ASRT, but, to our knowledge, their impact has not been systematically studied. In Study 4, we found that even the N-3rd stimulus has an effect on the reaction times measured on the Nth trial and that accounting for this (by using a filter that we named *quadfilter*) reverses the previously puzzling patterns observed for higher-order learning. Apart from affecting the magnitude of learning scores, higher individual variability could be detected by using this filter. This result indicates that the variability issue on this particular task could be attributed to people's susceptibility to pre-existing biases rather than to their similarity in their statistical learning capacity. We would like to underscore the need to address this question in every task that requires serial motor responding since the artifact that is introduced by these biases seems to be substantial.

VI/3. Strengths of the present Dissertation

One of the strengths of the present Dissertation (which is also one of its limitations) is that variations of the same task (i.e. the ASRT task) were used in all of the Studies. Although the results obtained could not be generalized to other implicit learning tasks, our work nevertheless expanded our knowledge about implicit statistical learning in visuomotor sequence learning tasks, and it could be a promising first step in the systematic exploration of implicit learning phenomena.

Although ASRT was used in all of the Studies presented in this Dissertation, we applied some modifications to the task to make it even more suitable for our purposes. For example, in Study 1 and Study 2 we introduced the so-called ASRT-RACE paradigm which was proven to be effective in separating perceptual and motor factors of implicit statistical learning. In Study 3 and Study 4 we did not make modifications to the task itself (as the Explicit variant of the ASRT has already been used by other authors, e.g. Song et al., 2007a), but we used atypical analysis methods to get the most out of the data. For example, in Study 3, we only analyzed random trials, so that any differences between groups performing the implicit vs. explicit variants of the task could not be attributed to explicit expectations about the upcoming stimuli. This made it possible to directly compare the implicit learning of all three groups (even those who performed the task explicitly). In Study 4 we suggested changes to the typical analysis methods of the ASRT task mainly concerning data filtering (data points that should be excluded from analysis) and data grouping (separating more than two groups of data based on subtler statistical properties). Our results confirmed that these modifications are worth the effort, and sometimes lead to qualitatively different results than the typical analyses.

In sum, we have highlighted many of the aspects of implicit statistical learning that need to be clarified if we aim to understand this process (or these processes) in detail. Although we did not (and could not) give definitive answers to all the emerging issues, we nevertheless provided some ideas on how to approach them. We have also shown that some practices that are currently widely applied need to be reconsidered due to methodological reasons; that the psychometric properties of the tasks need to be routinely assessed and reported; that pre-existing biases need to be considered when using reaction time based measures; and that small changes in the way of analysis sometimes lead to qualitatively different conclusions, emphasizing the need for prudent, methodologically sound analysis methods in order to get valid conclusions.

VI/4. Limitations and future directions

One of the strengths of the dissertation is also an important limitation: namely that only the ASRT task and its variants were used to examine implicit learning. In order to get to the ambitious goal presented earlier (i.e. to identify the possible components or submodules of implicit statistical learning), this kind of work that I presented in the current Dissertation needs to be done with different measures (different tasks) as well. For example, the complexity of the statistical structure embedded in the tasks could be varied in a systematic manner in the different types of tasks (e.g. the Weather Prediction task, Artificial Grammar Learning task, etc.) to see whether statistical complexity leads to similar phenomena in different paradigms and in different modalities.

As a second limitation, we have to emphasize that our results and conclusions are based on learning measures having relatively low reliability (as discussed in Study 4). In spite of having moderately large samples in our studies, we are aware of the need to treat our results with a reasonable amount of caution - only future studies will tell which results will stand the test of time.

Third, as part of a more general problem, we need to identify and understand the biases that shape our results with different methodologies – e.g. it is not enough if we admit that pre-existing biases exist in reaction time based studies, we should design experiments explicitly addressing these biases in order to get a better understanding. For example, we could ask participants to respond to a stream of statistically unstructured (random) stimuli and identify the reasons why some of the combinations are easy and others are hard. With such an understanding, we could design experiments that minimize such biases from the beginning (without the need to exclude a substantial amount of data from the analysis to deal with the problem after the fact).

VI/5. Conclusion

The research of implicit statistical learning (or implicit learning, in general) lead to very diverse – and sometimes contradictory – results. In order to understand the source of this variability, our duty is twofold: first, we need to improve the tasks that we use so that the results of measurements aid our theoretical understanding of implicit processes better; and second, based on our ever-expanding theoretical knowledge, we need to refine the tasks even more in order to narrow and specify their scope. Only this way could we get to the point where we know exactly what we intend to measure, and also have means to do it.

REFERENCES

Albouy, G., Sterpenich, V., Balteau, E., Vandewalle, G., Desseilles, M., Dang-Vu, T., … Maquet, P. (2008). Both the Hippocampus and Striatum Are Involved in Consolidation of Motor Sequence Memory. *Neuron*, *58*(2), 261–272. https://doi.org/10.1016/j.neuron.2008.02.008

Arciuli, J. (2017). The multi-component nature of statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711). https://doi.org/10.1098/rstb.2016.0058

Arciuli, J., & Conway, C. M. (2018). The Promise—And Challenge—Of Statistical Learning for Elucidating Atypical Language Development. *Current Directions in Psychological Science*, *27*(6), 492–500. https://doi.org/10.1177/0963721418779977

Barnes, K. A., Howard, J. H., Howard, D. V., Gilotty, L., Kenworthy, L., Gaillard, W. D., & Vaidya, C. J. (2008). Intact implicit learning of spatial context and temporal sequences in childhood autism spectrum disorder. *Neuropsychology*, *22*(5), 563.

Barnes, K. A., Howard, J. H., Howard, D. V., Kenealy, L., & Vaidya, C. J. (2010). Two forms of implicit learning in childhood ADHD. *Developmental Neuropsychology*, *35*(5), 494–505.

Bennett, I. J., Howard, J. H., & Howard, D. V. (2007). Age-Related Differences in Implicit Learning of Subtle Third-Order Sequential Structure. *The Journals of Gerontology: Series B*, *62*(2), P98–P103. https://doi.org/10.1093/geronb/62.2.P98

Bennett, I. J., Romano, J. C., Howard, J. H., & Howard, D. V. (2008). TWO FORMS OF IMPLICIT LEARNING IN YOUNG ADULT DYSLEXICS. *Annals of the New York Academy of Sciences*, *1145*, 184–198. https://doi.org/10.1196/annals.1416.006

Bergstrom, J. C. R., Howard, J. H., & Howard, D. V. (2012). Enhanced Implicit Sequence Learning in College-age Video Game Players and Musicians. *Applied Cognitive Psychology*, *26*(1), 91–96. https://doi.org/10.1002/acp.1800

Berry, D. C., & Broadbent, D. E. (1984). On the Relationship between Task Performance and Associated Verbalizable Knowledge. *The Quarterly Journal of*

Experimental Psychology Section A, *36*(2), 209–231. https://doi.org/10.1080/14640748408402156

Boyd, L. A., & Winstein, C. J. (2003). Impact of Explicit Information on Implicit Motor-Sequence Learning Following Middle Cerebral Artery Stroke. *Physical Therapy*, *83*(11), 976–989. https://doi.org/10.1093/ptj/83.11.976

Buchner, A., & Wippich, W. (2000). On the Reliability of Implicit and Explicit Memory Measures. *Cognitive Psychology*, *40*(3), 227–259. https://doi.org/10.1006/cogp.1999.0731

Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). The Pittsburgh Sleep Quality Index: A new instrument for psychiatric practice and research. *Psychiatry Research*, *28*(2), 193–213.

Christiansen, M. H. (2018). Implicit Statistical Learning: A Tale of Two Literatures. *Topics in Cognitive Science*. https://doi.org/10.1111/tops.12332

Chun, null. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences*, *4*(5), 170–178.

Cohen, A., Ivry, R., & Keele, S. (1990). Attention and Structure in Sequence Learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *16*(1), 17–30.

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*(1), 24–39. https://doi.org/10.1037/0278-7393.31.1.24

Conway, C. M., & Christiansen, M. H. (2009). Seeing and hearing in space and time: Effects of modality and presentation rate on implicit statistical learning. *European Journal of Cognitive Psychology*, *21*(4), 561–580. https://doi.org/10.1080/09541440802097951

Curran, T., & Keele, S. W. (1993). Attentional and nonattentional forms of sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(1), 189–202. https://doi.org/10.1037/0278-7393.19.1.189

Danner, D., Hagemann, D., & Funke, J. (2017). Measuring individual differences in implicit learning with artificial grammar learning tasks: Conceptual and methodological conundrums. *Zeitschrift Für Psychologie*, *225*(1), 5–19. https://doi.org/10.1027/2151-2604/a000280 de Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, *4*(4), 129–134. https://doi.org/10.1121/1.1613311

Debas, K., Carrier, J., Orban, P., Barakat, M., Lungu, O., Vandewalle, G., … Doyon, J. (2010). Brain plasticity related to the consolidation of motor sequence learning and motor adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(41), 17839–17844. https://doi.org/10.1073/pnas.1013176107

Dennis, N. A., & Cabeza, R. (2011). Age-related dedifferentiation of learning systems: An fMRI study of implicit and explicit learning. *Neurobiology of Aging*, *32*(12), 2318.e17-2318.e30. https://doi.org/10.1016/j.neurobiolaging.2010.04.004

Dennis, N. A., Howard, J. H., & Howard, D. V. (2006). Implicit sequence learning without motor sequencing in young and old adults. *Experimental Brain Research*, *175*(1), 153–164. https://doi.org/10.1007/s00221-006-0534-3

Deroost, N., Coomans, D., & Soetens, E. (2009). Perceptual load improves the expression but not learning of relevant sequence information. *Experimental Psychology*, *56*(2), 84–91.

Deroost, N., & Soetens, E. (2006). Perceptual or motor learning in SRT tasks with complex sequence structures. *Psychological Research*, *70*(2), 88–102. https://doi.org/10.1007/s00426-004-0196-3

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

Destrebecqz, Arnaud, Peigneux, P., Laureys, S., Degueldre, C., Del Fiore, G., Aerts, J., … Maquet, P. (2005). The neural correlates of implicit and explicit sequence learning: Interacting networks revealed by the process dissociation procedure. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *12*(5), 480–490. https://doi.org/10.1101/lm.95605

Dew, I. T. Z., & Cabeza, R. (2011). The porous boundaries between explicit and implicit memory: Behavioral and neural evidence. *Annals of the New York Academy of Sciences*, *1224*, 174–190. https://doi.org/10.1111/j.1749-6632.2010.05946.x

Dorfberger, S., Adi-Japha, E., & Karni, A. (2007). Reduced susceptibility to interference in the consolidation of motor memory before adolescence. *PloS One*, *2*(2), e240. https://doi.org/10.1371/journal.pone.0000240

Doyon, J. (2008). Motor sequence learning and movement disorders. *Current Opinion in Neurology*, *21*(4), 478–483. https://doi.org/10.1097/WCO.0b013e328304b6a3

Doyon, J., Bellec, P., Amsel, R., Penhune, V., Monchi, O., Carrier, J., … Benali, H. (2009). Contributions of the basal ganglia and functionally related brain structures to motor learning. *Behavioural Brain Research*, *199*(1), 61–75.

Doyon, J., Korman, M., Morin, A., Dostie, V., Hadj Tahar, A., Benali, H., … Carrier, J. (2009). Contribution of night and day sleep vs. Simple passage of time to the consolidation of motor sequence and visuomotor adaptation learning. *Experimental Brain Research*, *195*(1), 15–26. https://doi.org/10.1007/s00221-009-1748-y

Du, W., & Kelly, S. W. (2013). Implicit sequence learning in dyslexia: A within-sequence comparison of first- and higher-order information. *Annals of Dyslexia*, *63*(2), 154–170. https://doi.org/10.1007/s11881-012-0077-1

Du, Y., & Clark, J. E. (2017). New insights into statistical learning and chunk learning in implicit sequence acquisition. *Psychonomic Bulletin & Review*, *24*(4), 1225–1233. https://doi.org/10.3758/s13423-016-1193-4

E. Hick, W. (1952). On the rate of information gain. *Quarterly Journal of Experimental Psychology*, *4*, 11–26. https://doi.org/10.1080/17470215208416600

Edwards, J. R. (2001). Ten Difference Score Myths. *Organizational Research Methods*, *4*(3), 265–287. https://doi.org/10.1177/109442810143005

Emberson, L. L., Conway, C. M., & Christiansen, M. H. (2011). Timing is everything: Changes in presentation rate have opposite effects on auditory and visual implicit statistical learning. *Quarterly Journal of Experimental Psychology (2006)*, *64*(5), 1021–1040. https://doi.org/10.1080/17470218.2010.538972

Feeney, J. J., Howard, J. H., & Howard, D. V. (2002). Implicit learning of higher order sequences in middle age. *Psychology and Aging*, *17*(2), 351–355.

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality vs. modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences*, *19*(3), 117–125. https://doi.org/10.1016/j.tics.2014.12.010

Gao, J., Wong-Lin, K., Holmes, P., Simen, P., & Cohen, J. D. (2009). Sequential effects in two-choice reaction time tasks: Decomposition and synthesis of mechanisms.

Neural Computation, *21*(9), 2407–2436. https://doi.org/10.1162/neco.2009.09-08-866

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. https://doi.org/10.1037/0278-7393.33.1.34

Gerván, P., & Kovács, I. (2007). Sleep dependent learning in contour integration. *Journal of Vision*, *7*(9), 48–48. https://doi.org/10.1167/7.9.48

Ghilardi, M. F., Moisello, C., Silvestri, G., Ghez, C., & Krakauer, J. W. (2009). Learning of a Sequential Motor Skill Comprises Explicit and Implicit Components That Consolidate Differently. *Journal of Neurophysiology*, *101*(5), 2218–2229. https://doi.org/10.1152/jn.01138.2007

Goedert, K. M., & Willingham, D. B. (2002). Patterns of interference in sequence learning and prism adaptation inconsistent with the consolidation hypothesis. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *9*(5), 279–292. https://doi.org/10.1101/lm.50102

Goodwin, L. D., & Leech, N. L. (2006). Understanding Correlation: Factors That Affect the Size of r. *The Journal of Experimental Education*, *74*(3), 249–266. https://doi.org/10.3200/JEXE.74.3.249-266

Goschke, T. (1998). Implicit learning of perceptual and motor sequences: Evidence for independent learning systems. In *Handbook of implicit learning* (pp. 401–444). Thousand Oaks, CA, US: Sage Publications, Inc.

Goujon, A., Didierjean, A., & Poulet, S. (2014). The emergence of explicit knowledge from implicit learning. *Memory & Cognition*, *42*(2), 225–236. https://doi.org/10.3758/s13421-013-0355-0

Graf, P., & Schacter, D. L. (1985). Implicit and explicit memory for new associations in normal and amnesic subjects. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *11*(3), 501–518.

Grafton, S. T., Hazeltine, E., & Ivry, R. (1995). Functional mapping of sequence learning in normal humans. *Journal of Cognitive Neuroscience*, *7*(4), 497–510. https://doi.org/10.1162/jocn.1995.7.4.497

Grosjean, M., Rosenbaum, D. A., & Elsinger, C. (2001). Timing and reaction time. *Journal of Experimental Psychology. General*, *130*(2), 256–272.

Gucciardi, D. F., & Dimmock, J. A. (2008). Choking under pressure in sensorimotor skills: Conscious processing or depleted attentional resources? *Psychology of Sport and Exercise*, *9*(1), 45–59. https://doi.org/10.1016/j.psychsport.2006.10.007

Hallgató, E., Győri-Dani, D., Pekár, J., Janacsek, K., & Nemeth, D. (2013). The differential consolidation of perceptual and motor learning in skill acquisition. *Cortex*, *49*(4), 1073–1081. https://doi.org/10.1016/j.cortex.2012.01.002

Handa, A., Rhee, J., & Wright, D. L. (2016). The Structural Relationship Between Two Motor Sequences Practiced Close in Time Impacts Offline Facilitation. *Journal of Motor Behavior*, *48*(1), 47–56. https://doi.org/10.1080/00222895.2015.1035431

Hedenius, M., Persson, J., Tremblay, A., Adi-Japha, E., Veríssimo, J., Dye, C. D., … Ullman, M. T. (2011). Grammar predicts procedural learning and consolidation deficits in children with Specific Language Impairment. *Research in Developmental Disabilities*, *32*(6), 2362–2375. https://doi.org/10.1016/j.ridd.2011.07.026

Hedge, C., Powell, G., Bompas, A., Vivian-Griffiths, S., & Sumner, P. (2018). Low and variable correlation between reaction time costs and accuracy costs explained by accumulation models: Meta-analysis and simulations. *Psychological Bulletin*, *144*(11), 1200–1227. https://doi.org/10.1037/bul0000164

Heerey, E. A., & Velani, H. (2010). Implicit learning of social predictions. *Journal of Experimental Social Psychology*, *46*(3), 577–581. https://doi.org/10.1016/j.jesp.2010.01.003

Henke, K. (2010). A model for memory systems based on processing modes rather than consciousness. *Nature Reviews Neuroscience*, *11*(7), 523.

Hikosaka, O., Nakahara, H., Rand, M. K., Sakai, K., Lu, X., Nakamura, K., … Doya, K. (1999). Parallel neural networks for learning sequential procedures. *Trends in Neurosciences*, *22*(10), 464–471. https://doi.org/10.1016/S0166-2236(99)01439-3

Hikosaka, O., Nakamura, K., Sakai, K., & Nakahara, H. (2002). Central mechanisms of motor skill learning. *Current Opinion in Neurobiology*, *12*(2), 217–222. https://doi.org/10.1016/S0959-4388(02)00307-0

Horvath, K., Torok, C., Pesthy, O., Nemeth, D., & Janacsek, K. (2018). Explicit instruction differentially affects subcomponents of procedural learning and consolidation. *BioRxiv*, 433243.

Howard, D. V., & Howard, J. H. (1992). Adult age differences in the rate of learning serial patterns: Evidence from direct and indirect tests. *Psychology and Aging*, *7*(2), 232–241. https://doi.org/10.1037//0882-7974.7.2.232

Howard, D. V., & Howard, J. H. (2001). When it does hurt to try: Adult age differences in the effects of instructions on implicit pattern learning. *Psychonomic Bulletin & Review*, *8*(4), 798–805. https://doi.org/10.3758/BF03196220

Howard, D. V., Howard, J. H., Japikse, K., DiYanni, C., Thompson, A., & Somberg, R. (2004). Implicit sequence learning: Effects of level of structure, adult age, and extended practice. *Psychology and Aging*, *19*(1), 79–92. https://doi.org/10.1037/0882-7974.19.1.79

Howard, J. H., & Howard, D. V. (1997). Age differences in implicit learning of higher order dependencies in serial patterns. *Psychology and Aging*, *12*(4), 634–656.

Howard, J. H., Howard, D. V., Dennis, N. A., & Kelly, A. J. (2008). Implicit learning of predictive relationships in three-element visual sequences by young and old adults. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *34*(5), 1139–1157. https://doi.org/10.1037/a0012797

Howard, J. H., Howard, D. V., Dennis, N. A., & Yankovich, H. (2007). Event Timing and age deficits in higher-order sequence learning. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*, *14*(6), 647–668. https://doi.org/10.1080/13825580601186635

Howard, J. H., Howard, D. V., Japikse, K. C., & Eden, G. F. (2006). Dyslexics are impaired on implicit higher-order sequence learning, but not on implicit spatial context learning. *Neuropsychologia*, *44*(7), 1131–1144. https://doi.org/10.1016/j.neuropsychologia.2005.10.015

Janacsek, K., Ambrus, G. G., Paulus, W., Antal, A., & Nemeth, D. (2015). Right hemisphere advantage in statistical learning: Evidence from a probabilistic sequence learning task. *Brain Stimulation*, *8*(2), 277–282. https://doi.org/10.1016/j.brs.2014.11.008

Janacsek, K., Borbély-Ipkovich, E., Nemeth, D., & Gonda, X. (2018). How can the depressed mind extract and remember predictive relationships of the environment? Evidence from implicit probabilistic sequence learning. *Progress* *in Neuro-Psychopharmacology and Biological Psychiatry*, *81*, 17–24. https://doi.org/10.1016/j.pnpbp.2017.09.021

Janacsek, K., Fiser, J., & Nemeth, D. (2012). The best time to acquire new skills: Age-related differences in implicit sequence learning across the human lifespan. *Developmental Science*, *15*(4), 496–505. https://doi.org/10.1111/j.1467-7687.2012.01150.x

Japikse, K. C., Negash, S., Howard, J. H., & Howard, D. V. (2003). Intermanual transfer of procedural learning after extended practice of probabilistic sequences. *Experimental Brain Research*, *148*(1), 38–49. https://doi.org/10.1007/s00221-002-1264-9

Jiménez, L., Vaquero, J. M. M., & Lupiáñez, J. (2006). Qualitative differences between implicit and explicit sequence learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *32*(3), 475–490. https://doi.org/10.1037/0278-7393.32.3.475

Kalra, P. B., Gabrieli, J. D. E., & Finn, A. S. (2019). Evidence of stable individual differences in implicit learning. *Cognition*, *190*, 199–211. https://doi.org/10.1016/j.cognition.2019.05.007

Karni, A., Tanne, D., Rubenstein, B. S., Askenasy, J. J., & Sagi, D. (1994). Dependence on REM sleep of overnight improvement of a perceptual skill. *Science (New York, N.Y.)*, *265*(5172), 679–682.

Karzon, R. G. (1985). Discrimination of polysyllabic sequences by one- to four-month-old infants. *Journal of Experimental Child Psychology*, *39*(2), 326–342.

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. https://doi.org/10.1016/j.cognition.2010.05.011

Keele, S. W., Ivry, R., Mayr, U., Hazeltine, E., & Heuer, H. (2003). The cognitive and neural architecture of sequence representation. *Psychological Review*, *110*(2), 316–339.

Kim, R., Seitz, A., Feenstra, H., & Shams, L. (2009). Testing assumptions of statistical learning: Is it long-term and implicit? *Neuroscience Letters*, *461*(2), 145–149. https://doi.org/10.1016/j.neulet.2009.06.030

Kincses, T. Zs., Johansen-Berg, H., Tomassini, V., Bosnell, R., Matthews, P. M., & Beckmann, C. F. (2008). Model-free characterization of brain functional networks for motor sequence learning using fMRI. *NeuroImage*, *39*(4), 1950–1958. https://doi.org/10.1016/j.neuroimage.2007.09.070

Kirby, N. H. (1976). Sequential effects in two-choice reaction time: Automatic facilitation or subjective expectancy? *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 567.

Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *1*(2), 106–120.

