Studies in Second Language Acquisition (2021), 1–15
doi:10.1017/S027226312100019X

Critical Commentary

WHY IS THE COMPONENTIAL CONSTRUCT OF IMPLICIT LANGUAGE APTITUDE SO DIFFICULT TO CAPTURE?

A COMMENTARY ON THE SPECIAL ISSUE

Pierre Perruchet [*]

Université de Bourgogne

Abstract
Although this special issue reveals some promising achievements, most of the contributions show that tasks of implicit learning are not or are only weakly correlated with each other, and they have inconsistent predictive power on L2 acquisition. This commentary examines four possible explanations for this surprising pattern: The (suboptimal) selection of tasks, the low reliability of measures, the deep influence of the starting level even for nominally "new" implicit tasks, and the fact that the mastery of L2 may involve other implicit processes than implicit learning measured through laboratory tasks.

INTRODUCTION

The following comments do not attempt to recapitulate the main findings of the special issue, and I apologize in advance for not doing justice to the promising perspectives that this issue suggests. Rather, I will simply highlight a few points that may have constructive interactions with the implicit learning approach conducted in cognitive psychology.

TERMINOLOGICAL ISSUES

Although terminological questions are often peripheral to deeper theoretical issues, they can lead to unnecessary complications. For instance, calling the same task, namely the Serial Reaction Time (SRT) task, a measure of either implicit learning (Li & Qian, Yilmaz & Granena) or procedural memory (Fu & Li) is at best confusing. A general guideline for terminological issues could be to choose labels that are the least theoretically loaded, while being as close as possible to the common language. Regarding the terms "memory" and "learning," their use has long been well established: Memory involves the recovery of

[*] Correspondence concerning this article should be addressed to Pierre Perruchet, LEAD/CNRS, Pole AAFE, Université de Bourgogne, 21000 Dijon, France. E-mail: pierre.perruchet@u-bourgogne.fr

specific stimuli or events, whereas learning refers to the exploitation of rules or regularities emerging from the repetitions of the same or similar stimuli or events. In this regard, an SRT task is a learning task, and LLAMA-D, which is a standard recognition test, is a memory task. This obviously does not preclude the possibility that memory and learning share some of their mechanisms, but as argued by Yilmaz and Granena, it would be an impoverishment not to mark the structural differences of situations, all the more so as well-fitted words are part of everyday language.

The choice between the qualifiers "implicit," "statistical," or "procedural" is more controversial. In the general literature, when associated with "learning," "implicit," which was prevalent in the 1990s, tends to be replaced by "statistical" or "implicit/statistical." "Statistical" conveys a theoretical interpretation, namely the idea that faced with a rule-governed structure in incidental learning conditions, a learner is unable to infer the rules and proceeds instead by exploiting the statistical regularities emerging as a by-product of the rules. Because this interpretation is now almost consensual, this terminological substitution raises no major problem. However, "implicit" may be preferred (and will be used hereafter) for three reasons. First, "statistical learning" may be ambiguous because the term is sometimes used to designate the specific experimental setting historically linked to this terminological innovation (namely the segmentation in relevant units of continuous sequences). Second, "implicit" has the advantage of being also used in association with "memory," thus making immediately salient the commonalities between the areas of implicit learning and implicit memory. Finally, the assertion that the learners exploit statistical regularities is often tacitly taken as equivalent to the idea that learners compute statistics, such as frequency or transitional probabilities (TP). In fact, just as the sensitivity of a nail to the number of hammer blows received does not mean that it has counted them, the human sensitivity to the statistical properties of the input does not imply the computation of these statistics. This is just a theoretical option, often coined as the "TP-based" account. There are other interpretations, which have been implemented in various "chunk-based" models, in which relevant representations of the input emerge from a selection process, without any statistical computation (for a review, see Perruchet, 2019).

Fu and Li, Buffington et al., and other contributors in the second language domain of research refer to the declarative/procedural distinction, which played a key role in Anderson's ACT model (e.g., Anderson, 1983). I will come back to this distinction later when discussing the notion of proceduralization, but to anticipate, the theoretical grounding of the declarative/procedural distinction has been criticized as being rooted in a computational view of mind, in which everything is reducible to, or emerges from, a declarative version of knowledge (e.g., Kolers & Roediger III, 1984). If we take the words out of any theoretical anchorage, with the meaning they have for the layman, a "procedure" may be defined as a planned sequence of operations oriented toward a goal. Endorsing this meaning makes it difficult to see the reason for which an SRT task, for instance, would tap "procedural memory." Of course, performing an SRT task involves a procedure, but this procedure (looking at the screen, pressing a given button on a keyboard in response to a given stimulus, etc.) is not the object of researchers' interest. The object of interest is the discovery and exploitation of the repetition of some subsequences, and these operations are unrelated, if not opposite, to the notion of planned actions. Similar arguments apply to other tasks of implicit learning. In any event, equating "procedural" and “implicit” seems confusing because (a) implicit learning does not involve an initial declarative stage (Yilmaz & Granena) and (b) a procedure can clearly be explicit (but see Li & DeKeyser and Fu & Li for other standpoints).

IS THERE A CONSTRUCT OF “APTITUDE FOR IMPLICIT LANGUAGE LEARNING”?

