*Studies in Second Language Acquisition*, 2016, **0**, 1–17.
doi:10.1017/S0272263116000474

# ESTABLISHING EVIDENCE OF LEARNING IN EXPERIMENTS EMPLOYING ARTIFICIAL LINGUISTIC SYSTEMS

Phillip Hamrick
*Kent State University*

Rebecca Sachs
*Virginia International University*

Artificial linguistic systems (ALSs) offer many potential benefits for second language acquisition (SLA) research. Nonetheless, their use in experiments with posttest-only designs can give rise to internal validity problems depending on the baseline that is employed to establish evidence of learning. Researchers in this area often compare experimental groups' performance against (a) statistical chance, (b) untrained control groups' performance, and/or (c) trained control groups' performance. However, each of these methods can involve unwarranted tacit assumptions, limitations, and challenges from a variety of sources (e.g., preexisting perceptual biases, participants' fabrication of rules, knowledge gained during the test), any of which might produce systematic response patterns that overlap with the linguistic target even in the absence of learning during training. After illustrating these challenges, we offer some brief recommendations

We would like to thank Ronald Leow, Alison Mackey, Nick B. Pandža, Kelli Ryan, and the reviewers for their helpful comments on earlier versions of this paper. We are also grateful to Patrick Rebuschat for his valuable guidance on our early forays into implicit learning research, and for collaborations with him that have positively influenced our work in numerous productive ways. All remaining errors are our own. Both authors contributed equally to this paper.

Correspondence concerning this article should be addressed to Phillip Hamrick, Language and Cognition Research Laboratory, Department of English, Kent State University, 475 Janik Drive, Kent, Ohio 44242. E-mail: phamric1@kent.edu

regarding how triangulation and more sophisticated statistical approaches may help researchers to draw more appropriate conclusions going forward.

It is no secret that there is no one-size-fits-all solution for creating well-controlled experiments. Perhaps less often recognized, however, is the fact that commonly accepted methods in certain areas of research can lead to invalid conclusions in others, even when the fields are closely related and the studies' research questions and designs are very similar. In this paper, we examine some problems that have arisen in studies investigating language learning using artificial linguistic systems (ALSs) with posttest-only designs. ALSs range from miniature artificial languages (e.g., *Dia libro gin volak aceti*, in de Graaff's [1997] eXperanto, meaning "That's the book I'd like to buy") to semi-artificial languages that mix features of natural languages (e.g., *Vet-ga injection-o elephant-ni gave*, in Williams and Kuribara's [2008] Japlish, meaning "The vet gave the elephant an injection") or add artificial features to natural languages (e.g., *The circus performer covered herself with gi snakes*, from Williams [2005], in which *gi snakes* refers to "NEAR-ANIMATE snakes").[1] The fact that ALSs can be designed to target the learning of particular linguistic phenomena while controlling for certain types of prior knowledge offers considerable advantages in second language acquisition (SLA) research. However, the approach has been a double-edged sword: On one side, "upgrading" from non-meaning-bearing artificial grammars (AGs; e.g., *TPTXXVPX*, from Reber's [1967] AG) to the more complex, meaningful sentences characteristic of ALSs has had the benefit of more accurately simulating the cognitive processes involved in natural language learning (Ettlinger, Morgan-Short, Faretta-Stutenberg, & Wong, 2015). On the other, the fact that ALSs contain multiple layers of linguistic cues (e.g., phonological and semantic) means that certain features may interact with participants' first-language (L1) knowledge or other preexisting biases, creating additional challenges that have not yet sufficiently factored into methodological decision making in our field.

SLA researchers employing ALSs often attempt to establish evidence of learning by comparing the performance of experimental groups against (a) statistical chance, (b) untrained control groups, and (c) trained control groups. Each method offers potential advantages over the preceding one, but, in combination with the nature of ALSs, all three involve unwarranted tacit assumptions or other problems that can threaten a study's internal validity. To have confidence that we are truly establishing evidence of learning in such experiments, further measures will need to be taken. Before elaborating on specific challenges and making recommendations for addressing them, however, we must make explicit several factors that shape our exposition.

First, it is important to acknowledge that the concerns we raise bear some similarities to discussions in the literature on AGs in cognitive psychology, where researchers have also debated what constitutes an appropriate baseline for measuring learning (e.g., Dienes & Altmann, 2003; Perruchet & Reber, 2003; Reber & Perruchet, 2003; Redington & Chater, 1996). At the same time, while we have been informed by those debates, most of the insights we present here developed in direct response to challenges in our own research using ALSs (e.g., Hamrick, 2012, 2013, 2014a, 2014b, 2015; Hamrick & Rebuschat, 2012, 2014; Rebuschat, Hamrick, Riestenberg, Sachs, & Ziegler, 2015; Rebuschat, Hamrick, Sachs, Riestenberg, & Ziegler, 2013). Through subsequently recognizing these challenges in other SLA researchers' work (e.g., Morgan-Short, Steinhauer, Sanz, & Ullman, 2012; Rebuschat & Williams, 2012; Tagarelli, Borges Mota, & Rebuschat, 2015), we have come to understand that certain dilemmas extend beyond the issues discussed in the AG literature precisely because language learning—even artificial language learning—is different from AG learning in important respects. Having experienced firsthand how complicated it can be to address these challenges even when aware of them, and recognizing that other researchers' understandings of internal-validity problems also evolve over time in the process of grappling with them, we prefer to critique primarily our own work and that of our close colleagues in the spirit of promoting constructive dialogue.