Kóbor, A., Takács, Á., Kardos, Z., Janacsek, K., Horváth, K., Csépe, V., & Nemeth, D. (2018). ERPs differentiate the sensitivity to statistical probabilities and the learning of sequential structures during procedural learning. *Biological Psychology*, *135*, 180–193. https://doi.org/10.1016/j.biopsycho.2018.04.001

Kornblum, S. (1973). Sequential effects in choice reaction time: A tutorial review. *Attention and Performance IV*, 259–288.

Krakauer, J. W., & Shadmehr, R. (2006). Consolidation of motor memory. *Trends in Neurosciences*, *29*(1), 58–64. https://doi.org/10.1016/j.tins.2005.10.003

Lagarde, J., Li, L., Thon, B., Magill, R., & Erbani, E. (2002). Interactions between human explicit and implicit perceptual motor learning shown by kinematic variables. *Neuroscience Letters*, *327*(1), 66–70. https://doi.org/10.1016/s0304-3940(02)00380-4

Lebel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality & Social Psychology Bulletin*, *37*(4), 570–583. https://doi.org/10.1177/0146167211400619

Lee, J. C., Beesley, T., & Livesey, E. J. (2016). Sequential effects and sequence learning in a three-choice serial reaction time task. *Acta Psychologica*, *170*, 168–176. https://doi.org/10.1016/j.actpsy.2016.08.004

Li, X., Zhao, X., Shi, W., Lu, Y., & Conway, C. M. (2018). Lack of Cross-Modal Effects in Dual-Modality Implicit Statistical Learning. *Frontiers in Psychology*, *9*, 146. https://doi.org/10.3389/fpsyg.2018.00146

Marvel, C. L., Schwartz, B. L., Howard, D. V., & Howard, J. H. (2005). Implicit learning of non-spatial sequences in schizophrenia. *Journal of the International Neuropsychological Society*, *11*(6), 659–667. https://doi.org/10.1017/S1355617705050861

Masters, R. S. W. (1992). Knowledge, knerves and know-how: The role of explicit versus implicit knowledge in the breakdown of a complex motor skill under pressure. *British Journal of Psychology*, *83*(3), 343–358. https://doi.org/10.1111/j.2044-8295.1992.tb02446.x

Maxwell, J. P., Masters, R. S. W., & Eves, F. F. (2000). From novice to no know-how: A longitudinal study of implicit motor learning. *Journal of Sports Sciences*, *18*(2), 111–120. https://doi.org/10.1080/026404100365180

Mayr, U. (1996). Spatial attention and implicit sequence learning: Evidence for independent learning of spatial and nonspatial sequences. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *22*(2), 350–364.

Meier, B., & Cock, J. (2010). Are correlated streams of information necessary for implicit sequence learning? *Acta Psychologica*, *133*(1), 17–27. https://doi.org/10.1016/j.actpsy.2009.08.001

Negash, S., Boeve, B. F., Geda, Y. E., Smith, G. E., Knopman, D. S., Ivnik, R. J., … Petersen, R. C. (2007). Implicit Learning of Sequential Regularities and Spatial Contexts in Corticobasal Syndrome. *Neurocase*, *13*(3), 133–143. https://doi.org/10.1080/13554790701401852

Negash, S., Howard, D. V., Japikse, K. C., & Howard, J. H. (2003). Age-Related Differences in Implicit Learning of Non-Spatial Sequential Patterns. *Aging, Neuropsychology, and Cognition*, *10*(2), 108–121. https://doi.org/10.1076/anec.10.2.108.14462

Nemeth, D., Hallgató, E., Janacsek, K., Sándor, T., & Londe, Z. (2009). Perceptual and motor factors of implicit skill learning. *NeuroReport*, *20*(18), 1654. https://doi.org/10.1097/WNR.0b013e328333ba08

Nemeth, D., & Janacsek, K. (2011). The dynamics of implicit skill consolidation in young and elderly adults. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *66*(1), 15–22. https://doi.org/10.1093/geronb/gbq063

Nemeth, D., Janacsek, K., Balogh, V., Londe, Z., Mingesz, R., Fazekas, M., … Vetro, A. (2010). Learning in autism: Implicitly superb. *PloS One*, *5*(7), e11731. https://doi.org/10.1371/journal.pone.0011731

Nemeth, D., Janacsek, K., Csifcsak, G., Szvoboda, G., Howard, J. H., & Howard, D. V. (2011). Interference between Sentence Processing and Probabilistic Implicit

Sequence Learning. *PLOS ONE*, *6*(3), e17577. https://doi.org/10.1371/journal.pone.0017577

Nemeth, D., Janacsek, K., & Fiser, J. (2013). Age-dependent and coordinated shift in performance between implicit and explicit skill learning. *Frontiers in Computational Neuroscience*, *7*. https://doi.org/10.3389/fncom.2013.00147

Nemeth, D., Janacsek, K., Király, K., Londe, Z., Németh, K., Fazekas, K., … Csányi, A. (2013). Probabilistic sequence learning in mild cognitive impairment. *Frontiers in Human Neuroscience*, *7*, 318. https://doi.org/10.3389/fnhum.2013.00318

Nemeth, D., Janacsek, K., Londe, Z., Ullman, M. T., Howard, D. V., & Howard, J. H. (2010). Sleep has no critical role in implicit motor sequence learning in young and old adults. *Experimental Brain Research*, *201*(2), 351–358. https://doi.org/10.1007/s00221-009-2024-x

Nemeth, D., Janacsek, K., Polner, B., & Kovacs, Z. A. (2013). Boosting Human Learning by Hypnosis. *Cerebral Cortex*, *23*(4), 801–805. https://doi.org/10.1093/cercor/bhs068

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, *19*(1), 1–32. https://doi.org/10.1016/0010-0285(87)90002-8

Norman, E., & Price, M. C. (2012). Social intuition as a form of implicit learning: Sequences of body movements are learned less explicitly than letter sequences. *Advances in Cognitive Psychology*, *8*(2), 121–131. https://doi.org/10.2478/v10053-008-0109-x

Parshina, O., Obeid, R., Che, E. S., Ricker, T. J., & Brooks, P. J. (2018). SRT and ASRT: Similar Tasks Tapping Distinct Learning Mechanisms? *The 40th Annual Meeting of the Cognitive Science Society (CogSci 2018)*, Madison, Wisconsin (USA), July 25-28. 2018.

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Perruchet, P., Bigand, E., & Benoit-Gonin, F. (1997). The emergence of explicit knowledge during the early phase of learning in sequential reaction time tasks. *Psychological Research*, *60*(1), 4–13. https://doi.org/10.1007/BF00419676

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*(5), 233–238. https://doi.org/10.1016/j.tics.2006.03.006

Press, D. Z., Casement, M. D., Pascual-Leone, A., & Robertson, E. M. (2005). The time course of off-line motor sequence learning. *Brain Research. Cognitive Brain Research*, *25*(1), 375–378. https://doi.org/10.1016/j.cogbrainres.2005.05.010

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*(6), 855–863. https://doi.org/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., & Allen, R. (2000). Individual Differences in Implicit Learning Implications for the Evolution of Consciousness. In R. G. Kunzendorf & B. Wallace, *Individual differences in subjective experience: First-person constraints on theories of consciousness, subconsciousness, and self-consciousness*. Amsterdam: John Benjamins.

Reber, P. J. (2013). The neural basis of implicit learning and memory: A review of neuropsychological and neuroimaging research. *Neuropsychologia*, *51*(10), 2026–2042. https://doi.org/10.1016/j.neuropsychologia.2013.06.019

Remillard, G. (2003). Pure perceptual-based sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 581–597.

Remillard, G. (2008). Implicit learning of second-, third-, and fourth-order adjacent and nonadjacent sequential dependencies. *The Quarterly Journal of Experimental Psychology*, *61*(3), 400–424. https://doi.org/10.1080/17470210701210999

Remillard, G. (2009). Pure perceptual-based sequence learning: A role for visuospatial attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 528–541. https://doi.org/10.1037/a0014646

Remillard, G. (2010). Implicit learning of fifth- and sixth-order sequential probabilities. *Memory & Cognition*, *38*(7), 905–915. https://doi.org/10.3758/MC.38.7.905

Remillard, G. (2011). Pure perceptual-based learning of second-, third-, and fourth-order sequential probabilities. *Psychological Research*, *75*(4), 307–323. https://doi.org/10.1007/s00426-010-0309-0

Remington, R. J. (1969). Analysis of sequential effects on choice reaction times. *Journal of Experimental Psychology*, *82*(2), 250–257. https://doi.org/10.1037/h0028122

Robertson, E. M., Press, D. Z., & Pascual-Leone, A. (2005). Off-line learning and the primary motor cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *25*(27), 6372–6378. https://doi.org/10.1523/JNEUROSCI.1851-05.2005

Robertson, E. M., Tormos, J. M., Maeda, F., & Pascual-Leone, A. (2001). The role of the dorsolateral prefrontal cortex during sequence learning is specific for spatial information. *Cerebral Cortex*, *11*(7), 628–635.

Robertson, Edwin M. (2009). From Creation to Consolidation: A Novel Framework for Memory Processing. *PLOS Biology*, *7*(1), e1000019. https://doi.org/10.1371/journal.pbio.1000019

Robertson, Edwin M., Pascual-Leone, A., & Press, D. Z. (2004). Awareness modifies the skill-learning benefits of sleep. *Current Biology: CB*, *14*(3), 208–212. https://doi.org/10.1016/j.cub.2004.01.027

Rünger, D., & Frensch, P. A. (2008). How incidental sequence learning creates reportable knowledge: The role of unexpected events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1011–1026. https://doi.org/10.1037/a0012942

Rüsseler, J., & Rösler, F. (2000). Implicit and explicit learning of event sequences: Evidence for distinct coding of perceptual and motor representations. *Acta Psychologica*, *104*(1), 45–67.

Sævland, W., & Norman, E. (2016). Studying Different Tasks of Implicit Learning across Multiple Test Sessions Conducted on the Web. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00808

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Sanchez, D. J., & Reber, P. J. (2013). Explicit pre-training instruction does not improve implicit perceptual-motor sequence learning. *Cognition*, *126*(3), 341–351. https://doi.org/10.1016/j.cognition.2012.11.006

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(3), 501–518. https://doi.org/10.1037/0278-7393.13.3.501

Schendan, H. E., Searl, M. M., Melrose, R. J., & Stern, C. E. (2003). An FMRI study of the role of the medial temporal lobe in implicit and explicit sequence learning. *Neuron*, *37*(6), 1013–1025.

Schneider, W., Eschman, A., & Zuccolotto, A. (2012). E-Prime 2.0 reference guide manual. *Pittsburgh, PA: Psychology Software Tools*.

Schwartz, B. L., Howard, D. V., Howard, J. H., Hovaguimian, A., & Deutsch, S. I. (2003). Implicit learning of visuospatial sequences in schizophrenia. *Neuropsychology*, *17*(3), 517–533.

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, *81*, 105–120. https://doi.org/10.1016/j.jml.2015.02.001

Siengsukon, C. F., & Boyd, L. A. (2008). Sleep enhances implicit motor skill learning in individuals poststroke. *Topics in Stroke Rehabilitation*, *15*(1), 1–12. https://doi.org/10.1310/tsr1501-1

Simor, P., Zavecz, Z., Horváth, K., Éltető, N., Török, C., Pesthy, O., … Nemeth, D. (2019). Deconstructing Procedural Memory: Different Learning Trajectories and Consolidation of Sequence and Statistical Learning. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02708

Smith, J. G., & McDowall, J. (2004). Impaired higher order implicit sequence learning on the verbal version of the serial reaction time task in patients with Parkinson's disease. *Neuropsychology*, *18*(4), 679–691. https://doi.org/10.1037/0894-4105.18.4.679

Soetens, E., Boer, L. C., & Hueting, J. E. (1985). Expectancy or automatic facilitation? Separating sequential effects in two-choice reaction time. *Journal of Experimental Psychology: Human Perception and Performance*, *11*(5), 598.

Song, S. (2009). Consciousness and the consolidation of motor learning. *Behavioural Brain Research*, *196*(2), 180–186. https://doi.org/10.1016/j.bbr.2008.09.034

Song, S., Howard, J. H., & Howard, D. V. (2007a). Implicit probabilistic sequence learning is independent of explicit awareness. *Learning & Memory*, *14*(3), 167–176. https://doi.org/10.1101/lm.437407

Song, S., Howard, J. H., & Howard, D. V. (2007b). Sleep does not benefit probabilistic motor sequence learning. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *27*(46), 12475–12483. https://doi.org/10.1523/JNEUROSCI.2062-07.2007

Song, S., Howard, J. H., & Howard, D. V. (2008). Perceptual sequence learning in a serial reaction time task. *Experimental Brain Research. Experimentelle Hirnforschung. Experimentation Cerebrale*, *189*(2), 145–158. https://doi.org/10.1007/s00221-008-1411-z

Song, S., Marks, B., Howard, J. H., & Howard, D. V. (2009). Evidence for parallel explicit and implicit sequence learning systems in older adults. *Behavioural Brain Research*, *196*(2), 328–332. https://doi.org/10.1016/j.bbr.2008.09.022

Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences*, *93*(24), 13515–13522.

Stark-Inbar, A., Raza, M., Taylor, J. A., & Ivry, R. B. (2016). Individual differences in implicit motor learning: Task specificity in sensorimotor adaptation and sequence learning. *Journal of Neurophysiology*, *117*(1), 412–428.

Stickgold, R., & Walker, M. P. (2005). Sleep and memory: The ongoing debate. *Sleep*, *28*(10), 1225–1227.

Sun, R., Zhang, X., Slusarz, P., & Mathews, R. (2007). The interaction of implicit learning, explicit hypothesis testing learning and implicit-to-explicit knowledge extraction. *Neural Networks: The Official Journal of the International Neural Network Society*, *20*(1), 34–47. https://doi.org/10.1016/j.neunet.2006.07.002

Takács, Á., Kóbor, A., Chezan, J., Éltető, N., Tárnok, Z., Nemeth, D., … Janacsek, K. (2018). Is procedural memory enhanced in Tourette syndrome? Evidence from a sequence learning task. *Cortex*, *100*, 84–94. https://doi.org/10.1016/j.cortex.2017.08.037

Taylor, J. A., Krakauer, J. W., & Ivry, R. B. (2014). Explicit and Implicit Contributions to Learning in a Sensorimotor Adaptation Task. *The Journal of Neuroscience*, *34*(8), 3023–3032. https://doi.org/10.1523/JNEUROSCI.3619-13.2014

Thiessen, E. D. (2017). What’s statistical about learning? Insights from modelling statistical learning as a set of memory processes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *372*(1711). https://doi.org/10.1098/rstb.2016.0056

Thiessen, E. D., Kronstein, A. T., & Hufnagle, D. G. (2013). The extraction and integration framework: A two-process account of statistical learning. *Psychological Bulletin*, *139*(4), 792–814. https://doi.org/10.1037/a0030801

Török, B., Janacsek, K., Nagy, D. G., Orbán, G., & Nemeth, D. (2017). Measuring and filtering reactive inhibition is essential for assessing serial decision making and learning. *Journal of Experimental Psychology. General*, *146*(4), 529–542. https://doi.org/10.1037/xge0000288

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2008). Neural Evidence of Statistical Learning: Efficient Detection of Visual Regularities Without Awareness. *Journal of Cognitive Neuroscience*, *21*(10), 1934–1945. https://doi.org/10.1162/jocn.2009.21131

Urry, K., Burns, N. R., & Baetu, I. (2015). Accuracy-based measures provide a better measure of sequence learning than reaction time-based measures. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01158

Urry, K., Burns, N. R., & Baetu, I. (2018). Age-related differences in sequence learning: Findings from two visuo-motor sequence learning tasks. *British Journal of Psychology*, *109*(4), 830–849. https://doi.org/10.1111/bjop.12299

Verneau, M., van der Kamp, J., Savelsbergh, G. J. P., & de Looze, M. P. (2015). Proactive and retroactive transfer of middle age adults in a sequential motor learning task. *Acta Psychologica*, *156*, 57–63. https://doi.org/10.1016/j.actpsy.2015.01.009

Vervaeck, K. R., & Boer, L. C. (1980). Sequential effects in two-choice reaction time: Subjective expectancy and automatic after-effect at short response-stimulus intervals. *Acta Psychologica*, *44*(2), 175–190.

Walk, A. M., & Conway, C. M. (2016). Cross-Domain Statistical–Sequential Dependencies Are Difficult to Learn. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00250

Walker, C. M., Walker, M. F., & Sunderland, A. (2003). Dressing after a Stroke: A Survey of Current Occupational Therapy Practice. *British Journal of Occupational Therapy*, *66*(6), 263–268. https://doi.org/10.1177/030802260306600605

Walker, M. P., Brakefield, T., Hobson, J. A., & Stickgold, R. (2003). Dissociable stages of human memory consolidation and reconsolidation. *Nature*, *425*(6958), 616–620. https://doi.org/10.1038/nature01930

Walker, M. P., Brakefield, T., Morgan, A., Hobson, J. A., & Stickgold, R. (2002). Practice with sleep makes perfect: Sleep-dependent motor skill learning. *Neuron*, *35*(1), 205–211.

Weiermann, B., Cock, J., & Meier, B. (2010). What matters in implicit task sequence learning: Perceptual stimulus features, task sets, or correlated streams of information? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *36*(6), 1492–1509. https://doi.org/10.1037/a0021038

Willingham, D. B. (1999). Implicit motor sequence learning is not purely perceptual. *Memory & Cognition*, *27*(3), 561–572.

Willingham, D. B., & Koroshetz, W. J. (1993). Evidence for dissociable motor skills in Huntington's disease patients. *Psychobiology*, *21*(3), 173–182. https://doi.org/10.3758/BF03327132

Willingham, D. B., Nissen, M. J., & Bullemer, P. (1989). On the development of procedural knowledge. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *15*(6), 1047–1060.

Willingham, D. B., Wells, L. A., Farrell, J. M., & Stemwedel, M. E. (2000). Implicit motor sequence learning is represented in response locations. *Memory & Cognition*, *28*(3), 366–375. https://doi.org/10.3758/BF03198552

Woods, D. L., Wyma, J. M., Yund, E. W., Herron, T. J., & Reed, B. (2015). Factors influencing the latency of simple reaction time. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00131

Yotsumoto, Y., Watanabe, T., Chang, L.-H., & Sasaki, Y. (2013). Consolidated learning can be susceptible to gradually-developing interference in prolonged motor learning. *Frontiers in Computational Neuroscience*, *7*, 69. https://doi.org/10.3389/fncom.2013.00069

Zang, X., Zinchenko, A., Jia, L., Assumpção, L., & Li, H. (2018). Global Repetition Influences Contextual Cueing. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00402

Ziessler, M., & Nattkemper, D. (2001). Learning of event sequences is based on response-effect learning: Further evidence from a serial reaction task. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *27*(3), 595–613.

SUPPLEMETARY MATERIALS

Supplementary Materials for Study 2

ST-III/1.

Means and SDs for SLEs at the end of the Learning Phase, at the beginning of the Transfer Phase. SLE-change indicates the difference in SLEs between the two sessions.

Condition	Delay	Daytime	N	SLE (Learning Phase)		SLE (Transfer Phase)		SLE (Transfer – Learning)	
				Mean	SD	Mean	SD	Mean	SD
Perceptual	12-h	Morning-first	11	8.59	12.57	6.09	12.06	-2.50	17.45
		Evening-first	11	9.68	13.27	9.14	15.68	-.55	18.47
	24-h	Morning-first	14	14.82	19.39	12.11	8.70	-2.71	16.75
		Evening-first	14	18.64	25.42	8.86	13.64	-9.79	28.29
	Total		50	13.39	18.87	9.22	12.41	-4.17	20.78
Motor	12-h	Morning-first	12	8.13	22.51	16.29	15.20	8.17	19.14
		Evening-first	11	10.73	27.96	18.36	9.03	7.64	27.58
	24-h	Morning-first	12	11.63	18.27	16.67	12.18	5.04	24.64
		Evening-first	17	6.68	11.63	7.56	6.79	.88	12.63
	Total		52	9.01	19.53	13.96	11.54	4.95	20.46
Total			102	11.16	19.24	11.41	14.45	.48	21.02

ST-III/2.

Means and SDs for the first two sequence blocks of the Learning and Transfer Phase for perceptual and motor condition.

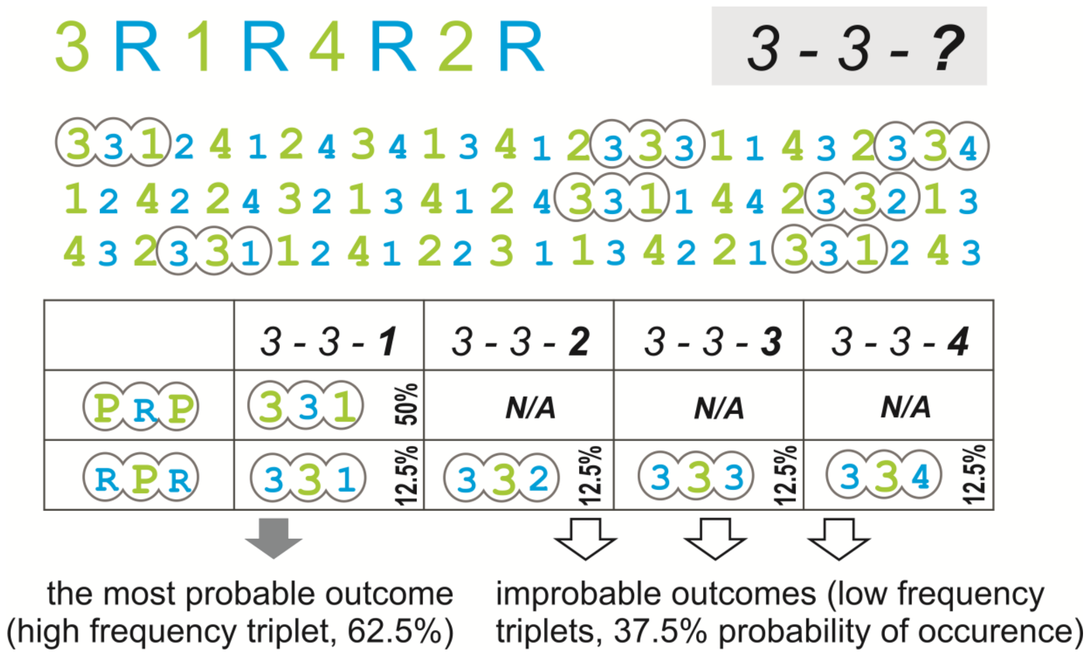
Condition	Phase	Mean	SD
Perceptual (N = 50)	Learning	-9.27	61.34
	Transfer	8.33	16.72
Motor (N = 52)	Learning	3.89	93.46
	Transfer	14.77	20.94

Supplementary Materials for Study 3

Supplementary Methods IV/1.