Most of the contributions to the special issue are aimed at evaluating whether performances on task of implicit memory or implicit learning can assess individual aptitudes, which would be potentially predictive of L2 acquisition. There is a consensus, both within and outside the field of Second Language Acquisition (SLA), that there is no *single* implicit ability. In their introduction to the special issue, Li and DeKeyser rightly defined the hypothetical construct of implicit aptitude in terms of clusters or components. Do the seven empirical contributions of the issue bring some support to this more plausible, componential framework? I am aware that, as a nonspecialist of SLA, I may miss the emergence of promising relationships. However, my own feeling is that even a componential view in which components would be defined as a manageable number of clearly identified dimensions predictive of specific aspects of L2 mastery has little support to date.

Several papers suggest a more optimistic conclusion, but, in my opinion, this optimism is at least partly underpinned by a tendency to treat performance in a specific task as if it were a component of a general ability. This tendency seemingly results from performing factor analysis and other multivariate analyses while using a small number of variables. For instance, Yilmaz and Granena labeled as “implicit learning ability” a factor from their Principal Component Analysis with a strong positive loading on an SRT task, and a much weaker *negative* loading on a letter span task. Without a theoretically grounded explanation of how a high implicit learning ability could elicit a low score in a letter span task,[1] this amounts to assess the implicit ability from the performance in the SRT task alone. The analysis reported in Godfroid and Kim is still clearer in this respect, as each of the first two factors emerging from their Exploratory Factor Analysis between four implicit learning tasks was defined by a single task (a third factor pooled together two identical tests of statistical learning except the involved sensory modality). Admittedly, there is nothing technically wrong here, but talking about a factor or a component instead of a task naturally suggests to the reader the existence of a latent construct that could be assessed with a set of more or less substitutable tasks. There is absolutely no evidence of such latent constructs. As a case in point, the tasks exploited by Yilmaz and Granena (SRT) and Godfroid and Kim (ASRT, with “A” standing for “Alternating”), although sharing a large number of features, are at best weakly correlated (see following text); and using the same general label for both wrongly suggests that they belong to a same, identified cluster. Moreover, in Godfroid and Kim, structural equation modeling leads the authors to conclude that the ASRT, taken as a measure of implicit learning aptitude, is specifically predictive of accuracy-based, timed language implicit tests of L2 mastery. Now, when looking at the correlation matrix, it appears that the correlations of the ASRT task with the four timed tests were .23, .14, .07, and −.06, respectively (with only the first one reaching significance). All this puts into question the existence of genuine components of implicit learning aptitude, and the idea that some components would be predictive of identified features of L2 performance. The data can alternatively be described as no more than a few (weak) correlations between specific tasks.

I will focus on what I see as the big picture, namely, to borrow Godfroid and Kim's formulation, the finding that "implicit-statistical learning aptitude does not generalize beyond the measure with which they were obtained." Of course, this also includes the observation that these measures are often not predictive of L2 acquisition or final attainment.

WHY IS THE BIG PICTURE SURPRISING?

It could be argued that low and sometimes negative correlations between implicit tasks can be easily explained by variations in the nature of the tasks (e.g., motor vs. cognitive; verbal vs. nonverbal), in measures (e.g., speed vs. accuracy), in sensory modalities (visual vs. auditory), by the dependence of performance on various fluctuating factors such as the level of attention or interest, and the list could be lengthened. Most of these commonsense arguments, however, break down in the face of the observation that correlations between explicit tasks involving roughly the same sources of variations are generally positive and much stronger. Everything happens as if the only shift from explicit to implicit was sufficient to dilute the correlational structure.

This raises a paradox. On the one hand, it is largely acknowledged that there are stable and enduring individual differences in L2 performances even after controlling for the conditions of immersion, and likewise for performances in most other complex domains. On the other hand, it is also increasingly acknowledged that implicit learning is essential for any complex acquisition in real life, and during the last three decades, a great deal of implicit learning studies following standard experimental approaches have led to the discovery of robust empirical phenomena, which are held as being largely independent of the specific settings in which they have been established. Regarding the acquisition of L2, the involvement of implicit learning processes is made all the more likely as it coincides with the usage-based approaches to language acquisition, as noted by Li and DeKeyser. Given these two premises, attributing stable individual differences in L2 acquisition to some implicit learning ability seems to be the natural next step. If such a conclusion turns out to be difficult to confirm empirically, then the validity of the premises must be questioned. More precisely, if the correlational structure is systematically stronger for explicit aptitudes than for their implicit counterparts, this may mean that explicit aptitudes are prevalent in real-world settings. Before endorsing such a disconcerting conclusion, it is worth examining whether the puzzling outcome of studies reported in the special issue and elsewhere could receive other explanations. In the text that follows I explore four possibilities.