Second, the concerns we raise do not necessarily apply to ALS experiments with pretest/posttest designs and comparison groups. Our focus is on ALS paradigms where participants are not pretested, but proceed directly to a training phase (also called an "exposure" or "learning" phase) that is followed by a testing (and sometimes delayed testing) phase. For researchers using ALSs to investigate phenomena such as implicit learning, such posttest-only designs are quite common and may be essential to avoid orienting participants toward more explicit modes of thinking. That said, the considerations we discuss are equally applicable to studies of explicit learning that employ such designs (e.g., Hamrick, 2013) because what matters is not the type of learning under investigation, but rather the validity of the baseline for comparison that is used to establish whether learning has taken place. Our concerns also do not necessarily apply to experiments with appropriate within-subjects controls, as are often used in production and eye-tracking studies, among others. We would emphasize, though, that this is due to the nature of the controls, and not to the method of data collection or the type of processing under investigation. For instance, if a study were to assess learning using a posttest designed within the visual-world paradigm (e.g., Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) and compare participants' visual "selections" of multiple-choice options against chance, it could also suffer from the internal-validity problems we describe in the following text. Given the space constraints of a squib, we must be selective about the number and nature of our examples, reusing them where possible. However, we attempt to be clear about the circumstances under which researchers will need to address issues of the sort we identify.

Finally, and perhaps most importantly, our arguments depend in large part on how we define learning. For our purposes, "learning" is the development of target knowledge—that is, knowledge of the target form(s) an experimenter is investigating—due to exposure to the training materials, and not, for example, knowledge gained during the test phase or response patterns attributable to prior experience, innate or learned biases, or some other source beyond the experiment. Other definitions of learning may alter the applicability of our insights.

## COMPARISONS AGAINST CHANCE

Several studies employing ALSs in SLA have used statistical chance as the primary baseline against which to assess learning (e.g., Hama & Leow, 2010; Hamrick & Rebuschat, 2012; Morgan-Short et al., 2012; Williams, 2005). The logic is essentially this: Participants have never been exposed to the ALS before, so if they do not learn anything during training, then they should be guessing at test. Guess-based performance should be equivalent to a random selection of answers (i.e., chance), which can be calculated by dividing the number of correct answers by the total number of possible answers. For instance, in a typical four-alternative multiple-choice test, chance is considered to be 25% accuracy. If participants perform above that baseline to a statistically significant degree, they are considered to have developed target knowledge.

Is this assumption justifiable? One way of exploring this empirically is to examine the results of experiments that have employed both chance and human control groups as comparisons. In some of these studies, control participants who were not exposed to the training have performed at chance overall (Hamrick, 2014a, 2014b; Rebuschat & Williams, 2012, Exp. 1), enabling the researchers to argue that chance had been validated as a baseline for their particular experiments. However, in very similar studies, control participants have been found to deviate from chance overall (Rebuschat & Williams, 2012, Exp. 2) or on specific items (Hamrick, 2014a; Rebuschat & Williams, 2012; Rebuschat et al., 2015), making it untenable to consider the corresponding experimental groups' above-chance performance alone as evidence of target learning. Extrapolating out to the area of research as a whole, without a significant body of work establishing the precise circumstances under which participants will simply guess and perform at chance in the absence of training/learning, repeated findings of nonchance performance in human controls underscore that it may be invalid, and therefore irresponsible, to employ chance as the sole baseline without direct experiment-specific evidence to support its use.

Some of the risks associated with using chance as a baseline derive from two assumptions that are likely to be unwarranted in the context of ALS (or L2) learning. The first is that an equivalent group of human control participants would have no systematic biases that might cause them to deviate from chance in the absence of training/learning. This first assumption has two interpretations: (a) participants have no (relevant) systematic biases, or (b) any biases they have are heterogeneous and will wash one another out across items or across participants, leaving overall performance at chance. The second assumption is that participants will not learn any aspect of the linguistic target during the test phase. Importantly, if only comparisons against chance are made in an experiment, there are few ways of checking these assumptions, and (as we discuss in the recommendations section) methods of doing so can be resource intensive.