SM- IV/1.1. The structure of the ASRT sequences

In the ASRT task (Howard & Howard, 1997) the probability of each stimulus location out of the four possible ones (0^{th} order probability) is equal (25%). For any stimulus *n*, the previous $n - 1$ trial (1^{st} order transitional probability) has no predictive value either (all pairs of stimuli are equally probable). The ASRT sequence is a 2^{nd} order probabilistic sequence because for any trial *n* there is a very probable and three less probable continuations of the sequence based on the $n - 2^{th}$ trial. The probabilities add up the following way: if a pattern trial comes up, the identity of this trial can be inferred with 100% certainty based on the previous pattern trial which occurred two trials before (thus 50% of all trials are predetermined as 50% of all trials are pattern trials). For example, in the case of a sequence such as 3-R-1-R-4-R-2-R, if the previous pattern trial was on the 3^{rd} location, the next pattern trial is going to be on the 1^{st}. If a random trial comes up, on the other hand, it could be any of the four possible stimuli with 25% probability (irrespectively of the stimulus that occurred two trials before) – as this 25% refers to random trials only, which makes up 50% of all trials, a particular outcome has an overall probability of 12.5% in this case. Taken together, there is always a probable outcome regarding the upcoming stimulus (50% + 12.5% = 62.5%) and three less probable outcomes (12.5% each) based on trial *n–2* (Supplementary Figure **SF-IV/1**). In the conventional (implicit) ASRT task individuals have no clue about the alternating nature of the random and pattern trials, they nevertheless learn that it is highly probable that they are going to encounter a stimulus on the 1st location if they encountered a stimulus on the 3rd location two trials before. Participants use this statistical knowledge on random and pattern trials alike.

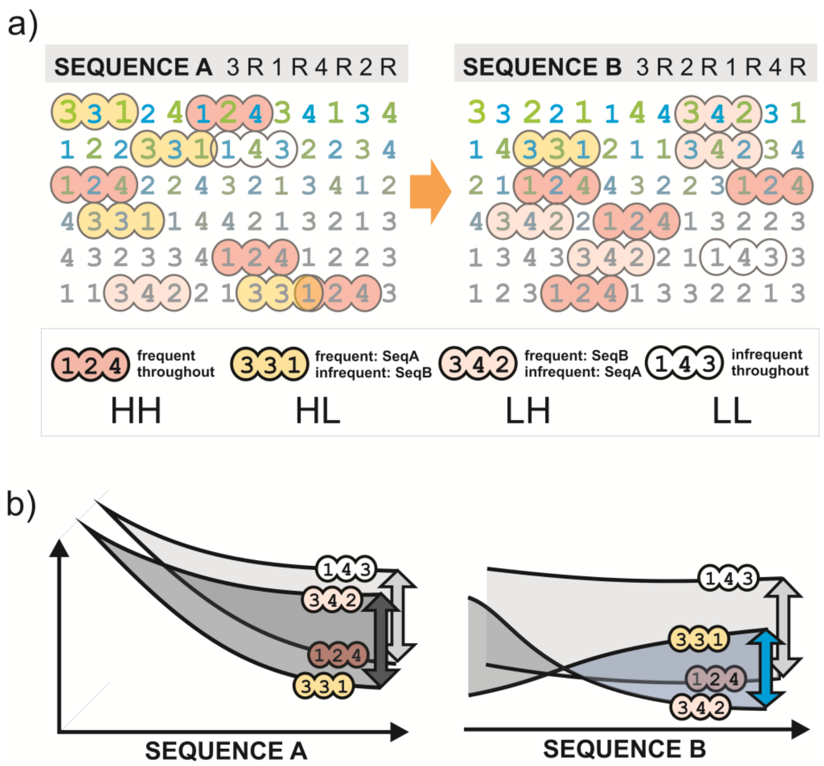


SF-IV/1. The statistical structure of the ASRT sequence. As a result of the alternation of pattern (green) and random („R", blue) trials, there are frequent (more probable) and infrequent (less probable) combinations of three consecutive stimuli. Whatever the first two elements are of such a combination (a so called triplet), there is always a probable continuation which occurs 62.5% of the time, and three less probable continuations with a probability of 12.5% each. Frequent combinations are called *high frequency triplets*, while the infrequent combinations are called *low frequency triplets*. In the above example after encountering two consecutive stimuli on the 3rd location, the most probable upcoming stimulus is one on the 1st location: if the upcoming stimulus is a pattern element (green), then it will be on the 1st location because of the embedded pattern. As only 50% of trials are pattern elements, there is 50% chance that the following stimulus will appear on the 1st location. However, there is another 50% chance that the upcoming stimulus is a random element (green). In this case each outcome is equally likely (12.5% each). Taken together, there is 50+12.5 = 62.5% chance that the upcoming stimulus will be on the 1st location – making this the most probable outcome. The comparison of high vs. low frequency triplets captures the 2nd order transitional probabilities embedded in the ASRT sequence. In the original (implicit) version of the ASRT pattern and random elements are shown in the same color, and participants are not told about the embedded regularity.

If the predicted stimulus comes up, we can categorize that stimulus as the final stimulus of a so called high frequency triplet, while the less probable continuations are categorized as the final stimulus of a low frequency triplet. Thus, when we use the terms *high frequency triplet* and *low frequency triplet*, we refer to the predictability of the final element of that triplet, and we quantify participants' suppositions as reaction times (RTs) to these final elements of triplets as a function of their probability (each element is categorized this way; the 3rd element of a triplet is also a second element of the following triplet, and so on). The typical result is shorter RTs to the last elements of high frequency triplets than to the last elements of low frequency triplets.

SM-IV/1.2. Unchanged and changed transitions across the Learning and Rewiring Phase

As each participant encountered two different sequences on the subsequent days of the experiment, the same stimuli could be a probable continuation of the same contexts during both sequences, or they could change their probability along with the change in sequence structure. For example, if the sequences were 3-R-1-R-4-R-2-R and 3-R-2-R-1-R-4-R, respectively, stimulus on the 4th location could be anticipated whenever a stimulus on the 1st location was encountered two trials earlier (1-1-**4**, 1-2-**4**, 1-3-**4** and 1-4-**4** are all high frequency triplets during both phases of the study). We can refer to such stimuli as „high-high" (HH), indicating their probability in the two subsequent phases of the study (Supplementary Figure **SF-IV/2/a**). There are also stimuli that are less probable during both phases, for example encountering a stimulus on the 3rd location after encountering a stimulus on the 1st location two trials before (i.e. 1-1-**3**, 1-2-**3**, 1-3-**3** or 1-4-**3**). This can only happen on random trials, with an overall probability of 12.5%. These trials can be categorized as „low-low" (LL). Finally, there are cases when the probability of a stimulus changes with the change in the sequence structure. For example, encountering a stimulus on the 1st location is highly probable during the first sequence when a participant encountered a stimulus on the 3rd location two trials before (i.e. 3-1-**1**, 3-2-**1**, 3-3-**1** or 3-4-**1**); however, the same stimulus is less probable during the second sequence (encountering a stimulus on the 2nd location would be probable in this case: 3-1-**2**, 3-2-**2**, 3-3-**2** or 3-4-**2**). If a stimulus is the probable continuation of its context during the first sequence, but less probable during the second, we can categorize it as „high-low" (HL); if it is less probable during the first sequence and probable during the second, we can categorize it as „low-high" (LH). Thus, the changes introduced between the Learning and the Rewiring Phase was based on 2nd order dependencies in the ASRT sequence that was captured by these four triplet types. Each pair of sequences had the same amount of shared transitions (i.e. the proportion of HH, HL, LH and LL triplets was constant across participants).



SF-IV/2. Stimulus types as a function of shared vs. not shared transitional probabilities in Sequence A and Sequence B. **(a)** Some of the triplets are frequent in both sequences – as they are „high frequency" during both Phases, we called them HH („high-high") triplets (e.g. 1-2-4). Other transitions are frequent only in one of the sequences; the ones that are of high frequency in Sequence A but of low frequency in Sequence B (e.g. 3-3-1) are called HL triplets („high-low"); the ones with the opposite pattern (e.g. 3-4-2) are called LH triplets („low-high"). Finally, some of the triplets are of low probability during both sequences (e.g. 1-4-3) – these are called LL („low-low") triplets. **(b)** Reaction times (RTs) to the final elements of the high frequency triplets are expected to be faster than RTs to the final elements of the low frequency triplets. Statistical Learning Effect (SLE) is the difference between RTs given to the two types of events (more probable vs. less probable) in both sequences. As particular transitions changed their frequency of occurrence when moving to Sequence B from Sequence A, different SLE-s were calculated. The SLE for the never changed transitional probabilities (HH vs. LL, e.g. 1-2-4 vs. 1-4-3) is shown by light grey arrows. SLE for the changed sequence parts is shown by two different colors; the dark grey arrow represents SLE before the swap in frequencies took place (3-3-1 vs. 3-4-2), while the blue arrow represents SLE after the change occurred (3-4-2 vs. 3-3-1). For the sake of clarity, we depicted these arrows at the end of learning, although average differences during the whole learning process were calculated.

SM-IV/1.3. Sequence combinations used in the current study

Different combinations of sequences were used in the experiment. Only 24 of the 30 possible combinations result in all four types of triplets (HH, LL, HL and LH) – or 12 out of 15 if we do not take order into account. The remaining sequence combinations would result in only LL, HL and LH triplets, no HH triplets (for example the sequences 1-R-2-R-3-R-4-R and 1-R-4-R-3-R-2-R have no high frequency triplets in common). All adequate combinations were used in the experiment (see Table S1) in a counterbalanced order. Importantly, we treated sequences 1-R-4-R-2-R-3-R-, 4-R-2-R-3-R-1-R-, 3-R-1-R-4-R-2-R- and 2-R-3-R-1-R-4-R- as being identical since they consist of the same triplets (just beginning at a different point of the sequence).

ST-IV/1. Sequence combinations used in the experiment.

Sequence combinations	HH	LH or HL (depending on the order)	LL
1-R-2-R-3-R-4-R and 1-R-2-R-4-R-3-R	1-X-2 - - -	- 2-X-3 / 2-X-4 3-X-4 / 3-X-1 4-X-1 / 4-X-3	1-X-1, 1-X-3, 1-X-4, 2-X-1, 2-X-2, 3-X-2, 3-X-3, 4-X-2, 4-X-4
1-R-3-R-2-R-4-R and 1-R-3-R-4-R-2-R	1-X-3 - - -	- 2-X-4 / 2-X-1 3-X-2 / 3-X-4 4-X-1 / 4-X-2	1-X-1, 1-X-2, 1-X-4, 2-X-2, 2-X-3, 3-X-1, 3-X-3, 4-X-3, 4-X-4
1-R-4-R-2-R-3-R and 1-R-4-R-3-R-2-R	1-X-4 - - -	- 2-X-3 / 2-X-1 3-X-1 / 3-X-2 4-X-2 / 4-X-3	1-X-1, 1-X-2, 1-X-3, 2-X-2, 2-X-4, 3-X-3, 3-X-4, 4-X-1, 4-X-4
1-R-3-R-4-R-2-R and 1-R-4-R-3-R-2-R	- 2-X-1 - -	1-X-3 / 1-X-4 - 3-X-4 / 3-X-2 4-X-2 / 4-X-3	1-X-1, 1-X-2, 2-X-2, 2-X-3, 2-X-4, 3-X-1, 3-X-3, 4-X-1, 4-X-4
1-R-2-R-3-R-4-R and 1-R-4-R-2-R-3-R	- 2-X-3 - -	1-X-2 / 1-X-4 - 3-X-4 / 3-X-1 4-X-1 / 4-X-2	1-X-1, 1-X-3, 2-X-1, 2-X-2, 2-X-4, 3-X-2, 3-X-3, 4-X-3, 4-X-4
1-R-2-R-4-R-3-R and 1-R-3-R-2-R-4-R	- 2-X-4 - -	1-X-2 / 1-X-3 - 3-X-1 / 3-X-2 4-X-3 / 4-X-1	1-X-1, 1-X-4, 2-X-1, 2-X-2, 2-X-3, 3-X-3, 3-X-4, 4-X-2, 4-X-4

1-R-2-R-4-R-3-R and 1-R-4-R-2-R-3-R	- - 3-X-1 -	1-X-2 / 1-X-4 2-X-4 / 2-X-3 - 4-X-3 / 4-X-2	1-X-1, 1-X-3, 2-X-1, 2-X-2, 3-X-2, 3-X-3, 3-X-4, 4-X-1, 4-X-4
1-R-3-R-2-R-4-R and 1-R-4-R-3-R-2-R	- - 3-X-2 -	1-X-3 / 1-X-4 2-X-4 / 2-X-1 - 4-X-1 / 4-X-3	1-X-1, 1-X-2, 2-X-2, 2-X-3, 3-X-1, 3-X-3, 3-X-4, 4-X-2, 4-X-4
1-R-2-R-3-R-4-R and 1-R-3-R-4-R-2-R	- - 3-X-4 -	1-X-2 / 1-X-3 2-X-3 / 2-X-1 - 4-X-1 / 4-X-2	1-X-1, 1-X-4, 2-X-2, 2-X-4, 3-X-1, 3-X-3, 3-X-3, 4-X-3, 4-X-4
1-R-2-R-3-R-4-R and 1-R-3-R-2-R-4-R	- - - 4-X-1	1-X-2 / 1-X-3 2-X-3 / 2-X-4 3-X-4 / 3-X-2 -	1-X-1, 1-X-4, 2-X-1, 2-X-2, 3-X-1, 3-X-3, 4-X-2, 4-X-3, 4-X-4
1-R-3-R-4-R-2-R and 1-R-4-R-2-R-3-R	- - - 4-X-2	1-X-3 / 1-X-4 2-X-1 / 2-X-3 3-X-4 / 3-X-1 -	1-X-1, 1-X-2, 2-X-2, 2-X-4, 3-X-2, 3-X-3, 4-X-1, 4-X-3, 4-X-4
1-R-2-R-4-R-3-R and 1-R-4-R-3-R-2-R	- - - 4-X-3	1-X-2 / 1-X-4 2-X-4 / 2-X-1 3-X-1 / 3-X-2 -	1-X-1, 1-X-3, 2-X-2, 2-X-3, 3-X-3, 3-X-4, 4-X-1, 4-X-2, 4-X-4

The first pattern trial in each block was chosen randomly, so for example if the ASRT sequence was 3-R-1-R-4-R-2, then some of the blocks started as 3-R-1-R-4-R-2-R-3-R-1-R-4-R-2-R- while others started as 1-R-4-R-2-R-3-R-1-R-4-2 or 4-R-3-R-1-R-4-R-2-R-3-R- or 2-R-3-R-1-R-4-R-2-R-3. Note that the pattern is repeated in the block (10 times in each), thus changing the starting point does not lead to a different sequence (just as 123412341234 is the same as 234123412341).

In each combination, there was a partial overlap between Sequence A and Sequence B. Twenty-five percent of the originally high frequency triplets in Sequence A remained high frequency in Sequence B as well, while the remaining 75% became low frequency triplets. For example, Sequence A was 1–R–3–R–4–R–2 and Sequence B was 1–R–4–R–3–R–2 (see Line 4 in the table above). Following this example, out of the 16 originally high frequency triplets (i.e., 2–X–1, 1–X–3, 3–X–4, 4–X–2; X indicates the middle element of the triplet; i.e., 2–1–1, 2–2–1, 2–3–1, 2–4–1), four remained unchanged (triplets 2–X–1; HH triplets) and 12 high frequency triplets became low frequency ones (HL triplets).

Beyond the 16 high frequency triplets, there were 48 low frequency triplets for a given sequence. Out of these 48 low frequency triplets, 12 became high frequency (in the above example: 1–X–4, 4–X–3, 3–X–2; LH triplets), and 36 remained low frequency (e.g., 2–X–3, 2–X–4, 1–X–2, 4–X–1; LL triplets). In Table S1 we included which triplets corresponded to the categories of HH, HL, LH, and LL for each sequence pair combinations.

SM-IV/1.4. Calculation of Statistical Learning Effect (SLE)

In accordance with the original way of analysis (Howard & Howard, 1997), RTs given to repetitions (e.g. 1-1-1) or trills (e.g. 1-3-1) were excluded, along with RTs of inaccurate responses and preparatory trials. Unlike the conventional analysis, we also excluded RTs given to pattern trials (except for the probe blocks, see **Figure IV/1** in the main text). If pattern trials were included, average RTs would have been lower for individuals performing the Explicit variant of the task, as they could explicitly anticipate stimuli on pattern trials (50% of all trials). Our aim was not to compare RTs when individuals knew in advance what the next stimulus was going to be versus when they had no explicit knowledge about this; we wanted to measure knowledge about the statistical structure which accompanies the alternating nature of random and pattern trials. Thus, by excluding pattern trials from analysis, we aimed to compare participants' statistical knowledge under similar conditions (when they could not explicitly anticipate the stimuli, irrespective of the type of ASRT) (see "pure statistical learning" in Nemeth, Janacsek, & Fiser, 2013).

We calculated median RTs given to random trials for each participant in each epoch for the four possible triplet types: HH, LL, HL and LH. Statistical Learning Effect (SLE) is the RT difference of responding to high frequency (probable) versus low frequency (less probable) trials. To get a positive value, we subtracted RTs given to high frequency trials from RTs given to low frequency trials that are usually slower. As we were specifically interested in the rewiring of learned sequences, we calculated 2 different SLE-s: one for those transitions (triplets) that did not change their frequency in the Rewiring Phase ($SLE_{NO\ REWIRING} = RT_{LL} - RT_{HH}$ in both phases), and one for those that changed their frequency in the Rewiring Phase ($SLE_{REWIRING} = RT_{LH} - RT_{HL}$ in the Learning Phase, and $SLE_{REWIRING} = RT_{HL} - RT_{LH}$ in the Rewiring Phase), so that we obtained a positive value whenever participants showed learning of the currently valid statistical structure (Supplementary Figure **SF-IV/2/b**). The higher the SLE, the bigger the difference between RTs given to the more probable stimuli in contrast to the less probable stimuli under the current circumstances. In the Learning Phase, theoretically, there was no reason for SLEs to differ in magnitude as a function of later rewiring ($SLE_{NO\ REWIRING}$ vs. $SLE_{REWIRING}$). However, in the Rewiring Phase the two types of SLEs may differ if participants experience difficulties modifying their skill.

SM-IV/1.5. Calculation of anticipatory errors

To test whether participants learned to anticipate the most probable endings of triplets we looked at erroneous responses and classified each error either as being *nonanticipatory* (resulting in a low frequency triplet), *anticipation of sequence A* (when the erroneous key press completed a triplet that was frequent during the Learning Phase, although the stimulus was on another location), *anticipation of sequence B* (when the erroneous key press completed a triplet that was frequent during the Rewiring Phase, although the stimulus was on another location) or *anticipation of both sequences* (in the rare case when the resulting triplet is high frequency during both phases). As before, we only analysed those errors that were given to random elements intervening the pattern elements (leaving out the initial preparatory random elements in each block); and we excluded those errors that were given on trills or repetitions. If all the errors were independent of learning (i.e. they occurred randomly), then by chance 16.67% of them were expected to be anticipatory of sequence A; another 16.67% of them were expected to be anticipatory of sequence B; 5.56 % were expected to be anticipatory of both Sequence A and Sequence B; and the remaining 61.11% were expected to be nonanticipatory. A different error proportion was expected in the case of the probe epochs: owing to inclusion of the pattern elements (which always corresponded to high frequency triplets of the current sequence), chance levels for anticipations of the alternative sequence grew substantially. Specifically, the chance level for anticipatory errors of Sequence A in the Learning Phase was only 7.41%, while the chance level for anticipations of the other sequence (Sequence B) was 21.42% - and the reversed pattern hold for the Rewiring Phase. Our critical measure was whether anticipatory errors were more numerous than expected by chance, and whether the proportion of anticipations of Sequence A and Sequence B corresponded to the expected pattern - so in our analysis we only included *anticipatory errors of Sequence A* and *anticipatory errors of Sequence B* (leaving out anticipatory errors of both sequences and nonanticipatory errors). It must be noted that both kinds of anticipations may be above chance level, but these measures are related (to each other and to the other two kinds of errors). If one kind of errors is more numerous, it lowers the proportion of other kinds. Also, this measure does not tell anything about the total number of errors. Participants without errors on some epochs are excluded from this analysis due its within subject nature, as error proportions cannot be calculated for these epochs.

SM-IV/1.6. Tests for assessing the explicit knowledge about the sequence structures

SM-IV/1.6.1. Free Generation Task

According to the process dissociation framework (Jacoby, 1991), intentional and automatic, nonintentional forms of memory can be separated by asking participants to a) intentionally include the learned material in their responses – the *inclusion condition*, and b) to intentionally exclude the previously learned material from their responses – the *exclusion condition*. If participants nevertheless include the learned material in the exclusion condition, knowledge of this material should be considered implicit; while performance in the inclusion condition is affected by both implicit and explicit knowledge. Comparing the performance under the two conditions can give us an estimate about the explicitness of the learning. Destrebecqz & Cleeremans (2001) proposed a 'free generation' task as a form of the process dissociation procedure specifically adapted to sequence learning paradigms. In the free generation task, stimuli appeared on the screen as a result of the corresponding buttonpresses, not the other way around. Thus, participants *generated the sequence,* both under inclusion and exclusion conditions.

As participants learned two sequences during the Learning and Rewiring Phase, they performed the free generation task twice on the third day, after the completion of the ASRT task: once for Sequence A and once for Sequence B, in random order. The stimuli that appeared in the free generation task looked exactly like those seen during the Learning and Rewiring Phases and mimicked the original task conditions: if a sequence was learned explicitly, stimuli appearing in the free generation task also alternated between the two colors. If a sequence was learned implicitly, stimuli appearing in the free generation task were always presented in the same color.

In our *free generation task* both the inclusion and exclusion conditions consisted of 4 blocks of 27 trials (that is, 25 triplets) each. Between the blocks, participants could pause for a few seconds if they needed. This way we obtained 100 generated triplets for both the inclusion and the exclusion conditions; the percentage of high frequency triplets could be easily calculated simply by counting the triplets that were high frequency transitions during learning. The question was whether participants differed in the explicitness they showed by this procedure as a function of the learning conditions of the particular sequence (explicit or implicit).