TASK SELECTION

A first possibility to account for the elusive correlational structure between implicit learning tasks is that the selection of tasks is inadequate. Five out of the seven empirical papers from the special issue used an SRT or ASRT task, which is legitimized by its status as one of the oldest and most exploited measures in the general literature. However, the other selected tasks are more surprising. As noted in the preceding text, the LLAMA-D, adopted in three studies, is a recognition task. A widespread view is that recognition relies on two components: recollection and familiarity, which are considered as explicit and implicit components, respectively. Therefore, the LLAMA-D can be construed at best as partially dependent on implicit memory. The other task that is used more than once is the Tower of London, a test devised to measure executive functions and whose relationship with implicit learning is unclear, to say the least. Except for SRT tasks, the selection of the other tasks is all the more unexpected as a number of theoretically well-grounded tasks are available in the literature, such as invariant learning (e.g., McGeorge & Burton, 1990), contextual cuing (e.g., Chun & Jiang, 1998), or the Hebb repetition task (e.g., Szmalec et al., 2009). Still more puzzling is the neglect of the tasks that have been elaborated to be the closest to language learning. Artificial grammar tasks were conceived as an approach to syntax acquisition, and statistical tasks in the line of Saffran et al. (1996) were conceived as an approach to lexical learning, but neither of them is exploited (Godfroid and Kim used statistical learning tasks but replaced the oral standard syllables with nonverbal sounds).

Whether exploiting more conventional tasks of implicit learning would have resulted in better correlations, especially between aptitude and L2 performance, is a matter for further investigation. However, I am somewhat pessimistic about the outcome. First, even those tasks that are devised to be the closest of language acquisition are still far from genuine linguistic tasks, notably because they involve semantically vacuous stimuli (Li & DeKeyser). Second, as noted by Hedge et al. (2018), "it should not be assumed that robust experimental paradigms will translate well to correlational studies" (p. 1177). This is because the reason ensuring easily replicable experimental effects—low between-participant variability—is also the reason that makes their use as correlational tools problematic.

The few available data do not favor the idea that standard tasks of implicit learning are correlated,[2] even when they look similar. Godfroid and Kim used two tasks of statistical learning involving the same formal types of regularities and differing only by the sensory modality (visual and auditory). The correlation was significant but reached only .49. This could attest to the importance of the sensory modality, but in the same study, the visual statistical learning task did not correlate ($r = .05$) with another visual implicit learning task (ASRT), ruling out the idea that the sensory modality is a major organizing principle.

Buffington and colleagues also report a significant, but moderate correlation (Spearman's $\rho = .38$) between two formally similar tasks, SRT and ASRT (note that Parshina et al. [2018], cited in Buffington et al., found a numerically negative correlation between the two tasks, $r = -.18$). Unsurprisingly, the situation becomes worse when the implicit learning tasks become more dissimilar. As often mentioned in the papers of the special issue, Gebauer and Mackintosh (2007) reported no correlations (r's from $-.03$ to .01) between three prototypical, but very different, implicit learning tasks. It is worth adding that these data were obtained under the standard, incidental instructions. With explicit instructions, in which participants were informed of the existence of rules to be discovered, Gebauer and Mackintosh observed that all the correlations increased, with the correlation between SRT and artificial grammar learning reaching significance. These data are only indicative, but nevertheless they lend support to the idea that the weakness of the correlations is due less to the choice of tasks than to their implicit nature.

THE LOW RELIABILITY

Most of the contributors to the special issue evoke the low reliability of tests as a possible explanation for low correlations. As a rule, the reliability of implicit tests is indeed lower than the reliability of their explicit counterparts. Suzuki reports that the mere fact of implementing incidental instructions, everything else being equal, resulted in a much lower internal consistency of LLAMA_D (Cronbach alpha = .20) compared to that obtained in other studies with intentional instructions (e.g., Yilmaz & Granena and Bokender & Bylund [2020] reported Cronbach alphas of .50 and .54, respectively), confirming the detrimental consequence of implicitness per se. The question arises: Is this attenuated reliability a psychometric glitch that could be easily corrected, or is it a more fundamental problem inherent in implicit learning, with only limited corrective actions possible?

Let us consider the ASRT task. Godfroid and Kim reported a reliability coefficient of .96, whereas Buffington et al. obtained a negative coefficient. Crucially, reliability was computed separately for the repeated trials and the random trials in Godfroid and Kim, and was computed on the difference score in Buffington et al. This suggests that computing a difference score is responsible for the discrepancy, in keeping with the well-known principle that difference scores have a lower reliability than the raw scores (given that error variances add up). However, this explanation is partly misleading or at least incomplete. Some implicit learning tasks do not require a difference score, and many explicit learning tasks imply some comparison with a baseline or a control situation. The most decisive factor is not the difference as such, but whether the raw scores correlate. In Buffington et al. the correlation between pattern and random scores was nearly perfect ($r = .99$). This correlation provides a technical explanation for the lack of reliability, the basis of which has been described in the statistical literature for a long time (e.g., McNemar, 1962, p. 157). However, the question of interest for our concern is only shifted by one step: Why are the raw scores so highly correlated? The response is straightforward. If the raw scores are correlated, this entails that pattern scores reflect the same sources of variation as random scores, such as visual acuity, speediness of motor responses, interest for the task, level of vigilance or fatigue during the experiment, and many other factors. These sources of variation may have good reliability (Godfroid & Kim), especially when reliability is assessed on a within-session basis, but they are a priori unrelated to learning about the sequential regularities. Crucially, the effect of learning, which should affect repeated trials exclusively and should therefore elicit a divergence between repeated and random trials, is missing or negligible. The hypothesis that learning would occur but would result in the very same amount of improvement for all learners is highly implausible, and in any cases refuted by the data. In Godfroid and Kim, the mean score for the ASRT task was numerically negative (−1.15 ms, Table 4), suggesting that no learning occurred (as an aside, this raises the question of the basis of the correlations between the ASRT task and two—out of nine—measures of L2 knowledge). In Buffington et al., the mean score was positive, but negligible (2.99 ms, Table 3). Given the reported variance and assuming a normal distribution, this implies that approximately a third of the subjects obtained a negative score, and if a negative score is attributed to random influences that can act either negatively or positively, then this means that two-thirds of the subjects failed to learn anything about the repetitions. This analysis points to the excessive difficulty of the task as a real cause for the low reliability of implicit learning scores assessed through a difference. Moreover, this analysis shows that taking the raw scores for the pattern sequence instead of a difference score is inappropriate: If the scores for the pattern and random trials are virtually confounded, this implies that the scores for the pattern trials does not reflect learning.