Even if a sequence remained entirely implicit, participants could have some general knowledge about it, for instance, that runs of three identical stimuli (e.g., 111, 222) were rare or runs of four identical stimuli (e.g., 1111, 2222) never occurred, etc. Performance in the free generation task could be affected by this knowledge. For example, if a participant pressed the same button in many consecutive trials, s/he could be sure that the resulting sequence was fundamentally different from the learned sequence. Such strategies result in data that could not be interpreted as reflecting explicit knowledge about the statistical structure of sequence (i.e., high vs. low frequency triplets), thus we excluded participants who pressed the same button in at least 50% of the free generation trials, and those who did not press at least one of the response buttons at all during the free generation trials (although these were somewhat arbitrary exclusion criteria; for example, when participants pressed the same button

49% of the trials had to be included by these criteria, even if we know that in the ASRT task all four stimuli occur equally often – 25% of trials). Our algorithm sure does not eliminate *all* strategies that may confound the measurement of implicit and explicit knowledge in the task, but at least eliminates those that most robustly affected the resulting sequences. This way 24 Implicit-Implicit, 19 Implicit-Explicit, and 18 Explicit-Explicit participants remained in the analysis.

SM-IV/1.6.2. Triplet Sorting Task

In our main analysis we were interested in participants' ability to learn the statistical structure resulting from an ASRT sequence (some triplets being frequent, other triplets being infrequent). Thus, we used a triplet sorting task (Song, Howard, & Howard, 2008) in which we presented all the 64 possible triplets (4x4x4) to participants – in each case, runs of three consecutive trials appeared on the screen, all initiated by the computer one after the other. Stimuli were identical in location and size to those seen during sequence learning – but the color of stimuli were always grey (independently of sequence learning conditions). When the presentation of the triplet was over, we asked participants to categorize that triplet either as a high frequency triplet or a low frequency triplet. As our participants actually learned two sequences (Sequence A and Sequence B), participants completed two triplet sorting tasks on the third day after the completion of the ASRT task: one for Sequence A and one for Sequence B, in random order.

SM-IV/1.7. Statistical analysis

Results were obtained using Mixed Design ANOVAs and Bonferroni-corrected post hoc tests if the omnibus ANOVA showed significant main effects or interactions. Sphericity was assessed with Mauchly's Test, and if this precondition was not met, degrees of freedom were adjusted with the Greenhouse-Geisser method. Partial eta squared effect sizes are reported for significant main effects and interactions in ANOVA. Cohen's d is reported for post hoc pairwise comparisons.

Supplementary Results IV/2.

SR-IV/2.1. Dynamics of the rewiring process in the experimental epochs compared across the Learning and Rewiring Phase

SR-IV/2.1.1. Statistical Learning Effect (SLE)

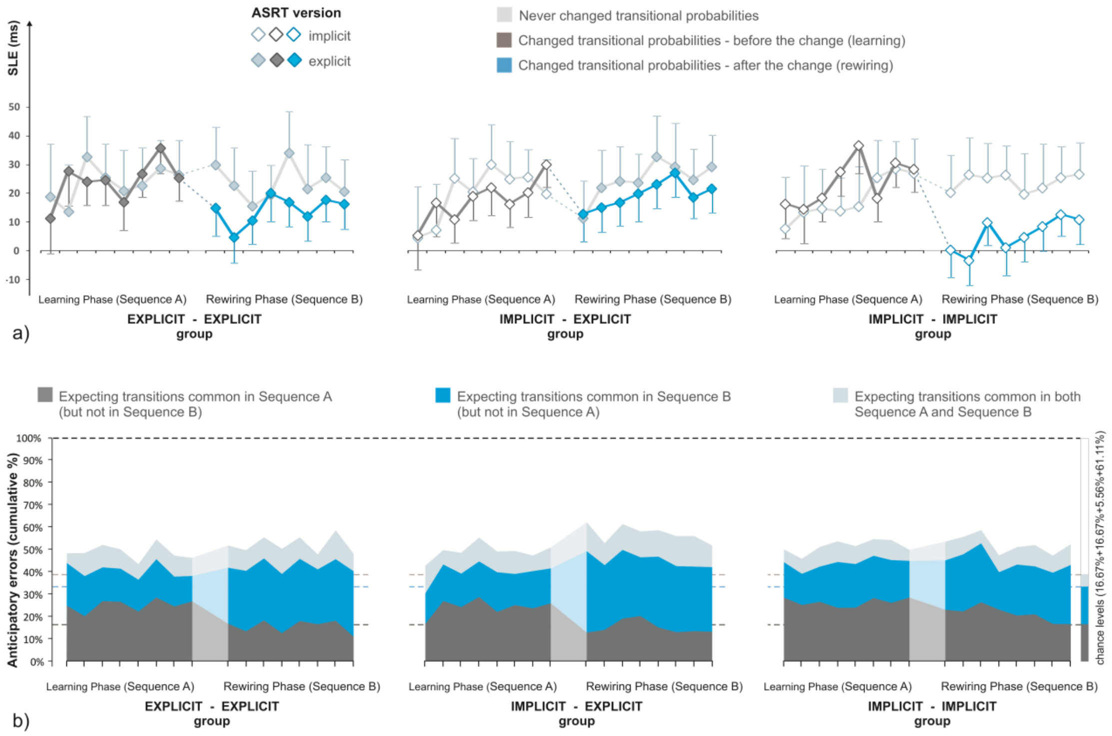
To analyse learning and rewiring on the first two days of the Experiment (see Supplementary Figure **SF-IV/3/a**), a 2 x 8 x 2 x 3 Mixed Design ANOVA was conducted on SLE-s with **PHASE** (Learning Phase or Rewiring Phase), **EPOCH** (1-8), **REWIRING** (change or no change in the frequency of particular transitions; $SLE_{NO\ REWIRING}$ vs. $SLE_{REWIRING}$) as within subject factors, and **GROUP** (Implicit-Implicit, Implicit-Explicit, Explicit-Explicit) as a between subject factor.

There was a significant main effect of **EPOCH**, $F(5.584, 446.718) = 6.554$, $MSE = 1073.133$, $p < .001$, $\eta_p^2 = .076$, as SLE-s increased as learning progressed in the two Phases. The significant main effect of **REWIRING**, $F(1, 80) = 9.604$, $MSE = 1649.252$, $p = .003$, $\eta_p^2 = .107$, showed that on average SLE-s of the changed part of the sequence were lower than SLEs for the unchanged part of the sequence ($d = .446$). The **PHASE x GROUP** interaction also reached significance, $F(2, 80) = 3.353$, $MSE = 1793.348$, $p = .040$, $\eta_p^2 = .077$. Post hoc tests showed that the average SLEs were significantly lower in the Rewiring Phase than in the Learning Phase in the case of the Implicit-Implicit group ($p = .031$, $d = .575$), and there was a trend toward the same effect in the Explicit-Explicit group ($p = .091$, $d = .457$) but not in the Implicit-Explicit group ($p = .241$, $d = .309$, the difference being in the other direction). Also, average SLEs of the Rewiring Phase in the Implicit-Implicit group were significantly lower than the same in the Implicit-Explicit group ($p = .022$, $d = .757$). No other paired comparisons reaching statistical significance (all *p*s > .373, all *d*s < .433).

There was a significant interaction of **PHASE x REWIRING**, $F(1, 80) = 19.604$, $MSE = 1283.176$, $p < .001$, $\eta_p^2 = .197$. Post hoc tests showed that there was no difference between average SLEs of the (later) changed transitions and the unchanged transitions of the sequence in the Learning Phase ($p = .568$, $d = .080$). In the Rewiring Phase, on the other hand, the SLEs for the changed transitions were significantly lower ($p < .001$, $d = .829$). Also, SLEs for the changed part of the sequence was on average lower in the Rewiring Phase than in the Learning Phase ($p < .001$, $d = .729$), that is, rewiring was not as effective as the original learning, while performance of the unchanged transitions was comparable in the two phases ($p = .151$, $d = .021$).

Most importantly, there was a significant **PHASE** x **REWIRING** x **GROUP** interaction, $F(2, 80) = 4.951$, $MSE = 1283.176$, $p = .009$, $\eta_p^2 = .110$, showing that the previously described difficulty in rewiring the original skill was not homogenous in the three groups. Post hoc tests revealed that SLEs for the rewired sequence part in the Rewiring Phase were smallest in the Implicit-Implicit group, being significantly lower than in the Explicit-Explicit ($p = .028$, $d = .743$) or Implicit-Explicit groups ($p < .001$, $d = 1.198$). The disadvantage of the rewired sequence part in the Rewiring Phase (in contrast to the unchanged sequence part in the same Phase) were apparent in the Implicit-Implicit ($p < .001$, $d = 1.425$) and the Explicit-Explicit groups ($p = .008$, $d = .737$), but not in the Implicit-Explicit group ($p = .128$, $d = .406$). Finally, the rewired SLEs in the Rewiring Phase were smaller than the original learning of these transitions in the Learning Phase for both the Implicit-Implicit ($p < .001$, $d = 1.562$) and Explicit-Explicit ($p = .028$, $d = .850$) groups, but again, such difficulty in rewiring was not apparent in the Implicit-Explicit group ($p = .561$, $d = .155$). There was no other significant main effect or interaction (all ps > .117, all $\eta_p^2 < .032$).

In summary, the Implicit-Implicit group had more difficulty in rewiring the original skill than the other groups – thus rewiring of such skills may benefit from explicit knowledge about the sequence's structure. As shown by 95% confidence intervals (see the blue line on Supplementary Figure **SF-IV/3/a**) rewiring started later in the Implicit-Implicit group, and was not statistically significant in the first half of the Rewiring Phase. In contrast, rewiring was evident as early as the first epoch of the Rewiring Phase in the case of the Implicit-Explicit and Explicit-Explicit groups.



SF-IV/3. Learning and rewiring – detailed graphs. **(a)** The magnitude of Statistical Learning Effect (SLE) indicates the difference of RTs given to frequent transitions (probable stimuli) in contrast to rare transitions (less probable stimuli). Some of the transitions had constant frequency in the Learning Phase and Rewiring Phase (unchanged transitions, dark grey line), while other transitions swapped their frequency – previously rare transitions became frequent in the Rewiring Phase, and vice versa (changed transitions). Adapting to the changed statistical structure in the Rewiring Phase was shown to be more difficult than learning the contingencies in the first place in the Learning Phase. This was shown by SLEs being – on average – lower for the changed transitions after the change in frequencies took place in the Rewiring Phase (blue line) than before the change (dark grey line). This difficulty was most pronounced in the Implicit-Implicit group, and less pronounced in the Implicit-Explicit group. Error bars represent 95% confidence intervals. **(b)** When less probable stimuli came up, participants sometimes errorneously pressed the key corresponding to the most probable stimulus – these errors are called anticipatory errors. As two (partly) different sequences were taught, we differentiated between anticipations of Sequence A's most probable stimuli (dark grey area), that of Sequence B's most probable stimuli (blue area), and those that could be considered as anticipations of both (light grey area). Chance levels for anticipatory errors are shown by the dotted lines. We were mainly interested in anticipations that either corresponded to Sequence A or to Sequence B (dark grey and blue areas). Each group showed adaptation to the current sequence, as anticipations for Sequence A were above chance level in the Learning Phase, while anticipations of Sequence B were above chance level in the Rewiring Phase. The Implicit-Implicit group additionaly showed above chance level anticipations of Sequence A during the Rewiring Phase, indicating the continuing influence of their knowledge gained in the Learning Phase.

SR-IV/2.1.2. Anticipatory Errors

To look at anticipations of Sequences A and B during the Learning Phase and the Rewiring Phase (see Supplementary Figure **SF-IV/3/b**), a 2 x 8 x 2 x 3 Mixed Design ANOVA was conducted with **PHASE** (Learning Phase or Rewiring Phase), **EPOCH** (1-8) and **ANTICIPATION** (anticipation of Sequence A vs. anticipation of Sequence B) as within subject factors and **GROUP** (Implicit-Implicit, Implicit-Explicit and Explicit-Explicit) as a between subjects factor.

There were no significant main effects (all *p*s > .113, all η_p^2 < .036). The interaction of **ANTICIPATION x GROUP** showed a trend towards significance, $F(2, 71) = 2.544$, $MSE = 904.244$, $p = .086$, $\eta_p^2 = .067$. Post hoc tests revealed that, overall, anticipations of Sequence A and Sequence B did not differ in either group (all *p*s > .106, all *d*s < .548); however, there was a difference between the Implicit-Implicit and Implicit-Explicit groups when anticipating transitions common in Sequence A: the Implicit-Implicit group made more such anticipations then the Implicit-Explicit group ($p = .050$, $d = .714$). No other paired comparison reached significance, all *p*s > .163, all *d*s < .600).

The **PHASE x ANTICIPATION** interaction was significant, $F(1, 71) = 86.707$, $MSE = 572.289$, $p < .001$, $\eta_p^2 = .550$. As expected (and as shown by post hoc comparisons), anticipations of Sequence A were more pronounced in the Learning Phase than anticipations of Sequence B, and vice versa in the Rewiring Phase (both $p < .001$, both $d > 1.061$). During the time course of learning, anticipations of Sequence A were more common in the Learning Phase than in the Rewiring Phase; while anticipations of Sequence B were more common in the Rewiring Phase than in the Learning Phase (both $p < .001$, both $d > .979$). Most importantly, the **PHASE x ANTICIPATION x GROUP** interaction was also significant, $F(2, 71) = 3.917$, $MSE = 572.289$, $p = .024$, $\eta_p^2 = .099$. Post hoc tests revealed that the previously described pattern was observed in all experimental groups, although effect sizes were substantially smaller in the case of the Implicit-Implicit group (both $d < 0.672$) than in the other groups (all *d*s > 1.226). Looking at the Learning Phase alone, groups did not differ in terms of anticipations of Sequence A (all *p*s > .999, *d*s < .263) and Sequence B (all *p*s > .999, *d*s < .285). In all groups, anticipations of Sequence A outnumbered anticipations of Sequence B in this Phase (all *p*s < .001, all *d*s > 1.058). Looking at the Rewiring Phase alone, however, showed us a more complicated pattern. In this phase anticipations of Sequence B outnumbered anticipations of Sequence A only in the Explicit-Explicit and Implicit-Explicit groups (both $p < .001$, both $d > .1.133$) but not in the Implicit-Implicit group ($p = .529$, $d = .225$). Anticipations of Sequence B were significantly less common in the Implicit-Implicit group than in the Implicit-Explicit group ($p = .047$, $d = .721$), while anticipations of Sequence A were significantly more common in the Implicit-Implicit group than both in the Implicit-Explicit and the Explicit-Explicit groups (both $p < .036$, both $d > .795$). The latter findings may be interpreted as a shift toward anticipating Sequence A at the expense of anticipating Sequence B in the case of the Implicit-Implicit group.

SR-IV/2.2. Testing the efficiency of the rewiring process in the experimental epochs of the Follow-up Phase

On the third day of the experiment, participants performed the sequence that was learned in the Learning Phase (Sequence A) as well as the sequence that was learned in the Rewiring Phase (Sequence B). The order of these mini-epochs was counterbalanced across participants. We were interested in the consolidation of their knowledge about the statistical structure of the task, and particularly, whether they could re-adapt to the characteristics of the Learning Phase or not (that is, whether the statistical learning on the first day was overwritten by the statistical learning on the second day, or they existed in parallel). This analysis provides information about retroactive interference effects (in addition to proactive effects assessed earlier).

SR-IV/2.2.1. Statistical Learning Effect (SLE)

We conducted a 2 x 2 x 3 Mixed Design ANOVA on SLEs with **SEQUENCE** (the same conditional probabilities as in the Learning Phase - that is, Sequence A; or the same conditional probabilities as in the Rewiring Phase - that is, Sequence B) and **REWIRING** (change or no change in the frequency of particular transitions) as within subject factors, and **GROUP** (Implicit-Implicit, Implicit-Explicit, and Explicit-Explicit) as a between subject factor.

There was a significant main effect of **SEQUENCE,** $F(1, 79) = 6.245$, $MSE = 891.475$, $p = .015$, $\eta_p^2 = .073$, as higher statistical knowledge was expressed for Sequence B (that is, for the conditional probabilities that corresponded to the Rewiring

Phase on the second day of the study), than for Sequence A (*d* = .404). There was also a significant main effect of **REWIRING**, *F*(1, 79) = 9.574, *MSE* = 943.423, *p* = .003, η_p^2 = .108, as higher statistical knowledge was expressed for those transitions that never changed their frequency of occurrence than for those that changed (*d* = .506). No other main effect or interaction reached significance, all *p*s > .284, η_p^2 < .031.

SR-IV/2.2.2. Anticipatory errors

There were two participants who made no errors when performing one of the Sequences - analysis was conducted on the remaining 27 (Explicit-Explicit), 28 (Explicit-Implicit) and 26 (Implicit-Implicit) participants. Chance level for both kinds of anticipations (anticipations of Sequence A and anticipations of Sequence B) was 16.67%. To assess anticipations on the third day of the study, a 2 x 2 x 3 Mixed Design ANOVA was conducted with **SEQUENCE** (Sequence A vs. Sequence B) and **ANTICIPATION** (anticipation of transitions common to Sequence A only vs. anticipation of transitions common to Sequence B only) as within subject factors and **GROUP** (Implicit-Implicit, Implicit-Explicit, and Explicit-Explicit) as a between subjects factor.

There was a trend towards a main effect of **GROUP**, *F*(2, 78) = 2.582, *MSE* = 292.754, *p* = .082, η_p^2 = .062. However, post hoc pairwise comparisons revealed no significant differences between groups, all *p*s >.141, *d*s < .574. We also found a significant interaction of **SEQUENCE x ANTICIPATION**, *F*(1, 78) = 13.815, *MSE* = 596.817, *p* < .001, η_p^2 = .150. Post hoc tests revealed that when performing Sequence A, anticipations of Sequence A were more common than anticipations of Sequence B (*p* = . 004, *d* = .533), and than what might have been expected by chance, *CI95%* [24.699, 34.391]. Anticipations of Sequence B did not differ from chance level, *CI95%* [15.318, 22.844]. When performing Sequence B, on the other hand, anticipations of Sequence B outnumbered anticipations of Sequence A (*p* = .009, *d* = .503), and were more numerous than expected by chance, *CI95%* [23.822, 32.291]. Anticipations of Sequence A did not differ from chance level, *CI95%* [14.024, 22.644]. From another point of view, anticipations of Sequence A were significantly more pronounced when performing Sequence A than when performing Sequence B, and vice versa (both *p* < .003, *d* > .494). This pattern of results indicate no proactive or retroactive interference, as participants were able to quickly adapt to the changed statistical conditions.

SR-IV/2.3. Dynamics of the rewiring process in the probe epochscompared across the Learning and Rewiring Phase

SR-IV/2.3.1. Statistical Learning Effect (SLE)

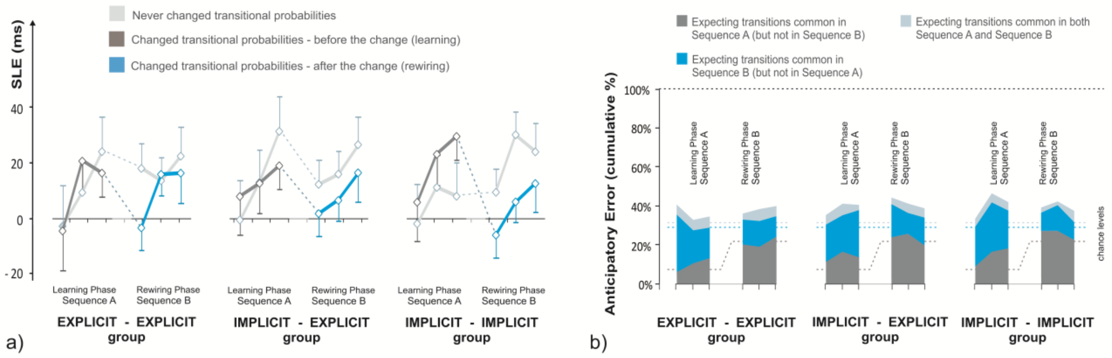
SLEs were calculated for the probe epochs of the Learning and Rewiring Phase. During these short epochs at the beginning, in the middle and at the end of each phase, trials were not cued, thus the task remained implicit for all participants. This manipulation made it possible to compare groups under uniform experimental conditions. Also, as there were no cued trials during these epochs, there was no need to exclude pattern trials. Correspondingly, although these epochs were substantially shorter than the experimental epochs, about the same number of trials were analysed. It must be noted, though, that the proportion of high frequency (more probable) and low frequency (less probable) combinations are different if the analysis includes pattern trials, as in this case there are more high frequency triplets than low frequency triplets, making the median RTs of the latter a bit noisier then the former.

Similarly as before, a 2 x 8 x 2 x 3 Mixed Design ANOVA was conducted on SLEs shown in Supplementary Figure **SF-IV/4/a** with **PHASE** (Learning Phase vs. Rewiring Phase), **PROBE EPOCH** (1-3) and **REWIRING** (change or no change in the frequency of particular transitions) as within subject factors, and **GROUP** (Implicit-Implicit, Implicit-Explicit, and Explicit-Explicit) as a between subject factor.

The main effect of **PROBE EPOCH** was significant, $F(1.633, 130.680) = 34.447$, $MSE = 947.431$, $p < .001$, $\eta_p^2 = .301$, as statistical knowledge indicated by SLEs became higher as learning progressed. The main effect of **REWIRING** showed a trend towards significance, $F(1, 80) = 2.870$, $MSE = 1157.559$, $p = .094$, $\eta_p^2 = .035$, as overall SLEs were higher for the never changed parts of the sequence than for the changed part ($d = .224$). We found a significant **PHASE x REWIRING** interaction, $F(1, 80) = 18.905$, $MSE = 858.840$, $p < .001$, $\eta_p^2 = .191$. Post hoc tests revealed that although there was no difference between the later to be changed transitions and the unchanged transitions during the Learning Phase ($p = .166$, $d = .202$), there was a substantial advantage of the unchanged transitions during the Rewiring Phase ($p < .001$, $d = .737$). During the time course of learning, there was a significant improvement in the performance on the unchanged parts of the sequence from the Learning Phase to the Rewiring Phase (learning for these transitions continued in the Rewiring Phase, $p = .001$, $d = .432$); while there was a significant drop in statistical knowledge for the rewired sequence parts in the Rewiring Phase (compared to the original learning in the Learning Phase, p = .009, d = .414).