Is using simpler tasks the solution for increasing reliability? An ASRT task as used in Buffington et al. and Godfroid and Kim is especially difficult because regularities are intermixed with noise, and several studies have shown that extracting regularities from noise is surprisingly challenging for humans (e.g., Rey et al., 2020). Performances obtained in standard SRT tasks are often much better, and as a consequence, reliability is generally acceptable. Fu and Li reports a Cronbach's alpha of .89, but it is unclear whether this value was obtained on the difference score or the original RTs. The reliability coefficients of difference scores in the other contributions to the special issue involving SRT were.79 (Yilmaz & Granena) and .76 (Buffington et al.). Further simplifying the to-be-learned regularities with the aim of further improving reliability is not advisable, as it would increase the likelihood that the learner would switch to an explicit style of thinking. The regularities embedded into an implicit learning setting must be sufficiently hidden to avoid the possibility that they pop out in learners' awareness. The conclusion is that the lower reliability obtained in implicit learning situations than in explicit learning situations is likely a consequence of the increased difficulty to learn in implicit conditions, a characteristic that is required to keep the implicit character.

This does not entail that the whole picture is only a matter of low reliability. As just mentioned, the reliability coefficients for the difference scores in SRT reported in the special issue were above .75. For the other implicit tasks, the coefficients of reliability as assessed by Cronbach alpha or split-half method, were .83 for the Frequency Following Response (FFR), a neural measure of auditory processing (Sun & Saito), .51 for syntactic priming (Li & Qian), .50 for LLAMA-D (Yilmaz & Granena), and .75 and .68 for VSL and ASL, respectively (Godfroid & Kim). These values do not prevent significant correlations. As an illustration, in Godfroid and Kim, given the coefficients of reliability for VSL and ASL (mean = .72) and the coefficients of reliability for the nine linguistic tests of L2 grammar knowledge (mean = .68), the theoretical maximum correlation between the two sets of tasks was .70 on average (the geometric mean of the two values). Now, the correlations between the tests of VSL and ASL, on the one hand, and the nine L2 tests, on the other hand, were virtually null (the mean of the 2 × 9 correlations was $r = .01$). Sun and Saito provided another illustration of the decoupling between fidelity and correlations. In their study, the implicit measure of auditory processing (FFR) and the explicit measure (musical memory) were roughly equally reliable (.83 and .86, respectively), but only the latter was predictive of gains in L2 prosody perception. Clearly, the observed pattern of correlations tells us something deeper about implicit learning than the low reliability of the tasks devised to measure it.

THE DEEP INFLUENCE OF THE STARTING LEVEL

Both in the general literature on implicit learning and SLA research, there is a tacit assumption that in front of a novel task, implicit learning processes starts from a blank slate. In this case, performance improvement should depend only on learning abilities, and if these abilities are common for a standard implicit learning task and for an L2 test, the two sets of data should correlate. But the reasoning no longer holds if implicit learning processes do not start from scratch, even in presumably new settings, and instead depend heavily on earlier experiences. Indeed, these experiences are obviously different for different individuals. I will examine first the empirical supports for such a hypothesis, and then I will elaborate on its explanatory power.

Empirical evidence is recent and still limited, but nevertheless highly compelling. Fu and Li studied the effect of feedback in the learning of the English past tense by Chinese students. They report a multiple regression analysis including as predictors three measures of what they called procedural memory, declarative memory, and working memory, respectively, and, crucially for our concern, pretest scores were included as a fourth predictor. The results were straightforward: The pretest scores were by far the best (and often the only significant) predictor of performance. It is worth adding that this finding was especially pronounced for the Elicited Imitation Test (EIT), which was designed to assess learners' implicit knowledge, and for the task-only group, which received no feedback (everything else being equal, feedback tends to invite explicit learning, so feedback is usually prohibited in implicit learning studies). In other terms, performances assumed to be the most dependent on implicit learning were also the most sensitive to pretest scores. Li et al. (2019) reported very similar data in a study using the English passive voice as a target structure (see also Yalçin & Spada, 2016). Somewhat surprisingly, the prevalent effect of the starting level occurred even though the performances in the pretest were very low. In Fu and Li, the mean pretest score for EIT was 3.32 out of 24, and the scores were still lower in Li et al. (2019) (from 1.19 to 2.28 out of 30, as a function of groups). Li et al. concluded that "even though the passive voice was a new structure, pretest scores were a strong and consistent predictor" (p. 717).