There was also a significant **PHASE x REWIRING x GROUP** interaction, F(2, 80) = 3.908, MSE = 858.840, p = .024, η_p^2 = .089, meaning that the previously described pattern was not homogenous in the three groups. Post hoc comparisons revealed that the Implicit-Implicit group showed an advantage of the later to be rewired sequence parts during the Learning Phase (p = .013, d = .654), while no difference between the later changed and unchanged sequence parts were observed in the other two groups (both p > .809, both d < .063). The difficulty of rewiring was apparent in all three groups, as participants responded faster to the unchanged transitions than to the recently changed transitional probabilities in the Rewiring phase (Explicit-Explicit group: p = .060, d = .541; Implicit-Explicit group: p = .023, d = .645, Implicit-Implicit group: p < .001, d = 1.097). It could also be observed that learning of the unchanged transitional probabilities continued in the Rewiring Phase, as statistical knowledge for these transitions were higher in the Rewiring Phase than in the Learning Phase for all groups (although this pattern was significant only in the Implicit-Implicit group: p = .001, d = .767; a trend was observed for the Explicit-Explicit group: p = .083, d = .389, and nonsignificant in the Implicit-Explicit group: p = .405, d = .182). Statistical learning for the recently changed transitional probabilities, on the other hand, was smaller than the original learning before rewiring for all groups (although it only reached significance in the Implicit-Implicit group: p = .001, d = .913; all other p > .287, d < .296). In addition, there was an interaction of **PROBE EPOCH x REWIRING x GROUP**, F(4, 160) = 2.969, MSE = 649.224, p = .021, η_p^2 = .069, indicating that the previously described effects varied as a function of probe epochs in the Learning and Rewiring Phases. A detailed graph depicting all levels of this interaction is shown in Supplementary Figure **SF-IV/4/a**.

In summary, analysis of the probe epochs strengthened our results observed in the experimental epochs. Rewiring of recently changed transitional probabilities was shown to be harder for the Implicit-Implicit group than for the other two groups who rewired with the help of explicit cues, although – by the end of the Rewiring Phase – all three groups showed adaptation to the new statistical structure (shown by 95% CIs on the blue lines on Supplementary Figure **SF-IV/4/a**).



SF-IV/4. Learning and Rewiring in the Probe epochs – detailed graphs. **(a)** Adapting to the changed statistical structure in the Rewiring Phase was shown to be more difficult than learning the contingencies in the first place in the Learning Phase. This was shown by SLEs being – on average – lower for the changed transitions after the change in frequencies took place in the Rewiring Phase (blue line) than before the change (dark grey line). Error bars represent 95% CIs. **(b)** Anticipatory errors either corresponded to Sequence A or to Sequence B (dark grey and blue areas). Participants could adapt to both statistical regularities as anticipations of Sequence A were higher than expected by chance when performing Sequence A in the Learning Phase; similarly, anticipations of Sequence B were higher then expected by chance when performing Sequence B in the Rewiring Phase. Chance levels for anticipatory errors are shown by the dotted lines.

SR-IV/2.3.2. Anticipatory errors

Participants without errors on some probe epochs were excluded from the analysis due its within subject nature, as error-proportions could not be calculated for these probe epochs. Analysis was performed on the remaining 23 (Explicit-Explicit), 26 (Implicit-Explicit) and 23 (Implicit-Implicit) participants. We conducted a 2 x 3 x 2 x 3 Mixed Design ANOVA on the anticipatory data shown in Supplementary Figure **SF-IV/4/b** with **PHASE** (Learning Phase vs. Rewiring Phase), **PROBE EPOCH** (1-3) and **ANTICIPATION** (anticipation of Sequence A vs. anticipation of Sequence B) as within subject factors and **GROUP** (Implicit-Implicit, Implicit-Explicit, and Explicit-Explicit) as a between subject factor.

The ANOVA revealed only one significant interaction: a **PHASE x ANTICIPATION** interaction, $F(1, 69) = 66.157$, $MSE = 308.436$, $p < .001$, $\eta_p^2 = .489$. Post hoc tests revealed that – as expected even by chance levels – there were significantly more anticipations of Sequence B than anticipations of Sequence A during the Learning Phase; and vice versa (both $p < .001$, $d > .765$); furthermore, 95%

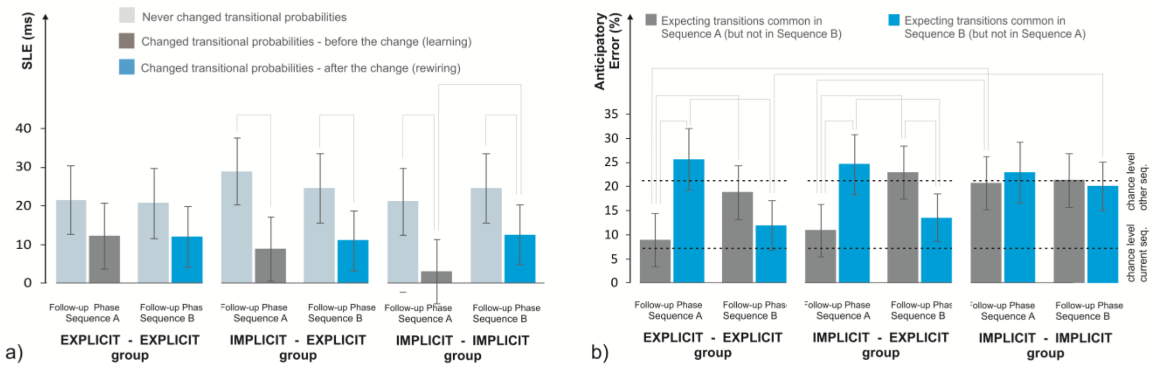
confidence intervals indicated that in the Learning Phase the anticipations of Sequence A were more pronounced than expected by chance, *CI95%* [11.139, 14.752], while anticipations of Sequence B did not differ from what we might have expected by chance, *CI95%* [18.091, 24.229]. In the Rewiring Phase, anticipatory errors of Sequence A did not differ from what we might have expected by chance, *CI95%* [20.760, 26.173]; while anticipations of Sequence B were more numerous than expected by chance, *CI95%* [9.894, 14.525]. Thus, anticipatory errors indicated that participants indeed learned to anticipate the most probable continuation of the previous trials both before and after rewiring the transitional probabilities. Post hoc tests also indicated that anticipations of Sequence A were of greater proportion during the Rewiring Phase than during the Learning Phase, and vice versa for anticipations of Sequence B (both $p < .001$, $d > .773$) – this pattern was expected even by chance, thus not providing additional information to our evaluation of the results. Critically, the interaction of **PHASE x ANTICIPATION x GROUP** was not significant, $F(2, 69) = 0.580$, $MSE = 308.436$, $p = .563$, $\eta_p^2 = .017$, suggesting that the previously described pattern was similar across the three experimental groups. This result did not rule out the possibility that differences existed, though, as the pattern can be very similar across groups even if some anticipations are above or below chance levels for some of the groups – this being the most important information we tried to uncover. Finally, there was a trend towards a **PROBE EPOCH x ANTICIPATION x GROUP** interaction, $F(4, 138) = 2.013$, $MSE = 300.555$, $p = .096$, $\eta_p^2 = .055$. A detailed graph depicting all levels of the interaction is shown in Supplementary Figure **SF-IV/4/b**.

SR-IV/2.4. Testing the efficiency of the rewiring process in the probe epochs of the Follow-up Phase

SR-IV/2.4.1 Statistical Learning Effect (SLE)

SLEs were calculated for probe epochs of the Follow-up Phase as previously described for the Learning and Rewiring Phase. We conducted a 2 x 2 x 3 Mixed Design ANOVA on SLEs shown in Supplementary Figure **SF-IV/5/a** with **SEQUENCE** (Sequence A vs. Sequence B) and **REWIRING** (change or no change in the frequency of particular transitions) as within subject factors, and **GROUP** (Implicit-Implicit, Implicit-Explicit, and Explicit-Explicit) as a between subject factor.

There was a significant main effect of **REWIRING**, $F(1, 80) = 26.449$, $MSE = 585.105$, $p < .001$, $\eta_p^2 = .248$, as higher statistical knowledge was expressed for those transitions that never changed their transitional probability than for those that changed ($d = .852$). No other main effect or interaction reached significance, all $ps > .313$, $\eta_p^2 < .030$. This pattern of results indicates only proactive but no retroactive interference.



SF-IV/5. Consolidation of learning in the Probe epochs of the Follow-up Phase. **(a)** The magnitude of SLE indicates the difference of RTs given to frequent transitions (probable stimuli) in contrast to rare transitions (less probable stimuli). Some of the transitions had constant frequency in both Sequences (unchanged transitions, dark grey line), while other transitions were frequent in only one of the Sequences, not in the other (changed transitions; dark grey bars for Sequence A and blue bars for Sequence B). Performance was better on the unchanged transitions than on changing transitions in all groups. No group differences were observed. **(b)** Anticipatory errors of Sequence A's most probable stimuli are shown in dark grey bars, and that of Sequence B's most probable stimuli are shown in blue bars. Chance levels for anticipatory errors are shown by the dotted lines (there is a lower chance level of anticipatory errors of Sequence A when performing Sequence A than when performing Sequence B, and vice versa). Participants could readapt to both statistical regularities as anticipations of Sequence A were higher than expected by chance when performing Sequence A in the Follow-up Phase; similarly, anticipations of Sequence B were higher than expected by chance when performing Sequence B in the Follow-up Phase. Error bars represent 95% CIs.

SR-IV/2.4.2 Anticipatory errors

Anticipatory errors for the probe epochs of the Follow-up Phase were calculated as previously described for the Learning and Rewiring Phase. A 2 x 2 x 3 Mixed Design ANOVA on the anticipatory errors shown in Supplementary Figure **SF-IV/5/b** was conducted with **SEQUENCE** (Sequence A vs. Sequence B) and **ANTICIPATION** (anticipation of transitions common to Sequence A only vs. anticipation of transitions common to Sequence B only) as within subject factors and **GROUP** (Implicit-Implicit, Implicit-Explicit, and Explicit-Explicit) as a between subjects factor.

The ANOVA revealed a significant main effect of **GROUP**, $F(2, 79) = 3.566$, $MSE = 193.010$, $p = .033$, $\eta_p^2 = .083$. Post hoc tests showed that this was caused by the Implicit-Implicit group making – on a trend level - more anticipations of Sequence A and B (overall) than the Explicit-Explicit group ($p = .070$, $d = .705$). No other paired comparison reached statistical significance, both $p > .434$, $d < .444$. We also found a significant **SEQUENCE x ANTICIPATION** interaction, $F(1, 79) = 17.558$, $MSE = 324.940$, $p < .001$, $\eta_p^2 = .182$. Post hoc tests revealed a pattern consistent with chance levels, that is, higher levels of anticipations of Sequence B when performing Sequence A, and vice versa (both $p < .011$, $d > .468$); and higher levels of Sequence B during Sequence A than during Sequence B - and vice versa (both $p < 0.03$, $d > .503$). Solely on the basis of the ANOVA we could not infer anything about anticipatory errors; confidence intervals, on the other hand, provide some interesting details. Anticipations of Sequence A when performing Sequence A exceeded chance levels, *CI95%* [10.856, 17.527], and so did anticipations of Sequence B when performing Sequence B, *CI95%*[11.598, 18.029]. Anticipations of Sequence A, on the other hand, did not exceed chance levels when performing Sequence B, *CI95%*[18.067, 25.015], and vice versa, *CI95%*[20.679, 28.599].

Finally, there was a trend towards a **SEQUENCE x ANTICIPATIONS x GROUP** interaction, $F(2, 79) = 2.777$, $MSE = 324.940$, $p = .068$, $\eta_p^2 = .066$, indicating that the previously described pattern of results was not the same in the three experimental groups. Post hoc tests revealed that the pattern expected by chance was apparent in the Explicit-Explicit and Implicit-Explicit groups (all *p*s < .039, all *d*s > .658), but not in the Implicit-Implicit group where anticipations of Sequence A and Sequence B were equally high both when performing Sequence A and when performing Sequence B (all *p*s > .551, *d*s < .213); additionally, 95% CIs indicated that the Implicit-Implicit group showed above-chance level of anticipations of Sequence A when performing Sequence A, *CI95%*[15.069, 27.189], and above-chance level of anticipations of Sequence B when performing Sequence B, *CI95%*[14.050, 25.736]. Based on confidence intervals, the Implicit-Explicit group also showed above-chance level anticipations of Sequence B when performing Sequence B, *CI95%* [8.377, 19.126].

In addition to these effects, there were differences between groups regarding the proportion of anticipations of Sequence A (when performing Sequence A); the Implicit-Implicit group showed significantly higher rates than the Explicit-Explicit group ($p =$

.023, d = 836), and – on a trend level - higher rates than the Implicit-Explicit group (p = .080, d = .680). Also, there were differences between groups regarding the proportion of anticipations of Sequence B when performing Sequence B; the Implicit-Implicit group showed higher rates than the Explicit-Explicit group (on a trend level, p = .080, d = 688). No other paired comparison reached significance, all ps > .381, all ds < .465.

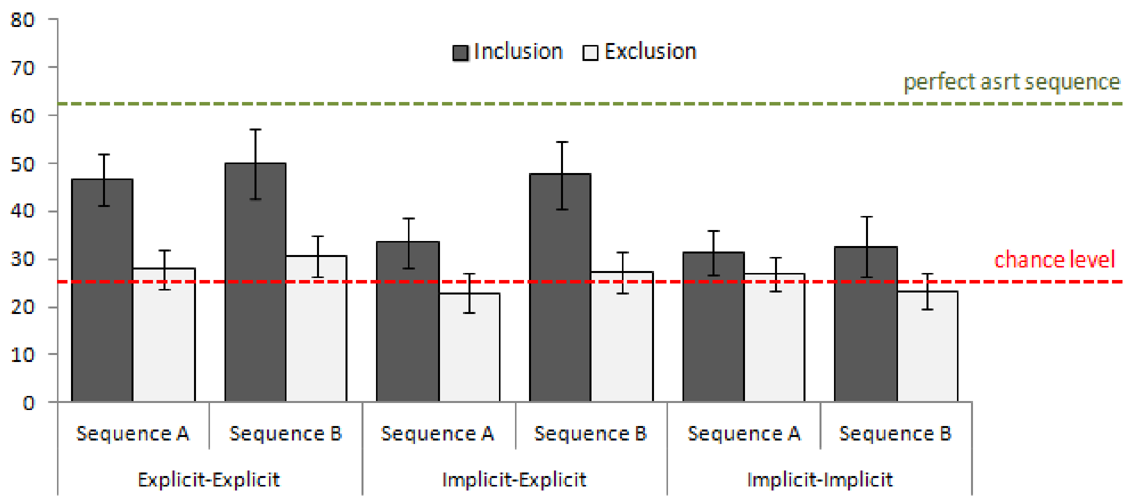
SR-IV/2.5 Testing the explicit knowledge acquired about the sequence structures

SR-IV/2.5.1. Free Generation Task

We conducted a Mixed Design ANOVA on the percentage of generated high frequency triplets with **SEQUENCE** (Sequence A vs. Sequence B) and **CONDITION** (Inclusion vs. Exclusion) as within subjects factors and **GROUP** (Implicit-Implicit, Implicit-Explicit, and Explicit-Explicit) as a between subject factor. To understand the results of such a factorial ANOVA, we have to keep in mind that the average explicitness of the task differed between the sequences and the conditions; the Explicit-Explicit group learned both sequences explicitly, thus the average explicitness was highest for this group; the Implicit-Explicit group learned Sequence A implicitly but Sequence B explicitly; while the Implicit-Implicit group learned both sequences implicitly (thus the average explicitness was lowest for this group). On the other hand, Sequence A was learned explicitly by only one of the groups (the Explicit-Explicit group), while Sequence B was learned explicitly by both the Implicit-Explicit and the Explicit-Explicit groups, making the average explicitness for Sequence B higher than that of Sequence A.

We found a significant main effect of **SEQUENCE**, $F(1, 58) = 6.611$, MSE = 122.210, p = .013, η_p^2 = .102, as participants generated – on average – more high frequency triplets for Sequence B than for Sequence A (d = .502). There was also a significant main effect of **GROUP**, $F(2, 58) = 24.295$, MSE = 88.008, p < .001, η_p^2 = .456, as the number of generated high frequency triplets was highest for the Explicit-Explicit group (*CI95%* [36.565, 40.991], significantly differing from the other groups, both p < .002, d > 1.327); and lowest for the Implicit-Implicit group (*CI95%* [26.667, 30.50], significantly differing from the other two groups, p < .014, d > .947).

More importantly, there was a significant **SEQUENCE x GROUP** interaction, $F(2, 58) = 4.623$, MSE = 122.210, p = .014, η_p^2 = .138. Post hoc tests revealed that the interaction was caused by higher explicitness if the learning was explicit than when learning was implicit (see Supplementary Figure **SF-IV/6**). That is, for Sequence A the Explicit-Explicit group outperformed the other two groups, both $p < .002$, $d > 1.272$ (the other groups did not differ from each other, $p > .999$, $d = .137$) – in accordance with this group being the only one to learn Sequence A explicitly. For Sequence B, on the other hand, a disadvantage of the Implicit-Implicit group was apparent – they generated significantly *less* high frequency triplets than the Implicit-Explicit and Explicit-Explicit groups, both $p < .002$, $d > 1.262$ (the latter two not differing from each other, $p = .832$, $d = .379$). The number of generated high frequency triplets for the two sequences did not differ from each other in the case of the Implicit-Implicit ($p = .620$, $d = .160$) and Explicit-Explicit ($p = .263$, $d = .428$) groups; but it was higher for Sequence B in the case of the Implicit-Explicit group ($p = .001$, $d = 1.330$) – as this was the only group where learning instructions differed for the two sequences. Thus, these results show that participants indeed gained more explicit knowledge about the regularities when they performed the explicit version of the task and were asked to keep track of these regularities.



SF-IV/6. Percentages of generating high frequency triplets in the free generation task. Participants in the implicit conditions performed well below of those in the explicit conditions when they were asked to generate similar sequences as the ones that they encountered during the experiment.

We also found a significant main effect of **CONDITION**, $F(1, 58) = 74.559$, $MSE = 151.926$, $p < .001$, $\eta_p^2 = .562$, as more high frequency triplets were generated under the inclusion condition than under the exclusion condition ($d = 1.785$). In addition, there was a significant **CONDITION x GROUP** interaction, $F(2, 58) = 5.615$, $MSE = 853.092$, $p = .006$, $\eta_p^2 = .162$. Post hoc tests revealed that the previously described effect was only apparent under the inclusion condition (all *p*s < .028, all *d*s > .933), and the groups generated similar number of high frequency triplets under the exclusion condition (all *p*s > .163, all *d*s < .652). Overall, all groups generated more high frequency triplets under the inclusion condition than under the exclusion condition (all *p*s < .010, all *d*s > .900). When participants learned the sequences implicitly, this result of more high frequency triplets in the inclusion vs. exclusion conditions can be caused by a more general knowledge/belief about the task structure, not the awareness of the particular high vs. low frequency transitional probabilities per se (although we cannot totally rule this out).

Inclusion and exclusion strategies may differ for implicit and explicit learners, and consequently, the difference between inclusion and exclusion performance may mean different things in the two groups. For the explicit group, the instructions are quite straightforward as the alternating sequence was explicitly cued for them during the Learning/Rewiring Phases. Consequently, they could more easily generate the known alternating sequence (inclusion condition) and also could know what sequence to not generate (exclusion condition). For the implicit learners the instructions might have been more puzzling: they might have generated a seemingly random sequence just as they experienced throughout the task (inclusion condition) vs. a sequence intentionally so that it does not seem random (exclusion condition). These cases can lead to similar outcomes, but it does not necessarily mean explicit knowledge about the alternating sequence (actually it may indicate that the sequence was perceived absolutely random, and thus the generation under exclusion condition is something (anything) non-random. In line with this argument, Fu and collegues showed that the difference between inclusion and exclusion scores could be based either on rules and memory (more explicit processes) but also on intuition (more implicit processes) (Fu, Dienes, & Fu, 2010). Consequently, this task *alone* is not suited to measure the conscious status of structural knowledge (Gaillard, Cleeremans, & Destrebecqz, 2014).

Importantly, significant group differences were found under the inclusion condition that reflect the effect of the explicit instructions. The results of the Triplet

Sorting Task (see the next section) further supports the interpretation that knowledge of the underlying structure in the ASRT task remained implicit for participants in the implicit conditions.

SR-IV/2.5.2 Triplet Sorting Task

We were interested in how accurately participants sorted transitions (triplets) for Sequence A and Sequence B. We conducted a Mixed design ANOVA with **SEQUENCE** (Sequence A vs. Sequence B) as a within-subject factor and **GROUP** (Implicit-Implicit, Implicit-Explicit, and Explicit-Explicit) as a between subject factor. Due to technical errors, data for 7 participants were lost, all from the Implicit-Implicit group; therefore the ANOVA was conducted on the data from the remaining participants.

Overall accuracy was 50.208% (*SEM* = 0.987, *CI95%* [48.238%, 52.178%]), thus not significantly different from what we would expect by chance. There was a trend towards a main effect of **SEQUENCE**, $F(1, 69) = 2.791$, $MSE = 42.823$, $p = .099$, $\eta_p^2 = .039$, participants being on average more accurate in the case of Sequence B than Sequence A ($d = .191$) – a fact possibly reflecting that knowledge of Sequence A in part became overwritten by the knowledge of Sequence B. In spite of the trend towards significance, 95% CIs showed that average accuracy for both Sequence A and Sequence B remained around (and not differed significantly from) the 50% chance level; accuracy for Sequence A: *CI95%* [47.310, 51.272], accuracy for Sequence B: *CI95%* [48.628, 53.624]. No other main effect or interaction reached significance, all ps > .271, all $\eta_p^2 < .037$, suggesting that the groups did not differ in their average accuracy, and that the effect of Sequence on accuracy was similar across groups.

The low performance on this measure could indicate at least two things: first, it may indicate that in spite of cueing the pattern trials under explicit task conditions, the knowledge of statistical structure (the relative frequency of different transitions) remained implicit for the participants. Second, participants may have had a knowledge about some triplets being more frequent than others, but they may have not been able to tell whether a particular transition was frequent during the Learning Phase (Sequence A) or the Rewiring Phase (Sequence B). Thus, low performance might have been caused by participants sorting high frequency trials as such, but not being able to correctly differentiate between Sequence A and Sequence B. This possibility should not be ignored considering that participants – on average – sorted 55.757% as high frequency trials, although in reality only 25% of them were truly high frequency (combined over the Learning and Rewiring Phases).

If a participant knew that a particular transition was frequent during one of the phases, but classified it as high frequency for the wrong sequence (e.g. classified it as being high frequency during Sequence A while in reality it was high frequency during Sequence B), this classification appears as a false alarm reducing overall accuracy scores. So we were interested in how many of such false alarms could be detected; we rerun the previously described ANOVA, but this time the dependent variable was this false alarm rate (correct classification of a transition but for the wrong sequence). The overall rate of such errors was 18.296% (*SEM* = 0.325, *CI95%* [17.647, 18.946]), thus not significantly different from chance level (18.75%). However, we found a significant main effect of **GROUP**, $F(2, 69) = 3.946$, $MSE = 15.027$, $p = .024$, $\eta_p^2 = 0.103$. Post hoc tests revealed that the percentage of such errors was significantly lower in the case of the Implicit-Explicit group than in the case of the Implicit-Implicit group ($p = .022$, $d = .829$). The Explicit-Explicit group's false alarm rate was between these two, not significantly different from either (both $p > .312$, $d < .480$). This result suggests that the interference was highest in the Implicit-Implicit group, and lowest in the Implicit-Explicit group. In spite of these differences, none of the groups showed more false alarms than expected by chance, as the 95% CIs included the 18.75 value.

In addition, there was a trend towards a main effect of **SEQUENCE**, $F(1, 69) = 3.438$, $MSE = 17.459$, $p = .068$, $\eta_p^2 = .047$, as false alarms were more numerous in the case of Sequence A than for Sequence B ($d = .323$). In other terms, retroactive interference was higher than proactive interference, possibly reflecting that knowledge for Sequence A might become partly overwritten by knowledge for Sequence B. But again, despite these differences, false alarm rates remained around chance level (and did not differ from it significantly, as the 95% CIs included the 18.75 value). The interaction of **SEQUENCE x GROUP** was not significant, $p = .834$, $\eta_p^2 = .005$, indicating that the previously described main effect of sequence was similar across groups.

In summary, the analysis of false alarms in the triplet sorting task shows that the interference caused by learning two, partly overlapping sequences was highest in the Implicit-Implicit group. This result suggests that the explicit cues indeed can help differentiate between the two sequences and use the acquired knowledge more appropriately in the relevant context.

Supplementary Materials for Study 4

Description of Supplementary Tables ST-V/1 to ST-V/4

Tables V/1-V/4 supplement the Results (Section 1, including Fig V/3-V/5) in the main text by providing supporting statistics.

ST-V/1 **Trial Type Proportions.** For each individual (N = 180) a Chi-Square test was run to assess whether random (R) and pattern (P) trials occur with the same relative frequency in the different categories present in a Model. Effect sizes (Cramer's V) were also calculated individually. These computations were repeated using Triplet Filtering (TF) and Quad Filtering (QF). Values in the table represent the percentage of participants where the result of the Chi-Squared test was significant (*χ^2 % participant significant*) and the mean Cramer's V values and the standard deviation of these values (*Cramer's V mean (SD)*).

ST-V/2 **Trial Probability**. To assess whether the distribution of trial probabilities (assessed either on triplet level or on quad level) were equal in the categories being contrasted for a particular learning score, Kolmogorov-Smirnov tests were run for each individual (N =180). To assess the direction of difference (if a difference was observed), an additional Mann-Whitney test was also run. Finally, the AUROC (Area Under the Reciever Operating Charachteristic Curve) was also calculated individually (AUC = Mann-Whitney U / (n_1 x n_2); where n_1 and n_2 refer to the two sample sizes). This statistics gives the probability that a randomly chosen value from one sample is higher that a randomly chosen value from the other sample, which we refer to as the *Probability of Superiority* in the Table.

The four values reported are then the following: The percentage of participants experiencing significantly different distributions of trial probabilities in the contrasted categories (*KS % participant significant*); the percentage of participants experiencing significantly higher trial probabilities in the first category (*MW % participant significant (a > b));* the percentage of participants experiencing significantly higher trial probabilities in the second category (*MW % participant significant (a > b)*), and the average probability that a randomly chosen member of category b is higher than a randomly chosen member of category A, along with the standard deviation of these values (*Probability of Superiority b > a mean % across participants (SD)*).

ST-V/3 **Combination Frequency.** The same statistics were calculated and presented that are described in *Table S2.*

ST-V/4 **The Abstract Structure of the Combinations**. The same statistics were calculated and presented as in *Table S1,* the only difference being that instead of trial proportions the relative frequency of abstact categories of combinations were compared.

ST-V/1. Trial Type Proportions.		χ^2 % participant significant			Cramer's V mean (*SD*)		
		NF	TF	QF	NF	TF	QF
M1	R (a) vs. P (b) Trial Type Effect	100.00	100.00	100.00	1.00 *0.00*	1.00 *0.00*	1.00 *0.00*
M2	L (a) vs. H (b) Sequence Spec. L.	100.00	100.00	100.00	0.78 *0.01*	0.73 *0.01*	0.58 *0.02*
M3	LR (a) vs. HR (b) Pure Statistical Learn.	0.00	0.00	0.00	0.00 *0.00*	0.00 *0.00*	0.00 *0.00*
	HR (a) vs. HP (b) Higher Order Seq.L.	100.00	100.00	100.00	1.00 *0.00*	1.00 *0.00*	1.00 *0.00*
	LR (a) vs. HP (b) Maximized Learning	100.00	100.00	100.00	1.00 *0.00*	1.00 *0.00*	1.00 *0.00*
M4	L (a) vs. H1 (b) Triplet L. (+ P. L.)	100.00	100.00	100.00	0.62 *0.00*	0.58 *0.01*	0.50 *0.01*
	H1 (a) vs. H2 (b) Quad L. (+ P. L.)	100.00	100.00	100.00	0.61 *0.00*	0.61 *0.00*	0.50 *0.01*
	L (a) vs. H2 (b) Maximized Learning	100.00	100.00	100.00	1.00 *0.00*	1.00 *0.00*	1.00 *0.00*
M5	LR (a) vs. H1R (b) Triplet Learning	0.00	0.00	0.00	0.00 *0.00*	0.00 *0.00*	0.00 *0.00*
	H1R (a) vs. H1P (b) Pattern Learning	100.00	100.00	100.00	1.00 *0.00*	1.00 *0.00*	1.00 *0.00*
	H1P (a) vs. H2P (b) Quad Learning	0.00	0.00	0.00	0.00 *0.00*	0.00 *0.00*	0.00 *0.00*
	LR (a) vs. H2P (b) Maximized Learning	100.00	100.00	100.00	1.00 *0.00*	1.00 *0.00*	1.00 *0.00*

NF = No Filter, TF = Triplet Filter, QF = Quad Filter

ST-V/2. Combination Frequencies.		Triplet Level												Quad Level											
		Kolmogorov-S. % Participant significant			Mann-Whitney % Participant significant						Probability of Superiority (b > a) mean % across participants (*SD*)			Kolmogorov-S. % Participant significant			Mann-Whitney % Participant significant						Probability of Superiority (b > a) mean % across participants (SD)		
					NF		TF		QF								NF		TF		QF				
		NF	TF	QF	*a>b*	*b>a*	*a>b*	*a>b*	*b>a*	*a>b*	NF	TF	QF	NF	TF	QF	*a>b*	*b>a*	*a>b*	*b>a*	*a>b*	*b>a*	NF	TF	QF
M1	R (a) vs. P (b) Trial Type Effect	100	100	100	0.0	100	0.0	100	0.0	100	87.3 *0.6*	83.1 *0.8*	74.9 *1.3*	31.7	100	50.6	0.0	5.6	91.1	0.0	4.4	17.2	51.3 *0.4*	46.2 *1.1*	50.8 *2.4*
M2	L (a) vs. H (b) Seq.Spec. L.	100	100	100	0.0	100	0.0	100	0.0	100	99.7 *0.2*	99.6 *0.2*	99.6 *0.3*	100	100	100	0.0	100	0.0	100	0.0	100	68.8 *1.3*	67.2 *2.0*	79.9 *3.7*
M3	LR (a) vs. HR (b) Pure Stat. Learn.	100	100	100	0.0	100	0.0	100	0.0	100	99.7 *0.2*	99.6 *0.3*	99.6 *0.3*	100	100	100	0.0	100	0.0	100	0.0	100	94.0 *2.0*	93.5 *2.3*	93.5 *2.8*
	HR (a) vs. HP (b) Higher O.Seq.L.	3.9	3.9	2.8	1.7	1.7	1.7	.17	0.6	2.2	50.1 *1.4*	50.1 *1.4*	50.1 *1.8*	100	100	100	100	0.0	100	0.0	100	0.0	17.2 *1.5*	17.2 *1.5*	28.4 *1.8*
	LR (a) vs. HP (b) Maximized Learning	100	100	100	0.0	100	0.0	100	0.0	100	99.6 *0.2*	99.6 *0.3*	99.6 *0.3*	100	100	100	0.0	100	0.0	100	0.0	100	62.7 *1.0*	60.8 *2.1*	73.3 *4.7*
M4	L (a) vs. H1 (b) Triplet L.(+P. L.)	100	100	100	0.0	100	0.0	100	0.0	100	99.7 *0.2*	99.6 *0.3*	99.6 *0.3*	100	100	100	0.0	100	0.0	100	0.0	100	94.0 *2.0*	93.4 *2.3*	93.4 *2.7*
	H1 (a) vs. H2 (b) Quad L. (+ P. L.)	8.9	8.9	7.2	3.9	5.6	3.9	5.6	1.1	3.3	50.3 *1.4*	50.3 *1.4*	50.4 *2.0*	100	100	100	100	0.0	100	0.0	100	0.0	6.3 *2.1*	6.3 *2.1*	7.1 *3.1*
	L (a) vs. H2 (b) Maximized Learning	100	100	100	0.0	100	0.0	100	0.0	100	99.6 *0.2*	99.6 *0.3*	99.6 *0.3*	57.2	45.0	76.7	0.0	47.2	14.4	13.9	18.3	45.6	52.2 *0.6*	49.9 *2.4*	53.2 *8.4*
M5	LR(a) vs. H1R(b) Triplet Learning	100	100	100	0.0	100	0.0	100	0.0	100	99.7 *0.2*	99.6 *0.3*	99.6 *0.3*	100	100	100	0.0	100	0.0	100	0.0	100	94.0 *2.1*	93.5 *2.3*	93.5 *2.8*
	H1R(a) vs. H1P(b) Pattern Learning	1.1	1.1	3.9	0.6	1.1	0.6	1.1	1.7	2.2	49.9 *1.4*	49.9 *1.4*	49.8 *2.1*	1.1	1.1	1.7	0.6	0.0	0.6	0.0	0.6	0.6	49.8 *1.3*	49.8 *1.3*	49.8 *2.0*
	H1P(a) vs. H2P(b) Quad Learning	4.4	4.4	8.3	1.1	3.9	1.1	3.9	2.8	4.4	50.3 *1.6*	50.3 *1.6*	50.5 *2.3*	100	100	100	100	0.0	100	0.0	100	0.0	6.4 *2.1*	6.4 *2.1*	7.2 *3.2*
	LR(a) vs. H2P(b) Maximized Learning	100	100	100	0.0	100	0.0	100	0.0	100	99.6 *0.2*	99.6 *0.3*	99.6 *0.3*	57.2	45.0	76.7	0.0	47.2	14.4	13.9	18.3	45.6	52.2 *0.6*	49.9 *2.4*	53.2 *8.4*

NF = No Filter, TF = Triplet Filter, QF = Quad Filter

ST-V/3. Conditional Probabilities (Trial Probabilities).		Triplet Level												Quad Level											
		Kolmogorov-S. % Participant significant			Mann-Whitney % Participant significant						Probability of Superiority (b > a) mean % across participants (*SD*)			Kolmogorov-S. % Participant significant			Mann-Whitney % Participant significant						Probability of Superiority (b > a) mean % across participants (SD)		
		NF	TF	QF	NF		TF		QF		NF	TF	QF	NF	TF	QF	NF		TF		QF		NF	TF	QF
					a>b	*b>a*	*a>b*	*a>b*	*b>a*	*a>b*							*a>b*	*b>a*	*a>b*	*b>a*	*a>b*	*b>a*			
M 1	R (a) vs. P (b) Trial Type Effect	100	100	100	0.0	100	0.0	100	0.0	100	87.3 *0.6*	83.1 *0.7*	74.9 *1.3*	100	100	100	0.0	100	0.0	100	0.0	100	94.3 *0.5*	93.1 *0.6*	84.6 *1.2*
M 2	L (a) vs. H (b) Sequence Spec. L.	100	100	100	0.0	100	0.0	100	0.0	100	99.7 *0.2*	99.7 *0.2*	99.7 *0.2*	100	100	100	0.0	100	0.0	100	0.0	100	97.3 *0.8*	97.0 *0.9*	96.2 *1.7*
M 3	LR (a) vs. HR (b) Pure Statistical L.	100	100	100	0.0	100	0.0	100	0.0	100	99.7 *0.2*	99.7 *0.2*	99.7 *0.3*	100	100	100	0.0	100	0.0	100	0.0	100	95.7 *1.8*	95.1 *2.1*	95.2 *2.5*
	HR (a) vs. HP (b) Higher Order S.L.	6.7	6.7	5.6	1.7	1.1	1.7	1.1	3.3	1.7	49.9 *1.5*	49.9 *1.5*	50.1 *2.0*	100	100	100	0.0	100	0.0	100	0.0	100	84.4 *1.0*	84.4 *1.0*	72.6 *1.6*
	LR (a) vs. HP (b) Maximized L.	100	100	100	0.0	100	0.0	100	0.0	100	99.7 *0.2*	99.7 *0.2*	99.7 *0.2*	100	100	100	0.0	100	0.0	100	0.0	100	97.7 *0.6*	97.5 *0.7*	96.6 *1.4*
M 4	L (a) vs. H1 (b) Triplet L. (+ P)	100	100	100	0.0	100	0.0	100	0.0	100	99.8 *0.2*	99.7 *0.2*	99.7 *0.2*	100	100	100	0.0	100	0.0	100	0.0	100	95.7 *1.9*	95.1 *2.1*	95.2 *2.6*
	H1 (a) vs. H2 (b) Quad L. (+ P. L.)	3.3	3.3	2.8	2.8	1.1	2.8	1.1	0.6	4.4	50.0 *1.2*	50.0 *1.2*	50.6 *1.9*	100	100	100	0.0	100	0.0	100	0.0	100	96.0 *0.9*	96.0 *0.9*	96.2 *1.6*
	L (a) vs. H2 (b) Maximized L.	100	100	100	0.0	100	0.0	100	0.0	100	99.7 *0.2*	99.7 *0.2*	99.7 *0.3*	100	100	100	0.0	100	0.0	100	0.0	100	98.3 *0.4*	98.3 *0.4*	98.2 *0.7*
M 5	LR (a) vs. H1R (b) Triplet Learning	100	100	100	0.0	100	0.0	100	0.0	100	99.8 *0.2*	99.7 *0.2*	99.7 *0.3*	100	100	100	0.0	100	0.0	100	0.0	100	95.7 *1.8*	95.1 *2.1*	95.2 *2.5*
	H1R (a) vs. H1P (b) Pattern Learning	7.2	7.2	6.1	0.0	2.2	0.0	2.2	3.9	1.7	49.9 *1.9*	49.9 *1.9*	49.7 *2.3*	0.6	0.6	1.1	0.6	1.7	0.6	1.7	1.7	0.6	49.9 *1.7*	49.9 *1.7*	49.8 *2.0*
	H1P (a) vs. H2P (b) Quad Learning	2.8	2.8	4.4	0.6	2.8	0.6	2.8	1.1	6.1	50.1 *1.5*	50.1 *1.5*	50.7 *2.3*	100	100	100	0.0	100	0.0	100	0.0	100	96.1 *0.8*	96.1 *0.8*	95.2 *1.6*
	LR(a) vs. H2P(b) Max.Learning	100	100	100	0.0	100	0.0	100	0.0	100	99.7 *0.2*	99.7 *0.2*	99.7 *0.3*	100	100	100	0.0	100	0.0	100	0.0	100	98.3 *0.4*	94.3 *0.4*	98.2 *0.6*

NF = No Filter, TF = Triplet Filter, QF = Quad Filter

ST-V/4. Abstract Structure of the Combination.		Triplet Level						Quad Level					
		χ^2 % participant significant			Cramer's V mean (*SD*)			χ^2 % participant significant			Cramer's V mean (SD)		
		NF	TF	QF	NF	TF	QF	NF	TF	QF	NF	TF	QF
M1	R (a) vs. P (b) Trial Type Effect	100.00	10.56	1.67	0.38 *0.01*	0.03 *0.01*	0.02 *0.01*	100.00	100.00	7.22	0.50 *0.01*	0.35 *0.01*	0.04 *0.02*
M2	L (a) vs. H (b) Sequence Spec. L.	100.00	3.89	4.44	0.49 *0.01*	0.02 *0.01*	0.02 *0.02*	100.00	100.00	1.11	0.61 *0.01*	0.41 *0.01*	0.03 *0.02*
M3	LR (a) vs. HP (b) Pure Statistical Learn.	100.00	6.11	5.56	0.34 *0.01*	0.03 *0.02*	0.03 *0.03*	100.00	100.00	8.89	0.71 *0.01*	0.66 *0.01*	0.05 *0.03*
	HR (a) vs. HP (b) Higher Order Seq.L.	11.11	11.11	6.11	0.03 *0.02*	0.03 *0.02*	0.03 *0.02*	100.00	100.00	14.44	0.47 *0.01*	0.47 *0.01*	0.05 *0.03*
	LR (a) vs. HP (b) Maximized Learning	100.00	5.56	2.78	0.47 *0.01*	0.03 *0.01*	0.02 *0.02*	100.00	100.00	0.00	0.59 *0.01*	0.40 *0.01*	0.03 0.02
M4	L (a) vs. H1 (b) Triplet L. (+ P. L.)	100.00	7.78	6.11	0.41 *0.01*	0.03 *0.02*	0.03 *0.02*	100.00	100.00	2.78	0.76 *0.01*	0.71 *0.01*	2.78 *0.02*
	H1 (a) vs. H2 (b) Quad L. (+ P. L.)	5.00	5.00	5.56	0.02 *0.02*	0.02 *0.02*	0.03 *0.02*	100.00	100.00	3.33	0.76 *0.01*	0.76 *0.01*	0.04 *0.02*
	L (a) vs. H2 (b) Maximized Learning	100.00	2.22	2.22	0.45 *0.01*	0.02 *0.01*	0.03 *0.02*	100.00	100.00	0.56	0.58 *0.01*	0.41 *0.01*	0.04 *0.02*
M5	LR (a) vs. H1R (b) Triplet Learning	100.00	6.11	5.56	0.34 *0.01*	0.03 *0.02*	0.03 *0.03*	100.00	100.00	8.89	0.71 *0.01*	0.66 *0.01*	0.05 *0.03*
	H1R (a) vs. H1P (b) Pattern Learning	15.00	15.00	5.00	0.05 *0.03*	0.05 *0.03*	5.00 *0.02*	26.11	26.11	19.44	0.07 *0.04*	0.07 *0.04*	0.07 *0.04*
	H1P (a) vs. H2P (b) Quad Learning	5.56	5.56	4.44	0.03 *0.02*	0.03 *0.02*	0.03 *0.02*	100.00	100.00	6.67	0.71 *0.01*	0.71 *0.01*	0.05 *0.03*
	LR (a) vs. H2P (b) Maximized Learning	100.00	2.22	2.22	0.45 *0.01*	0.02 *0.01*	0.03 *0.02*	100.00	100.00	0.56	0.58 *0.01*	0.41 *0.01*	0.04 0.02

NF = No Filter, TF = Triplet Filter, QF = Quad Filter

Description of Supplementary Tables ST-V/5-15

ST-V/5 Mean reaction times were calculated for each Models' each subcategory in each epoch for each participant. Then the nine values belonging to the nine epochs were averaged yielding a single value per participant (RT mean). The Standard Deviation (RT SD) and Coefficient of Variation (RT CV, SD/mean, %) was also computed based on the spread of these means. Similarly, error percentages were calculated for each Models' each subcategory in each epoch for each participant. Then the nine values belonging to the nine epochs were averaged yielding a single value per participant (Error %). The Standard Deviation (Error SD) and Coefficient of Variation (Error CV) were then computed based on the spread of these averages. All of these calculations were done for all three filtering types (NF = No Filter, TF = Triplet Filter, QF = Quad Filter).

ST-V/6 Learning scores were calculated for each Model. Learning scores based on reaction times are Cohen's d values, computed separately for each individual and each epoch, and the nine values corresponding to the nine epochs were averaged to yield a single value for each participant (d mean). The spread of these means were quantified as Standard Deviations (d SD) and Coefficients of Variation (d CV).

In the case of error rates, data from the nine epochs was collapsed into a single category due to the low overall error rates. Cramer's V values were computed for each Models' each learning score individually, and the mean of these means is presented in the table (V mean). The spread of these means were also assessed (V SD and V CV).

All of these calculations were done for all three filtering types (NF = No Filter, TF = Triplet Filter, QF = Quad Filter).

ST-V/7-12 In these tables the same descriptive statistics are shown as in **Table ST-V/6** (each new table corresponding to the subsequent row in Table ST-V/6), but here they are broken down by the ASRT sequence being used (the notations P1-P6 referring to the six pattern types).

ST-V/13 Within-subject variability of reaction times were quantified as Standard Deviations (SD mean (ms)) and as Coefficients of Variation (CV mean (%)). These descriptive statistics were computed for each Models' each subcategory in each epoch for each participant. Then the nine values belonging to the nine epochs were averaged yielding a single value per participant. All of these calculations were done for all three filtering types (NF = No Filter, TF = Triplet Filter, QF = Quad Filter).

ST-V/14-15 In these tables the same descriptive statistics are shown as in **Table ST-V/13** (separate tables corresponding to SDs and CVs), but here the statistics are broken down by the ASRT sequence being used (the notations P1-P6 referring to the six pattern types).

ST-V/5. Descriptive statistics of mean reaction times and error percentages with standard deviations and coefficients of variations for each Models' each subcategory.