The strong dependence of learners' performances on prior knowledge provides an obvious explanation for the low correlations observed with aptitudes. The presence of a strong predictor makes the existence of other independent strong predictors impossible. Bokander (2020) compared the learning of Swedish as L2 by students whose L1 was either typologically similar (the Germanic language group) or distant (non-Germanic language group). Unsurprisingly, the former group performed better than the latter group in a test of Swedish proficiency performed at the end of training, an effect known as cross-linguistic influence. The point is that all participants also completed a language aptitude test (LLAMA test battery) before training. A regression analysis on the whole sample revealed that only the effect of L1 background was significant (beta = .41; the betas for the aptitude measures were between .11 and .19). L1 is usually fixed in SLA research, but individual differences in the predisposition to learn a specific L2 might still persist and reduce the predictive value of standard tests of aptitude.

The effect of the initial level is not only a matter of methodological considerations: It is also a valuable source of information for our understanding of implicit learning. Of course, the strict empiricist idea of the brain as a *tabula rasa* has been invalidated a long time ago, but the laboratory approach of implicit learning may have created the illusion that a learner could be placed in an entirely new situation, on which previous experiences would have no or only minor impact. Because the standard experimental approaches rely on averaged performances, the influence of the learners' individual starting level cannot be evaluated. Indirect evidence nevertheless exists. Perruchet and Tillmann (2010) used a standard design of word segmentation with an artificial language (played as a continuous succession of syllables without any prosodic markers), except that they recorded the probability for the artificial words of the language to be perceived as words from the outset even though all of them were random sequence of syllables. In this design, the pretraining level does not depend on the characteristics of participants, but on the characteristics of the artificial words. The key point, nevertheless, is that differences existed at the very beginning to the learning session. The results showed that the tendency to perceive a string of syllable as a word from the outset significantly favored the *learning* of this word, the initial effect being significantly amplified with training.[3]

The powerful effect of the starting level as revealed in the studies on individual differences in L2 acquisition is not easy to encompass within the historically prevalent model of implicit learning, in which learning is the results of statistical computations. For instance, for Shukla et al. (2007), transitional probability computations (or other forms of statistical computation) over syllabic representations of speech rely on encapsulated, automatic processes, which proceed irrespective of the material. If such was the case, learners' performance should directly depend on the statistical structure of the current environment. I alluded in the preceding text to competing accounts of implicit learning, based on chunk selection. The strong effect of the starting level is much easier to explain in a chunk-based framework because the perceptual and cognitive units that have been shaped in previous implicit learning episodes naturally guide the coding of any new situation (e.g., Perruchet et al., 1997).

These observations could provide an answer to the opening question: Is it reasonable to attribute to implicit learning a crucial adaptive function generating stable and enduring individual differences in real-world situations, if implicit learning abilities measured in artificial settings appear so volatile? An apparently paradoxical response could be that the two statements, which seem irreconcilable, are in fact causally dependent. If implicit learning has powerful and long-lasting effects on *all* aspects of an individual's life, these effects must pervade situations that the investigator considers to be new. As a consequence, performance in the presumably new situations will depend more on the coding units shaped throughout prior individual's experiences than on the results of tests devised to assess a general learning aptitude.

THE INVOLVEMENT OF OTHER PROCESSES

The last possible explanation for the low predictive value of implicit learning tasks I will discuss is that the implicit factors involved in L2 acquisition may comprise different processes than those involved in standard implicit learning tasks. Indeed, the recent interest in implicit learning should not overshadow the relevance of older literature. As mentioned in several contributions to the special issue, acquiring a second language certainly engages different processes at different stages of training, with a general shift from explicit to implicit processes. This shift has been previously observed in different research streams. Studies focusing on sensorimotor skills deserve to be mentioned first. They have given rise to the now classic model of Fitts and Posner (e.g., Fitts & Posner, 1967), in which performance is characterized by three sequential stages, termed the cognitive, associative, and autonomous stages. Another influential stream of research, which is certainly more relevant for L2 acquisition, followed the seminal papers by

Schneider and Shiffrin (e.g., Schneider & Shiffrin, 1977) around the notion of automatisms. The main distinction was between controlled and automatic processes, also conceived as modes of treatment operating in succession during extensive practice on tasks involving consistent stimulus-response mapping. This framework has been deeply challenged by Logan (1988), hence resulting in two major conceptions about the formation of automatisms. Both describe the initial stage of any supervised learning in complex domains as being explicit, attention-based, and involving sequential, algorithmic operations. The divergence arises about the changes induced by intensive practice.

In the first conception, the rule or the algorithm continues to be applied, but with a progressive withdrawal of attentional processes. Such a view—automatization as attention- withdrawal of an otherwise unchanged chain of operations—underlies the studies spurred by the Schneider and Shiffrin papers. This conception fits well with the computer metaphor of "compilation" proposed by Anderson (e.g., Anderson, 1983), or still, in my understanding, with the concept of "proceduralization" largely used in SLA research (DeKeyser, 2020). To illustrate with spelling: English-native children learn the morphological rule that all regular plurals are spelled with *s*, even when they are pronounced /z/. This rule would be first used intentionally, then the very same rule would be used unconsciously and without attentional control after extensive exposure to written material.