		Model 1		Model 2		Model 3			Model 4			Model 5			
		P	R	H	L	HP	HR	LR	H2	H1	L	H2P	H1P	H1R	LR
RT mean	NF	373.59	381.18	372.07	386.51	373.59	365.92	386.51	375.72	366.77	386.51	375.72	367.57	365.92	386.51
	TF	373.59	379.14	372.07	385.99	373.59	365.92	385.99	375.72	366.77	385.99	375.72	367.57	365.92	385.99
	QF	377.32	385.83	377.43	394.28	377.32	377.57	394.28	375.74	378.29	394.28	375.74	378.90	377.57	394.28
RT SD	NF	27.80	28.39	27.73	29.00	27.80	27.81	29.00	27.99	27.64	29.00	27.99	27.66	27.81	29.00
	TF	27.80	28.24	27.73	28.93	27.80	27.81	28.93	27.99	27.64	28.93	27.99	27.66	27.81	28.93
	QF	29.96	31.52	30.30	32.58	29.96	31.36	32.58	29.79	30.83	32.58	29.79	30.56	31.36	32.58
RT CV	NF	7.44	7.45	7.45	7.50	7.44	7.60	7.50	7.45	7.54	7.50	7.45	7.53	7.60	7.50
	TF	7.44	7.45	7.45	7.49	7.44	7.60	7.50	7.45	7.54	7.49	7.45	7.53	7.60	7.49
	QF	7.94	8.17	8.03	8.26	7.94	8.31	8.26	7.93	8.15	8.26	7.93	8.07	8.31	8.26
Error (%)	NF	3.95	6.15	3.89	6.99	3.95	3.65	6.99	4.05	3.70	6.99	4.05	3.69	3.65	6.99
	TF	3.95	5.60	3.89	6.59	3.95	3.65	6.59	4.05	3.70	6.59	4.05	3.69	3.65	6.59
	QF	4.45	6.17	4.57	7.56	4.45	4.77	7.56	4.15	4.78	7.56	4.15	4.75	4.77	7.56
Error SD	NF	2.68	3.50	2.60	3.97	2.68	2.51	3.97	2.80	2.48	3.97	2.80	2.63	2.51	3.97
	TF	2.68	3.29	2.60	3.89	2.68	2.51	3.89	2.80	2.48	3.89	2.80	2.63	2.51	3.89
	QF	3.01	3.73	3.04	4.70	3.01	3.35	4.70	2.94	3.26	4.70	2.94	3.42	3.35	4.70
Error CV	NF	67.89	56.91	66.68	56.81	67.89	68.72	56.81	69.05	67.08	56.81	69.05	71.43	68.72	56.81
	TF	67.89	58.71	66.68	58.96	67.89	68.72	58.96	69.05	67.08	58.96	69.05	71.43	68.72	58.96
	QF	67.73	60.49	66.51	62.12	67.73	70.28	62.12	70.76	68.20	62.12	70.76	71.96	70.28	62.12

RT = reaction times, SD = standard deviation, CV = Coefficient of Variation (SD/mean, %), NF = No Filter, TF = Triplet Filter, QF = Quad Filter

ST-V/6 Mean individual effect sizes of learning, SD of these effect sizes and the CV of these effect sizes.

		Model 1	Model 2	Model 3			Model 4			Model 5			
		R-P	L-H	HR-HP	LR-HR	LR-HP	H1-H2	L-H1	L-H2	H1P-H2P	H1R-H1P	LR-H1R	LR-H2P
d mean	NF	.153	.283	-.153	.407	.256	-.174	.385	.217	-.164	-.032	.407	.217
	TF	.113	.272	-.153	.396	.244	-.174	.375	.206	-.164	-.032	.396	.206
	QF	.172	.340	-.004	.350	.343	.047	.328	.376	.056	-.028	.350	.376
d SD	NF	.106	.136	.093	.156	.137	.100	.150	.141	.112	.089	.156	.141
	TF	.082	.115	.093	.147	.114	.100	.138	.117	.112	.089	.147	.117
	QF	.100	.171	.104	.192	.170	.121	.180	.182	.134	.116	.192	.182
d CV	NF	69.36	47.95	60.41	38.41	53.51	57.44	38.95	64.97	68.57	276.19	38.41	64.97
	TF	72.62	42.31	60.41	37.09	46.73	57.44	36.73	56.91	68.57	276.19	37.09	56.91
	QF	57.90	50.15	2543.49	54.87	49.65	265.81	54.86	48.41	241.79	409.74	54.87	48.41
V mean	NF	.050	.068	-.005	.059	.067	-.009	.069	.064	-.008	.000	.059	.064
	TF	.039	.058	-.005	.058	.058	-.009	.065	.056	-.008	.000	.058	.056
	QF	.037	.058	.006	.057	.063	.012	.057	.069	.012	.001	.057	.069
V SD	NF	.031	.036	.026	.031	.037	.028	.036	.036	.029	.038	.031	.036
	TF	.029	.034	.026	.037	.035	.028	.038	.036	.029	.038	.037	.036
	QF	.037	.049	.036	.055	.053	.037	.051	.058	.047	.045	.055	.058
V CV	NF	61.55	52.37	-468.25	53.58	54.69	-306.23	51.74	56.95	-358.67	-30905.4	53.58	56.95
	TF	73.28	58.70	-468.25	62.50	60.94	-306.23	58.36	64.32	-358.67	-30905.4	62.50	64.32
	QF	98.50	84.58	581.77	97.46	83.45	311.89	90.50	83.15	382.23	5738.97	97.46	83.15

RT = reaction times, SD = standard deviation, CV = Coefficient of Variation (SD/mean, %), NF = No Filter, TF = Triplet Filter, QF = Quad Filter

ST-V/7. Mean individual effect sizes (Cohen's d) computed for each Models' learning scores broken down by the ASRT sequences being taught (P1-P6), RTs.

		Model 1	Model 2	Model 3			Model 4			Model 5			
		R-P	L-H	HR-HP	LR-HR	LR-HP	H1-H2	L-H1	L-H2	H1P-H2P	H1R-H1P	LR-H1R	LR-H2P
P1	NF	.153	.297	-.178	.438	.265	-.190	.407	.226	-.168	-.055	.438	.226
	TF	.110	.285	-.178	.427	.253	-.190	.396	.213	-.168	-.055	.427	.213
	QF	.133	.320	-.049	.357	.309	-.110	.326	.316	.013	-.055	.357	.316
P2	NF	.223	.375	-.146	.493	.348	-.159	.466	.315	-.143	-.037	.493	.315
	TF	.164	.346	-.146	.465	.319	-.159	.438	.286	-.143	-.037	.465	.286
	QF	.234	.467	.006	.473	.468	.106	.434	.538	.130	-.057	.473	.538
P3	NF	.123	.243	-.152	.362	.216	-.212	.366	.163	-.226	.011	.362	.163
	TF	.093	.242	-.152	.362	.215	-.212	.365	.161	-.226	.011	.362	.161
	QF	.189	.318	.038	.294	.337	.018	.313	.342	-.011	.048	.294	.342
P4	NF	.141	.257	-.139	.372	.233	-.136	.339	.206	-.114	-.053	.372	.206
	TF	.108	.251	-.139	.365	.227	-.136	.333	.200	-.114	-.053	.365	.200
	QF	.173	.330	.002	.340	.333	.102	.302	.396	.124	-.058	.340	.396
P5	NF	.124	.238	-.144	.354	.212	-.186	.348	.166	-.189	-.006	.354	.166
	TF	.079	.214	-.144	.330	.188	-.186	.324	.142	-.189	-.006	.330	.142
	QF	.156	.277	.015	.271	.287	.040	.267	.311	.034	.004	.271	.311
P6	NF	.158	.294	-.162	.428	.266	-.164	.393	.234	-.144	–.052	.428	.234
	TF	.128	.300	-.162	.435	.271	-.164	.399	.240	-.144	-.052	.435	.240
	QF	.150	.334	-.039	.371	.326	.024	.333	.355	.041	-.055	.371	.355

P1-P6 = Pattern1-Pattern6, NF = No Filter, TF = Triplet Filter, QF = Quad Filter

ST-V/8. Individual effect size (Cohen's d) SDs computed for each Models' learning scores broken down by the ASRT sequences being taught (P1-P6), RTs.

		Model 1	Model 2	Model 3			Model 4			Model 5			
		R-P	L-H	HR-HP	LR-HR	LR-HP	H1-H2	L-H1	L-H2	H1P-H2P	H1R-H1P	LR-H1R	LR-H2P
P1	NF	.109	.134	.092	.138	.138	.104	.137	.145	.111	.074	.138	.145
	TF	.082	.107	.092	.132	.108	.104	.131	.111	.111	.074	.132	.111
	QF	.105	.173	.099	.180	.180	.136	.174	.211	.144	.091	.180	.211
P2	NF	.093	.128	.064	.145	.126	.077	.150	.120	.094	.071	.145	.120
	TF	.069	.111	.064	.138	.108	.077	.134	.106	.094	.071	.138	.106
	QF	.092	.151	.099	.173	.152	.113	.169	.143	.129	.118	.173	.143
P3	NF	.090	.134	.103	.186	.128	.107	.174	.124	.115	.099	.186	.124
	TF	.066	.110	.103	.167	.103	.107	.153	.097	.115	.099	.167	.097
	QF	.090	.160	.116	.201	.150	.116	.183	.145	.127	.132	.201	.145
P4	NF	.104	.121	.084	.132	.126	.084	.123	.133	.099	.093	.132	.133
	TF	.082	.103	.084	.126	.105	.084	.120	.107	.099	.093	.126	.107
	QF	.100	.152	.089	.161	.158	.094	.156	.161	.115	.115	.161	.161
P5	NF	.099	.107	.115	.128	.114	.121	.119	.128	.122	.084	.128	.128
	TF	.081	.087	.115	.126	.092	.121	.114	.102	.122	.084	.126	.102
	QF	.070	.144	.101	.185	.130	.118	.166	.129	.112	.094	.185	.129
P6	NF	.116	.151	.091	.165	.153	.090	.167	.151	.102	.091	.165	.151
	TF	.090	.130	.091	.155	.130	.090	.149	.131	.102	.091	.155	.131
	QF	.112	.194	.098	.201	.199	.105	.205	.199	.116	.107	.201	.199

P1-P6 = Pattern1-Pattern6, NF = No Filter, TF = Triplet Filter, QF = Quad Filter

ST-V/9. Individual effect size (Cohen's d) CVs computed for each Models' learning scores broken down by the ASRT sequences being taught (P1-P6), RTs.

		Model 1	Model 2	Model 3			Model 4			Model 5			
		R-P	L-H	HR-HP	LR-HR	LR-HP	H1-H2	L-H1	L-H2	H1P-H2P	H1R-H1P	LR-H1R	LR-H2P
P1	NF	71.07	45.00	51.65	31.59	51.89	55.05	33.77	63.97	66.53	134.01	31.59	63.97
	TF	74.12	37.65	51.65	30.88	42.83	55.05	32.98	51.92	66.53	134.01	30.88	51.92
	QF	78.93	54.17	201.13	50.46	58.21	1370.95	53.25	66.65	1106.34	166.04	50.46	66.65
P2	NF	41.87	34.07	44.25	29.05	36.21	48.09	32.22	38.22	65.38	193.63	29.50	38.22
	TF	41.79	32.19	44.25	29.61	33.78	48.09	30.52	36.90	65.38	193.63	29.61	36.90
	QF	39.25	32.39	1532.23	36.58	32.55	106.32	38.92	26.53	98.98	206.21	36.58	26.53
P3	NF	73.29	55.24	67.94	51.37	59.18	50.36	47.54	76.00	51.03	895.22	51.37	76.00
	TF	71.05	45.55	67.94	46.15	47.88	50.36	42.07	59.99	51.03	895.22	46.15	59.99
	QF	47.36	50.49	307.93	68.34	44.53	684.08	58.46	42.24	1163.74	273.37	68.34	42.24
P4	NF	74.18	47.15	60.57	35.66	54.12	62.01	36.28	64.61	86.36	173.94	35.66	64.61
	TF	76.66	41.14	60.57	34.59	46.52	62.01	36.03	53.25	86.36	173.94	34.59	53.25
	QF	57.92	45.94	4440.50	47.40	47.49	91.45	51.65	40.74	92.82	198.34	47.40	40.74
P5	NF	80-39	45.17	80.20	36.26	53.86	65.05	34.17	77.18	64.87	1413.96	36.26	77.18
	TF	102.48	40.80	80.20	38.07	48.82	65.05	35.03	72.30	64.87	1413.96	38.07	72.30
	QF	45.19	51.89	655.02	68.29	45.29	293.38	62.21	41.63	328.69	2205.81	68.29	41.63
P6	NF	72.93	51.27	55.98	38.51	57.40	54.91	42.62	64.51	70.84	174.94	38.51	64.51
	TF	70.70	43.23	55.98	35.58	48.00	54.91	37.31	54.51	70.84	174.94	35.58	54.51
	QF	74.74	57.98	248.52	54.35	60.95	429.84	61.61	56.11	282.94	196.45	54.35	56.11

P1-P6 = Pattern1-Pattern6, NF = No Filter, TF = Triplet Filter, QF = Quad Filter

ST-V/10. Mean individual effect sizes (Cramer's V) computed for each Models' learning scores broken down by the ASRT sequences being taught (P1-P6), errors.

		Model 1	Model 2	Model 3			Model 4			Model 5			
		R-P	L-H	HR-HP	LR-HR	LR-HP	H1-H2	L-H1	L-H2	H1P-H2P	H1R-H1P	LR-H1R	LR-H2P
P1	NF	.047	.066	-.009	.060	.064	-.020	.074	.057	-.020	.006	.060	.057
	TF	.035	.054	-.009	.059	.054	-.020	.069	.049	-.020	.006	.059	.049
	QF	.038	.058	.007	.057	.064	.002	.061	.063	-.002	.009	.057	.063
P2	NF	.054	.073	-.006	.062	.072	-.008	.073	.069	-.006	-.004	.062	.069
	TF	.043	.063	-.006	.063	.063	-.008	.070	.062	-.006	-.004	.063	.062
	QF	.045	.073	.003	.075	.078	.014	.072	.087	.017	-.005	.075	.087
P3	NF	.048	.063	-.002	.052	.063	-.004	.062	.061	-.005	.003	.052	.061
	TF	.040	.056	-.002	.053	.057	-.004	.060	.057	-.005	.003	.053	.057
	QF	.029	.039	.010	.034	.044	.017	.036	.054	.017	.004	.034	.054
P4	NF	.058	.078	-.005	.065	.077	.000	.073	.076	.004	-.011	.065	.076
	TF	.047	.068	-.005	.066	.069	.000	.069	.071	.004	-.011	.066	.071
	QF	.043	.072	.001	.074	.076	.018	.069	.087	.025	-.011	.074	.087
P5	NF	.048	.065	-.004	.056	.064	-.010	.067	.060	-.011	.004	.056	.060
	TF	.039	.057	-.004	.058	.057	-.010	.065	.055	-.011	.004	.058	.055
	QF	.041	.058	.013	.054	.066	.018	.055	.075	.016	.006	.054	.075
P6	NF	.044	.062	-.007	.055	.060	-.013	.066	.056	-.012	.002	.055	.056
	TF	.029	.045	-.007	.049	.048	-.013	.055	-.043	-.012	.002	.049	.043
	QF	.026	.043	.003	.045	.047	.001	.045	.047	-.001	.003	.045	.047

P1-P6 = Pattern1-Pattern6, NF = No Filter, TF = Triplet Filter, QF = Quad Filter

ST-V/11. Individual effect size (Cramer's V) SDs computed for each Models' learning scores broken down by the ASRT sequences being taught (P1-P6), errors.

		Model 1	Model 2	Model 3			Model 4			Model 5			
		R-P	L-H	HR-HP	LR-HR	LR-HP	H1-H2	L-H1	L-H2	H1P-H2P	H1R-H1P	LR-H1R	LR-H2P
P1	NF	.034	.036	.026	.028	.039	.029	.033	.039	.027	.035	.028	.039
	TF	.030	.030	.026	.029	.033	.029	.032	.034	.027	.035	.029	.034
	QF	.043	.055	.038	.059	.061	.038	.057	.065	.045	.044	.059	.065
P2	NF	.025	.033	.027	.035	.031	.027	.037	.029	.028	.041	.035	.029
	TF	.023	.034	.027	.044	.033	.027	.042	.032	.028	.041	.044	.032
	QF	.033	.050	.038	.060	.051	.035	.054	.053	.043	.044	.060	.053
P3	NF	.036	.044	.025	.037	.045	.031	.043	.045	.033	.037	.037	.045
	TF	.032	.040	.025	.042	.042	.031	.045	.044	.033	.037	.042	.044
	QF	.026	.039	.030	.050	.040	.044	.045	.049	.058	.042	.050	.049
P4	NF	.026	.030	.025	.028	.030	.021	.027	.030	.021	.036	.028	.030
	TF	.026	.032	.025	.035	.032	.021	.033	.034	.021	.036	.035	.034
	QF	.039	.052	.037	.058	.056	.031	.052	.062	.037	.044	.058	.062
P5	NF	.029	.033	.029	.033	.034	.031	.038	.032	.030	.038	.033	.032
	TF	.028	.030	.029	.037	.032	.031	.039	.032	.030	.038	.037	.032
	QF	.032	.041	.032	.047	.044	.036	.046	.046	.051	.044	.047	.046
P6	NF	.033	.037	.023	.028	.039	.025	.036	.038	.031	.044	.028	.038
	TF	.031	.034	.023	.031	.037	.025	.037	.037	.031	.044	.031	.037
	QF	.044	.045	.041	.046	.055	.032	.046	.058	.041	.051	.046	.058

SD = Standard Deviation, P1-P6 = Pattern1-Pattern6, NF = No Filter, TF = Triplet Filter, QF = Quad Filter

ST-V/12. Individual effect size (Cramer's V) CVs computed for each Models' learning scores broken down by the ASRT sequences being taught (P1-P6), errors.

		Model 1	Model 2	Model 3			Model 4			Model 5			
		R-P	L-H	HR-HP	LR-HR	LR-HP	H1-H2	L-H1	L-H2	H1P-H2P	H1R-H1P	LR-H1R	LR-H2P
P1	NF	72.80	55.00	-291.30	45.84	60.81	-142.75	44.91	68.33	-130.71	621.41	45.84	68.33
	TF	85.58	55.59	-291.30	48.85	62.06	-142.75	46.51	70.75	-130.71	621.41	58.85	70.75
	QF	112.75	94.26	540.84	103.45	95.07	2349.56	93.89	103.91	-2026.90	507.06	103.45	103.91
P2	NF	46.27	44.85	-411.41	55.70	43.57	-325.35	51.15	41.96	-462.69	-1117.24	55.70	41.96
	TF	53.95	54.31	-411.41	69.01	51.67	-325.35	60.96	51.10	-462.96	-1117.24	69.01	51.10
	QF	73.17	68.05	1169.08	80.61	64.52	259.27	75.02	61.24	249.07	-851.74	80.61	61.24
P3	NF	75.92	69.30	-1592.84	70.86	71.10	-702.42	69.71	73.75	-714.74	1213.91	70.86	73.75
	TF	81.10	71.74	-1592.84	78.60	73.44	-702.42	73.85	77.24	-714.74	1213.91	70.86	77.24
	QF	90.11	101.76	300.39	146.74	90.96	262.01	125.25	91.46	347.87	1149.73	146.74	91.46
P4	NF	44.24	38.11	-518.79	42.78	39.23	14006.08	37.42	39.81	487.49	-343.98	42.78	39.81
	TF	55.09	46.98	-518.79	53.01	47.43	14006.08	46.91	48.17	487.49	-343.98	53.01	48.17
	QF	90.13	72.17	2799.04	78.49	72.96	168.33	74.73	71.12	148.94	-411.30	78.49	71.12
P5	NF	60.97	51.03	-691.72	59.53	53.12	-306.07	56.33	53.55	-280.88	1035.03	59.53	53.55
	TF	71.44	53.52	-691.72	63.42	56.06	-306.07	59.40	57.32	-280.88	1035.03	63.42	57.32
	QF	77.89	71.17	255.72	88.57	67.39	198.97	83.27	60.80	313.34	738.28	88.57	60.80
P6	NF	74.96	59.57	-338.12	50.03	64.37	-195.99	55.23	66.92	-256.93	2125.72	50.03	66.92
	TF	106.58	73.85	-338.12	63.48	81.32	-195.99	68.05	86.32	-256.93	2125.72	63.48	86.32
	QF	167.83	103.88	1429.56	102.13	116.29	3529.74	102.11	125.08	-3486.94	1699.65	102.13	125.08

CV = coefficient of variation (SD/mean, %), P1-P6 = Pattern1-Pattern6, NF = No Filter, TF = Triplet Filter, QF = Quad Filter

ST-V/13. Mean within-subject variability of reaction times computed for each Models` each subcategory

		Model 1		Model 2		Model 3			Model 4			Model 5			
		P	R	H	L	HP	HR	LR	H2	H1	L	H2P	H1P	H1R	LR
SD mean (ms)	NF	51.38	54.72	51.64	54.34	51.38	51.09	54.34	50.77	51.76	54.34	50.77	51.39	51.09	54.34
	TF	51.38	54.40	51.64	54.03	51.38	51.09	54.03	50.77	51.76	54.03	50.77	51.39	51.09	54.03
	QF	49.59	52.14	49.61	52.73	49.59	48.53	52.73	48.98	49.40	52.73	48.98	49.06	48.53	52.73
CV mean (%)	NF	13.67	14.28	13.80	13.97	13.67	13.88	13.97	13.42	14.03	13.97	13.42	13.89	13.88	13.97
	TF	13.67	14.28	13.80	13.91	13.67	13.88	13.91	13.42	14.03	13.91	13.42	13.89	13.88	13.91
	QF	13.04	13.42	13.05	13.28	13.04	12.74	13.28	12.95	12.95	13.28	12.95	12.82	12.74	13.28

SD = Standard Deviaton, CV = Coefficient of Variation (SD/mean, %)

ST-V/14. Mean within-subject variability (SD) of reaction times computed for each Models' each subcategory broken down by the ASRT sequences (P1-P6).