The second conception, developed in Logan (1988), is that intensive practice induces a shift from rule-based to memory-based processing. Automatized responses would be nothing more than the direct retrieval in memory of stored instances. To return to the previous example, in a memory-based model, an *s* would be added to, say, "bee" in the plural because "bees" was read in the past (with possible generalization based of surface similarities). In some sense, this interpretation posits the reverse of what the term of proceduralization suggests: Automatization would imply the *removal* of the initial procedure, and its substitution by a direct access to the result of the procedure stored in memory.

It is important to understand the differences between automatization, whatever the interpretation of the phenomenon, and implicit learning. Unfortunately, the relevant literature is nearly inexistent in fundamental research because automatization and implicit learning have been investigated using different experimental settings, and within historically separated schools of thought. By contrast, the problem of unraveling influences arises naturally in all applied settings in which initial explicit teaching is associated with or followed by extensive exposure to the raw data.

Let us return to spelling. Kemp and Bryant (2003) showed that children and adults partially base their spellings of plural words endings on the frequency with which letters co-occur in English. They wrote about their results: "They suggest that despite the existence of a morphology-based rule that is simple, reliable, and discussed at school, neither children nor adults consistently make (full) use of it in their spelling" (p. 73). This example highlights several points. First, it shows that instructed learning, which is generally conveyed as a set of simple rules or propositions, does not prevent the involvement of implicit learning processes. Second, the example illustrates a crucial aspect, which is related to a question raised for instance by Godfroid and Kim about the difference between what the authors call "explicit automatized" and "implicit" knowledge. The knowledge that results from the automatization or proceduralization of earlier declarative knowledge does not change in content over time, and hence remains primarily rule-based, or at least representational in nature (e.g., Anderson, 1993; Smith et al., 1992). After all, a compiled program follows the same algorithm and gives the same result as its interpreted counterpart. By contrast, the outcome of implicit learning cannot come from the proceduralization of some prior declarative knowledge, simply because frequency-based regularities exploited by implicit learning processes are difficult, and often impossible to translate into a declarative form. The impossibility to convey the skills acquired after extensive practice into a propositional format is a ubiquitous observation for any expert trying to pass on his or her expertise. Calling "procedural" any form of implicit knowledge thus leaves aside an essential part of the functioning of the mind, which cannot be conceived as deriving from an initial declarative form, because it does not belong to the propositional sphere.

Using a compacted rule, relying on memorized exemplars, and exploiting statistical regularities are not exclusive of one another, and these accounts are certainly more or less relevant as a function of domains. For instance, the Logan's instance-based model has strong supporting evidence for mental calculation, where the number of possible instances is limited (although the idea of automated counting procedures is still advocated [Thevenot & Barrouillet, 2020]). The role of implicit learning is ascertained in more complex domains, such as spelling (for a review, see Pacton et al., 2019). Regarding SLA, my intuition as a nonexpert would be that none of the potential mechanisms evoked in the preceding text can be eliminated a priori, and the problem of attributing a given performance to one process rather than another arises. In an overwhelming proportion of situations, the three kinds of hypothetical processes lead to the same output, namely the correct response. The key solution, which was adopted by Kemp and Bryant (2003) and many others, consists in constructing atypical situations likely to generate a type of error predicted by one account and not by the other two (in the same way as perceptual illusions or cognitive biases are exploited in other domains).

For the present concern, the potential plurality of mechanisms engaged in L2 acquisition may have damaging consequences on the correlation patterns. In general psychology, the acquisition of automatisms is commonly assessed through a variety of tasks including dual-task paradigms, stop-signal paradigms, or Stroop-like tasks. These tasks have little, if any, resemblance with implicit learning tasks, and conversely, the latter are in no way designed for the evaluation of automatisms. It is probable that aptitudes required in each case are different. For instance, it has been observed in the literature of automatisms that an important source of individual differences is the subjects' propensity to let go of controlled processing (e.g., Schneider & Fisk, 1982), a factor certainly influential in L2 acquisition. By contrast, such a propensity to let automatisms take over is certainly not relevant for implicit learning because the sensitivity to the statistical properties of the material develops independently of any strategy. If, for instance, at a given stage, the level of automatization of the declarative knowledge gained in instructed sessions is the main determinant of individual differences in L2 production, then the predictive value of implicit learning tasks on L2 achievement would necessarily be restricted.

CONCLUSION

Starting from the big picture emerging from this special issue and earlier related studies, according to which tasks of implicit/statistical learning are not, or only weakly, correlated with each other, and have inconsistent predictive power for L2 acquisition, this article has examined four possible explanations for this surprising pattern: the (suboptimal) selection of tasks, the low reliability of measures, the deep influence of the starting level even for nominally "new" implicit tasks, and the fact that L2 achievement could also depend on the automatized exploitation of the knowledge initially conveyed through explicit teaching. This analysis suggests that some methodological adjustments could help to improve the global outcome. For instance, a gain should follow the selection of tasks clearly designed to capture implicit learning aptitudes (this should eliminate ToL, for instance), not too difficult to ensure satisfactory reliability (this should eliminate ASRT[4]), and preferably as close as possible to the field of language. Also, some control over learners' starting levels is usual (e.g., in most studies learners share the same L1), but more stringent criteria should be encouraged.