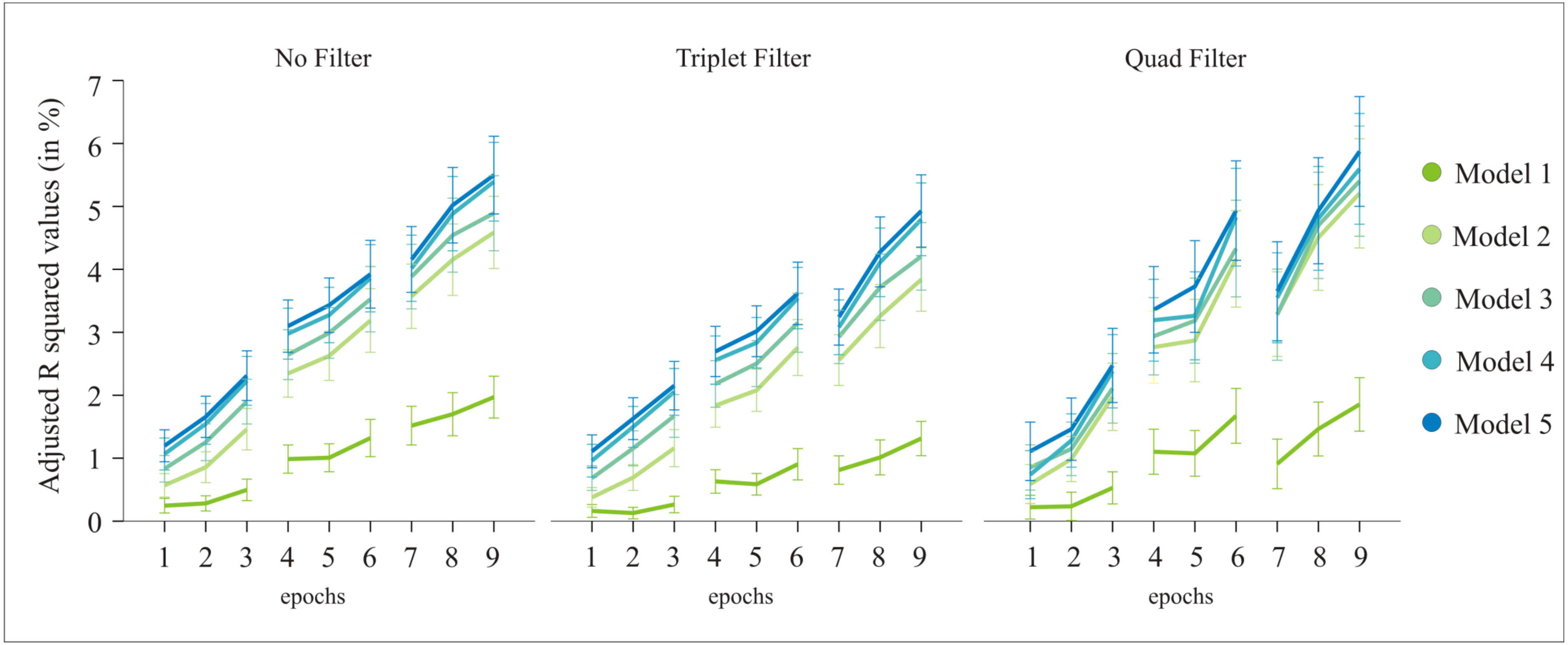
		Model 1		Model 2		Model 3			Model 4			Model5			
		P	R	H	L	HP	HR	LR	H2	H1	L	H2P	H1P	H1R	LR
P1	NF	50.01	53.48	50.23	53.03	50.01	49.59	53.03	49.43	50.17	53.03	49.43	49.87	49.59	53.03
	TF	50.01	53.00	50.23	52.54	50.01	49.59	52.54	49.43	50.17	52.54	49.43	49.87	49.59	52.54
	QF	48.18	50.65	48.19	51.08	48.18	47.11	51.08	48.03	47.87	51.08	48.03	47.56	47.11	51.08
P2	NF	51.13	55.59	51.14	55.41	51.13	50.00	55.41	50.46	51.19	55.41	50.46	51.62	50.00	55.41
	TF	51.13	54.82	51.14	54.76	51.13	50.00	54.76	50.46	51.19	54.76	50.46	51.62	50.00	54.76
	QF	50.11	52.89	49.64	53.65	50.11	48.04	53.65	48.72	49.51	53.65	48.72	50.12	48.04	53.65
P3	NF	54.12	57.33	54.90	56.18	54.12	56.34	56.18	53.44	55.64	56.18	53.44	53.89	56.34	56.18
	TF	54.12	57.26	54.90	55.80	54.12	56.34	55.80	53.44	55.64	55.80	53.44	53.89	56.34	55.80
	QF	51.68	55.20	54.72	54.11	51.68	53.48	54.11	50.58	53.19	54.11	50.58	51.60	53.48	54.11
P4	NF	50.56	53.78	50.61	53.95	50.56	49.13	53.95	50.13	50.44	53.95	50.13	50.64	49.13	53.95
	TF	50.56	53.98	50.61	54.59	50.56	49.13	54.59	50.13	50.44	54.59	50.13	50.64	49.13	54.49
	QF	48.35	50.85	47.98	53.03	48.35	45.97	23.03	47.93	47.54	53.03	47.93	47.83	45.97	53.03
P5	NF	51.05	53.52	51.41	53.08	51.05	51.16	53.08	50.49	51.41	53.08	50.49	50.61	51.16	53.08
	TF	51.05	53.13	51.41	52.60	51.05	51.16	52.60	50.49	51.41	52.60	50.49	50.61	51.16	52.60
	QF	49.17	51.30	49.39	51.65	49.17	48.90	51.65	48.65	49.23	51.65	48.65	48.40	48.90	51.65
P6	NF	51.49	54.71	51.62	54.45	51.49	50.29	54.45	50.72	51.77	54.45	50.72	51.83	50.29	54.45
	TF	51.49	54.22	51.62	53.86	51.49	50.29	53.86	50.72	51.77	53.86	50.72	51.83	50.29	53.86
	QF	50.28	52.05	49.94	52.93	50.28	47.74	52.93	50.24	49.18	52.93	50.24	48.97	47.74	52.93

P1-P6 = Pattern1-Pattern6, NF = No Filter, TF = Triplet Filter, QF = Quad Filter

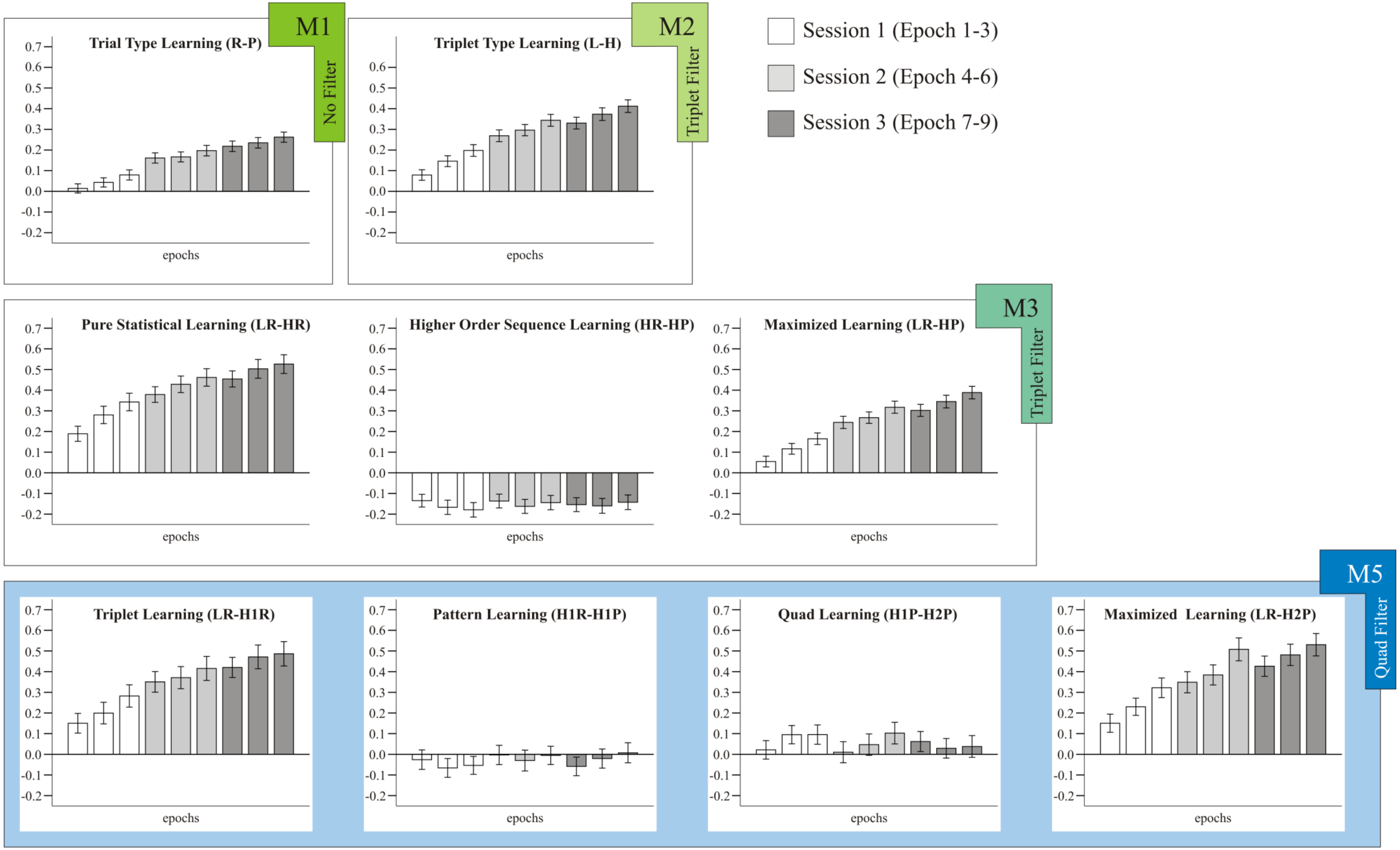
ST-V/15. Mean within-subject variability (CV) of reaction times computed for each Models' each subcategory broken down by the ASRT sequences (P1-P6).

		Model 1		Model 2		Model 3			Model 4			Model5			
		P	R	H	L	HP	HR	LR	H2	H1	L	H2P	H1P	H1R	LR
P1	NF	13.75	14.43	13.88	14.08	13.75	13.96	14.08	13.51	14.07	14.08	13.51	13.94	13.96	14.08
	TF	13.75	14.40	13.88	14.00	13.75	13.96	14.00	13.51	14.07	14.00	13.51	13.94	13.96	14.00
	QF	13.06	13.52	13.08	13.33	13.06	12.81	13.33	13.04	12.98	13.33	13.04	12.85	12.81	13.33
P2	NF	13.63	14.38	13.69	14.07	13.63	13.60	14.07	13.38	13.88	14.07	13.38	13.93	13.60	14.07
	TF	13.63	14.31	13.69	13.97	13.63	13.60	13.97	13.38	13.88	13.97	13.38	13.93	13.60	13.97
	QF	13.29	13.61	13.17	13.34	13.29	12.73	13.34	13.04	13.06	13.34	13.04	13.15	12.73	13.34
P3	NF	14.06	14.67	14.32	14.20	14.06	14.92	14.20	13.76	14.77	14.20	13.76	14.33	14.92	14.20
	TF	14.06	14.71	14.32	14.09	14.06	14.92	14.09	13.76	14.77	14.09	13.76	14.33	14.92	14.09
	QF	13.32	13.87	13.55	13.37	13.32	13.63	13.37	13.03	13.63	13.37	13.03	13.28	13.63	13.37
P4	NF	13.17	13.73	13.24	13.58	13.17	13.06	13.58	13.01	13.33	13.58	13.01	13.30	13.06	13.58
	TF	13.17	13.85	13.24	13.75	13.17	13.06	13.75	13.01	13.33	13.75	13.01	13.30	13.06	13.75
	QF	12.40	12.79	12.32	13.05	12.40	11.80	13.05	12.42	12.14	13.05	12.42	12.13	11.80	13.05
P5	NF	13.54	14.01	13.70	13.71	13.54	13.89	13.71	13.28	13.95	13.71	13.28	13.71	13.89	13.71
	TF	13.54	13.99	13.70	13.62	13.54	13.89	13.62	13.28	13.95	13.62	13.28	13.71	13.89	13.62
	QF	12.93	13.24	12.97	13.13	12.93	12.78	13.13	12.85	12.89	13.13	12.85	12.67	12.78	13.13
P6	NF	13.92	14.50	14.02	14.22	13.92	13.89	14.22	13.64	14.25	14.22	13.64	14.19	13.89	14.22
	TF	13.92	14.43	14.02	14.05	13.92	13.89	14.05	13.64	14.25	14.05	13.64	14.19	13.89	14.05
	QF	13.36	13.57	13.29	13.51	13.36	12.72	13.51	13.39	13.06	13.51	13.39	12.94	12.72	13.51

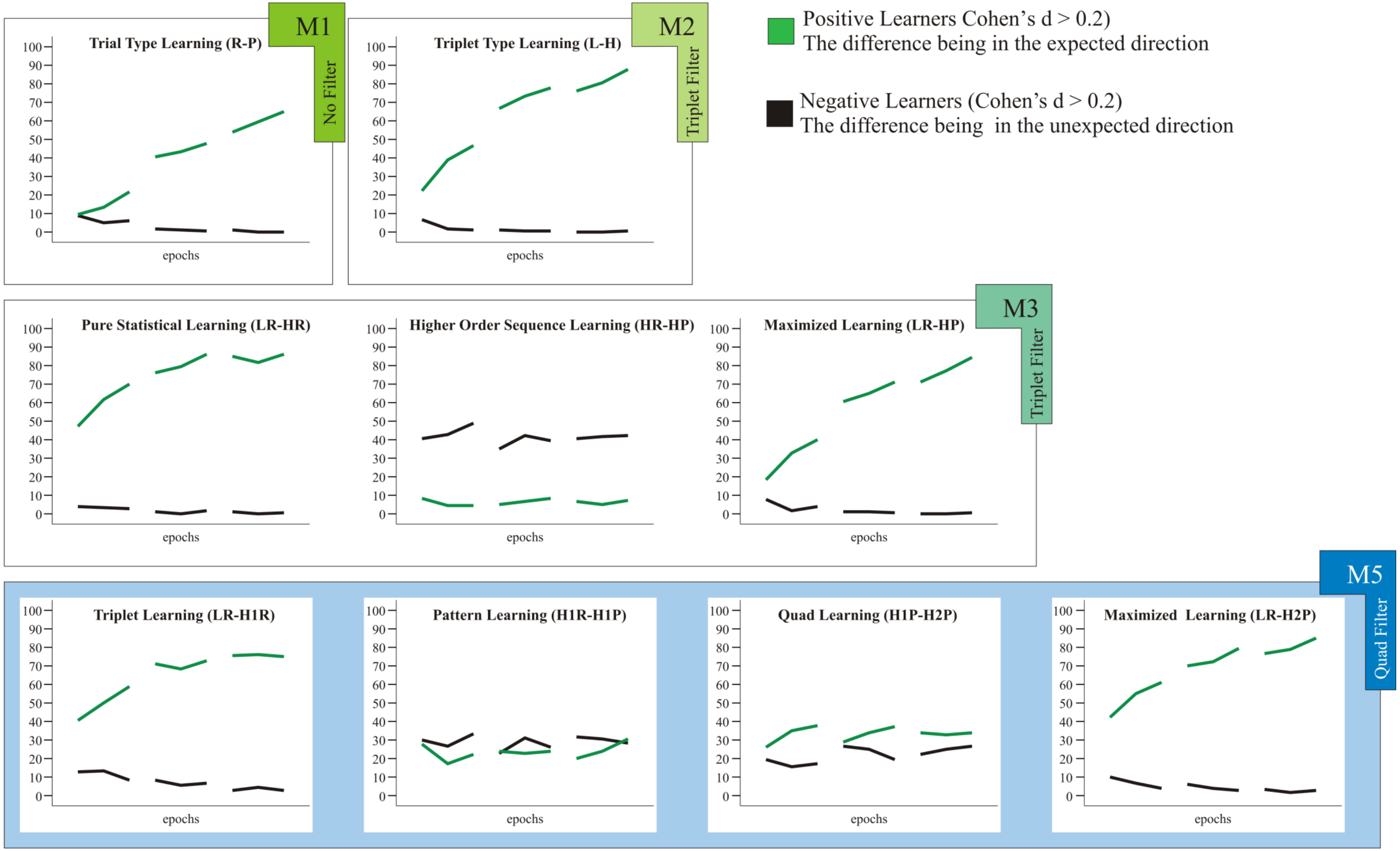
P1-P6 = Pattern1-Pattern6, NF = No Filter, TF = Triplet Filter, QF = Quad Filter



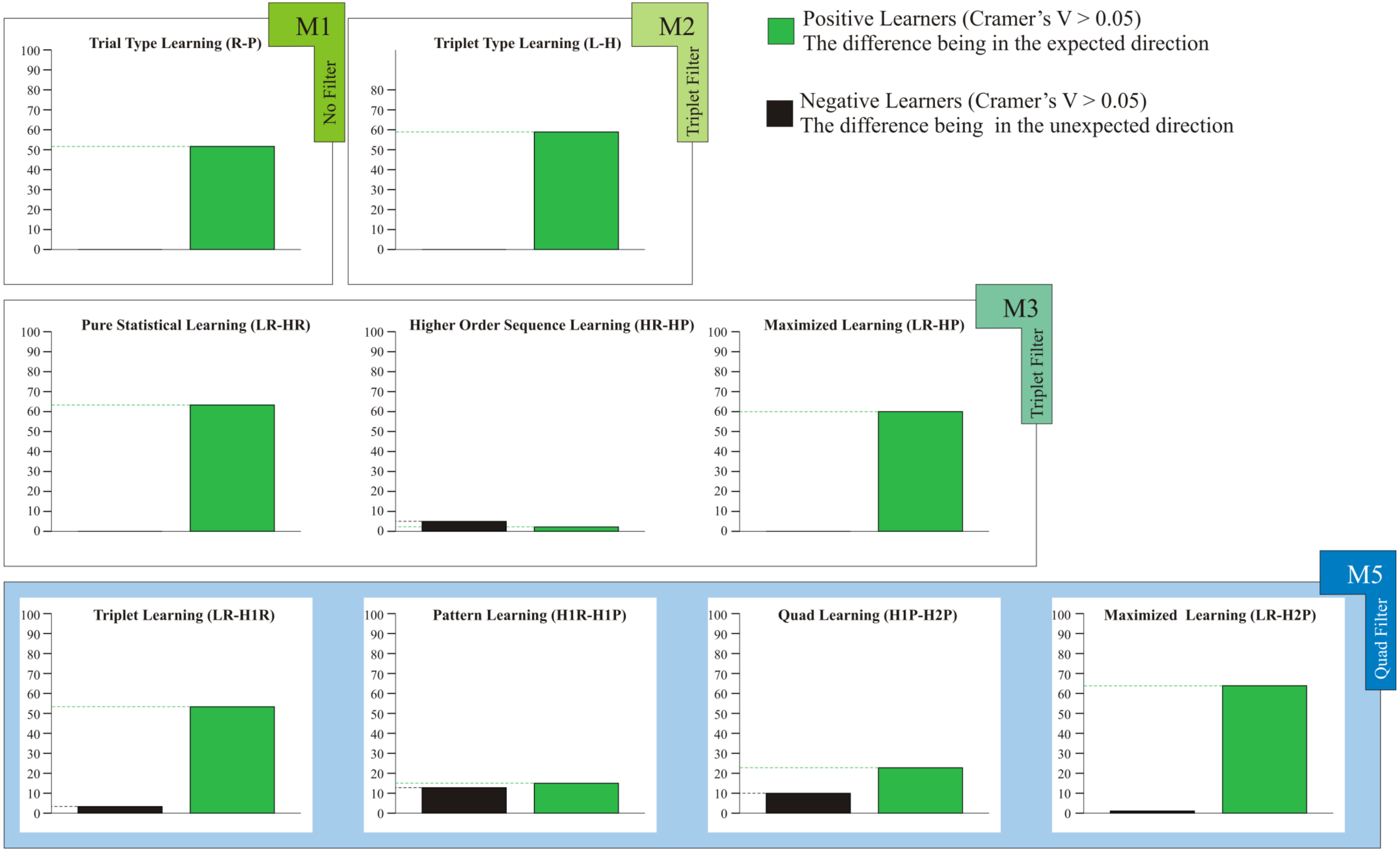
SF-V/1. Goodness of fit indicators (Adjusted R^2s) of each Model (Model 1-5) and each filtering method (No Filter, Triplet Filter and Quad Filter) as a function of epochs (1-9). Discontinuities of the lines indicate pauses between the three Sessions (Epochs 1-3, Epochs 4-6 and Epochs 7-9). Error bars represent the 95% CI.



SF-V/2. Reaction time based learning scores calculated the typical way (M1 nofilter, M2 triplet filter and M3 triplet filter) and the proposed way (M5 quadfilter) in each of the nine epochs. There was a longer pause between epochs 3 and 4, and between epochs 6 and 7 (creating three sessions, indicated by different colors). Bars represent the mean of the individual learning scores. Error bars represent 95% CI.



SF-V/3. The percentage of participants showing learning based on reaction times with an effect size of Cohen's d > 0.2, separately in each of the nine epochs. Green lines represent participants whose learning scores were positive (i.e. the observed difference was in the expected direction). Black lines represent participants whose learning scores were negative (i.e. in the unexpected direction). The discontinuity of the lines indicate pauses during data collection (i.e. there were three sessions).



SF-V/4. The percentage of participants showing learning based on error rates with an effect size of Cramer's V > 0.05 (data of the nine epochs were collapsed into a single category due to low overall error rates). Green bars represent participants whose learning scores were positive (i.e. the observed difference was in the expected direction). Black bars represent participants whose learning scores were negative (i.e. in the unexpected direction).

ST-V/16. Correspondance between the „reliably positive learner" status of participants with the currently proposed method vs. the usual analysis methods. Phi coefficients were calculated and are shown in the table. The scores are based on reaction times.

		M1 No Filter	M2 Triplet Filter	M3 Triplet Filter		
		R-P *TrialType effect*	**L-H** *Seq. Spec. L.*	**LR-HR** *Pure Stat. L.*	**HR-HP** *Higher Ord. L.*	**LR-HP** *Max. Learning*
M5 Quad Filter	**LR-H1R** *Triplet Learning*	.301**	.389**	.442**	N/A	.442**
	H1R-H1P *Pattern Learning*	.141	-.092	-.243**	N/A	-.031
	H1P-H2P *Quad Learning*	-.100	-.010	-.087	N/A	-.064
	LR-H2P *Max. Learning*	.242**	.392**	.257**	N/A	.293**

M1-M5: Model1 – Model 5

N/A – statistics not available due to the lack of reliable *M3 Higher Order Learning* learners

** significant, p < .005

Supplementary References

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

Fu, Q., Dienes, Z., & Fu, X. (2010). Can unconscious knowledge allow control in sequence learning? *Consciousness and Cognition*, *19*(1), 462–474. https://doi.org/10.1016/j.concog.2009.10.001

Gaillard, V., Cleeremans, A., & Destrebecqz, A. (2014). Dissociating conscious and unconscious learning with objective and subjective measures. *Clinical EEG and Neuroscience*, *45*(1), 50–56. https://doi.org/10.1177/1550059413516757

Howard, J. H., & Howard, D. V. (1997). Age differences in implicit learning of higher order dependencies in serial patterns. *Psychology and Aging*, *12*(4), 634–656.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541. https://doi.org/10.1016/0749-596X(91)90025-F

Nemeth, D., Janacsek, K., & Fiser, J. (2013). Age-dependent and coordinated shift in performance between implicit and explicit skill learning. *Frontiers in Computational Neuroscience*, *7*. https://doi.org/10.3389/fncom.2013.00147

Song, S., Howard, J. H., & Howard, D. V. (2008). Perceptual sequence learning in a serial reaction time task. *Experimental Brain Research. Experimentelle Hirnforschung. Experimentation Cerebrale*, *189*(2), 145–158. https://doi.org/10.1007/s00221-008-1411-z