However, the expected improvement is clearly limited. Even standard implicit learning tasks have little or no correlation with each other and making them less difficult to perform must be done with caution because easy tasks can trigger the involuntary involvement of explicit processes. As mentioned by Li and DeKeyser, Siegelman et al. (2017) proposed to explore the various facets of implicit statistical learning starting from a mapping sentence such as: "Statistical Learning is the ability to pick-up {transitional/ distributional} statistics from the sensory environment, in the {visual/auditory} modality, when contingencies are {adjacent/ non-adjacent}, over {verbal/ non-verbal} material, across{time/ space}, {with/ without} motor involvement, thereby shaping behavior" (p. 4). Note that because the interactions are certainly relevant, the exploration of all the 64 (2^6) combinations is needed, each potentially pointing toward a specific implicit learning aptitude. Note also that this preliminary mapping sentence can be lengthened at will with a number of additional distinctions, such as whether performance is assessed through speed or accuracy criteria (Godfroid & Kim; Li & Qian), and with the number of resulting aptitudes increasing exponentially. Such an approach may turn out to be fruitful (see Growns et al., 2020, for promising results regarding correlations between visual statistical learning tasks). However, whoever engages in such a project must be aware that its realization appears to be extremely time-consuming because, in addition to all the constraints mentioned already, a single implicit learning task (i.e., the implementation of one out of the possible sentences mentioned) takes a long time to complete, and the number of tasks given to each learner is necessarily limited to avoid transfer or interference effects. Moreover, the current state of knowledge makes it likely that the conclusion will be somewhat disappointing, with the number of aptitudes hardly lower than the number of tasks tested.

Testing more specific hypotheses could be an alternative, and more manageable objective. Among the points raised in the special issue, the strong effect of even small differences in the starting level in L2 mastery (e.g., Fu & Li) appears to be intriguing and one whose exploration could deepen our understanding of implicit learning. An obvious question is whether this effect increases or decreases with prolonged L2 practice. More generally, the idea that L2 acquisition proceeds in successive stages with different aptitudes being involved for initial and later stages is evoked in several contributions. A common claim is that progression goes from a prevalence of explicit processes to the prevalence of implicit processes. The Skill Acquisition Theory proposed by DeKeyser (Li & DeKeyser), potentially amended by the principle of a distinction between the proceduralization of mostly rule-based declarative knowledge acquired in instructed settings, and the effect of never taught frequency-based regularities could provide a good start to further explore this issue. This would require the repeated measurements of L2 achievement throughout L2 exposure, with tests exploring the probability of errors in carefully selected situations, as correct responses under usual conditions are usually noninformative about the underlying processes. Of course, such an approach does not allow bypassing the problem of low correlations between tests, which plagues the entire field. However, repeated exposures, although primarily devised to explore the dynamic of the processes at play, could at least partially alleviate this problem. As an illustration, Fleischman and Rich (1963) administered to a group of subjects a standard test of spatial-visual abilities, and a second test that was especially designed to measure proprioceptive sensitivity. Then the same subjects received extended practice on a perceptual motor task involving two-hand coordination. The correlations between the performance on this task and the spatial ability measure systematically decreased throughout learning, whereas the correlations with the kinesthetic sensitivity measure systematically increased at the same time, as expected from the model proposed by Fitts (1951). Note that in this famous study, only 3 out of 20 correlations (10 repeated measures for each test) reached the conventional threshold of significance. Nevertheless, the data were clear-cut because even low correlations become informative when they are inserted into a systematic (here: monotonic or even linear) trend.

In addition to its potential interest in SLA research, dissociating the long-term consequences of formal teaching, whether it is conceived as the automatization of declarative knowledge or the memorization of instances, and the consequences of direct interactions of the learner with the raw data, leading to the exploitation of patterns of regularities that cannot be taught, would be an essential contribution to cognitive psychology as a whole. Indeed, such an objective is beyond the reach of a standard, laboratory approach, and can only be conducted in the few real-life situations in which the learners can benefit from both formal teaching and prolonged immersion in the appropriate environment.

NOTES

[1] The authors suggest an interpretation for their bipolar factor in note 6, according to which the attention-driven memory processes involved in the Letter span test would be detrimental for implicit learning. However, this account is inconsistent with the well-documented fact that implicit learning also requires attention. Moreover, in the so-called Hebb repetition task, which is a standard task of implicit learning, the to-be-learned sequences are displayed in span tasks, and learning occurs very quickly. This strongly suggests that the relation between span tests and implicit learning, if any, should be positive.

[2] This is also the case for implicit memory tests. I think I was the first one, more than 30 years ago (Perruchet & Baveux, 1989) to examine the correlations between memory tests, with tasks including two tests of explicit memory (recall and recognition) and four tests that were then the most largely used to investigate implicit memory. Correlations between implicit memory tests ranged from −.249 to .346, and a factor analysis, followed by Varimax rotation, returned two factors that did not match with the explicit/implicit distinction. A number of subsequent studies has confirmed that correlations between implicit memory tasks are low, inconstant, and elusive.

[3] The amplification of the initial difference implies that using a gain score does not eliminate the influence of the initial level. See Li et al. (2019) for a discussion.

[4] Another troublesome aspect is that I surmise (the point deserves verification) that none of the current computational models of implicit learning, whether TP-based or chunk-based, would be able of learning the regularities embedded in the ASRT task.

REFERENCES

Anderson, J. R. (1983). *The architecture of cognition*. Harvard University Press.

Anderson, J. R. (1993). *Rules of mind*. Erlbaum.

Bokander, L. (2020). Language aptitude and crosslinguistic influence in initial L2 learning. *Journal of the European Second Language Association*, *4*, 35–44. https://doi.org/10.22599/jesla.69

Bokender, L., & Bylund, E. (2020). Probing the internal validity of the LLAMA language aptitude tests. *Language Learning*, *70*, 11–47.

Chun, M. M., & Jiang, Y. H. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, *36*, 28–71.

DeKeyser, R. (2020). Skill acquisition theory. In B. VanPatten, G. Keating, & S. Wulff (Eds.), *Theories in second language acquisition* (pp. 83–104). Routledge.

Fitts, P. M. (1951). Engineering psychology and equipment design. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. Wiley.

Fitts, P. M., & Posner, M. I. (1967). *Human performance*. Brooks/Cole Publishing Company.

Fleischman, E. A., & Rich, S. (1963). Role of kinesthetic and spatial-visual abilities in perceptual-motor learning. *Journal of Experimental Psychology*, *66*, 6–11.

Gebauer, G. F., & Mackintosh, N. J. (2007). Psychometric intelligence dissociates implicit and explicit learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 34–54. https://doi.org/10.1037/0278-7393.33.1.34

Growns, B., Siegelman, N., & Martire, K. A. (2020). The multi-faceted nature of visual statistical learning: Individual differences in learning conditional and distributional regularities across time and space. *Psychonomic Bulletin & Rev*iew, *27*, 1291–1299. https://doi-org.insb.bib.cnrs.fr/10.3758/s13423-020-01781-0

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research and Methods*, *50*, 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Kemp, N., & Bryant, P. (2003). Do beez buzz? Rule-based and frequency-based knowledge in learning to spell plural-s. *Child Development*, 74, 63–*74*.

Kolers, P. A., & Roediger III, H. L. (1984). Procedures of mind. *Journal of Verbal Learning and Verbal Behavior*, *23*, 425–449.

Li, S., Ellis, R., & Zhu, Y. (2019). The associations between cognitive ability and L2 development under five different instructional conditions. *Applied Psycholinguistics*, *40*, 693–722. https://doi.org/10.1017/S0142716418000796

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*, 492–527. https://doi.org/10.1037/0033-295X.95.4.492

McGeorge, P., & Burton, A. M. (1990). Semantic processing in an incidental learning task. *Quarterly Journal of Experimental Psychology*, *42A*, 597–609.

McNemar, Q. (1962). *Psychological statistics* (3rd ed.). Wiley.

Pacton, S., Fayol, M., Nys, M., & Peereman, R. (2019). Implicit statistical learning of graphotactic knowledge and lexical orthographic acquisition. In C. Perret & T. Olive (Eds.), *Spelling and writing words* (Vol. *39*, pp. 40–66). Brill.

Parshina, O., Obeid, R., Che, E. S., Ricker, T. J., & Brooks, P. J. (2018). SRT and ASRT: Similar tasks tapping distinct learning mechanisms? In *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 2202–2207). Cognitive Science Society.

Perruchet, P. (2019). What mechanisms underlie implicit statistical learning? Transitional probabilities versus chunks in language learning. *Topics in Cognitive Sciences*, *11*, 520–535. https://doi.org/10.1111/tops.12403

Perruchet, P., & Baveux, P. (1989). Correlational analyses of explicit and implicit memory. *Memory and Cognition*, *17*, 77–86.

Perruchet, P., & Tillmann, B. (2010). Exploiting multiple sources of information in learning an artificial language: Human data and modeling. *Cognitive Science*, *34*, 255–285.

Perruchet, P., Vinter, A., & Gallego, J. (1997). Implicit learning shapes new conscious percepts and representations. *Psychonomic Bulletin and Review*, *4*, 43–48.

Rey, A., Bogaerts, L., Tosatto, A., Bonafos, G., Franco, A., & Favre, B. (2020). Detection of regularities in a random environment. *Quarterly Journal of Experimental Psychology*, *73*, 2106–2118. https://doi.org/10.1177/1747021820941356

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

Schneider, W., & Fisk, A. D. (1982). Degree of consistent training: Improvements in search performance and automatic process development. *Perception & Psychophysics*, *31*, 160–168.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: 1. Detection, search, and attention. *Psychological Review*, *84*, 1–66.

Shukla, M., Nespor, M., & Mehler, J. (2007). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology*, *54*, 1–32.

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society B*, *372*, 20160059. http://dx.doi.org/10.1098/rstb.2016.0059

Smith, E. B., Langston, C., & Nisbett, R. E. (1992). The case for rules in reasoning. *Cognitive Science*, *16*, 1–40.

Szmalec, A., Duyck, W., Vandierendonck, A., Mata, A. B., & Page, M. P. A. (2009). The Hebb repetition effect as a laboratory analogue of novel word learning. *Quarterly Journal of Experimental Psychology*, *62*, 435–443.

Thevenot, C., & Barrouillet, P. (2020). Are small additions solved by direct retrieval from memory or automated counting procedures? A rejoinder to Chen and Campbell (2018). *Psychonomic Bulletin & Review*, *27*, 1416–1418.

Yalçin, S., & Spada, N. (2016). Language aptitude and grammatical difficulty: An EFL classroom-based study. *Studies in Second Language Acquisition*, *38*, 239–263. https://doi.org/10.1017/S0272263115000509