In relation to the first assumption, it is widely agreed that humans are not bias-free blank slates at birth, much less in adulthood. But how likely is it that their biases will be relevant in ALS experiments? Research conducted within a variety of frameworks, from Competition Model studies (MacWhinney, Bates, & Kliegl, 1984) to statistical learning experiments (Onnis & Thiessen, 2013), has demonstrated that adults' L1s give them consistent preferences for using particular cues to parse novel linguistic input. For instance, MacWhinney et al. (1984) found that L1-English speakers systematically relied on word order to identify the actor in English sentences with modified syntax (e.g., NNV: *the pig the eraser chases*; VNN: *chases the pig the eraser*), whereas German and Italian speakers relied on animacy and/or agreement cues to interpret sentences with equivalent modifications in their L1s. Considering that ALS experiments often utilize stimuli similar to those of MacWhinney et al., such findings suggest that researchers cannot assume unbiased guessing would occur in the absence of training/learning.[2] Without empirically established, robust, converging evidence of directly relevant processing phenomena in prior research or exact replications sampled from the same population, it may not be possible to predict how human participants would behave, even if researchers are armed with detailed knowledge of how L1 speakers tend to process a language in its unmanipulated form.

To elaborate on the consequences of this problem, if participants' language-related biases correlate (positively or negatively) with an intended linguistic target, they may perform above or below chance even in the absence of training/learning, making chance meaningless as a sole baseline for establishing evidence of learning. One illustration of this comes from Rebuschat's (2008; Rebuschat & Williams, 2012) investigation of the implicit learning of syntax in a semi-artificial language consisting of L1-English words arranged in German syntactic structures (e.g., *After the instructor a sword brandished, focused Brian more on his defensive stance*). In the study's exposure phase, without having been notified of the rule-governed nature of the stimuli, experimental participants listened to sentences in the ALS and judged whether they made sense; then, a surprise posttest asked them to judge whether new sentences seemed to follow the same rules. Their performance was significantly above chance, suggesting learning. However, by including a control group that was not exposed to any training materials, Rebuschat was able to discover unexpected systematicity in control participants' responses that could not have been learned during the experiment. The untrained controls' test performance sometimes departed significantly from chance, and the experimental group sometimes did not differ from the untrained controls, begging the question: What if some of the systematicity in experimental participants' responses was also due to factors independent of the training? Findings such as these highlight some of the difficulties involved in interpreting apparent learning effects in ALS experiments (see also Reber & Perruchet, 2003, p. 113) and raise red flags about the use of chance as the sole baseline in other studies.

Regarding interpretation (b) of the "no systematic bias" assumption, while we agree it is possible that variations in participants' response biases may cancel one another out, we would argue that it is, at this point, an empirical question requiring direct investigation within the contexts of particular studies. Considering, moreover, that SLA researchers tend to collect data from rather homogeneous samples (Plonsky, 2015), and that homogeneous samples are more likely to have internal correlations manifesting as systematic group-level behavior (Field, 2009), it seems reasonable to expect participants to show at least some similar response tendencies independently of training/learning, thereby making chance an inappropriate baseline. Additional examples of this (from Hamrick, 2013, 2014a; Rebuschat et al., 2015) will be presented in the following text in the sections on untrained and trained controls.

Finally, the other unwarranted assumption underlying the use of chance as a baseline is that participants will not learn any aspect of a linguistic target during the test phase. The fact that participants have never been exposed to a target previously does not exempt ALS research from this potential problem, as noted by Redington and Chater (1996) in the AG literature some 20 years ago. Human controls in ALS experiments do not always show evidence of learning during the test phase (e.g., Grey et al., 2014; Tagarelli et al., 2015), but sometimes they do—as, for example, when features of a posttest unexpectedly spur participants in all groups to search for rules (Rebuschat et al., 2015). The critical point is that if researchers are not able to say for certain that participants' response tendencies are the product of exposure to the training materials (and may be due instead to reactivity during a test, for instance), they cannot be taken to reflect learning in the study (Perruchet & Reber, 2003, p. 127).

In sum, using chance as a baseline entails unwarranted assumptions that may invalidate claims about learning in an experimental group. Untrained controls may help to address some risks associated with these assumptions; however, they come with their own limitations, as we discuss next.

## UNTRAINED CONTROL GROUPS

In SLA research, participants who are tested without having been exposed to training materials might be referred to as a maturation or testing control group depending on the nature of the study. Within the context of ALS experiments, we call such a group an untrained control group to contrast it with the type of trained control group we discuss in the next section. As illustrated in the preceding text, untrained controls offer certain advantages over the use of chance as a baseline: By allowing researchers to check whether preexisting biases or learning during a test might account for (at least some) systematic response tendencies underlying nonchance performance, they enable better-informed evaluations of whether apparent evidence of learning in an experimental group is genuine. However, the method is not fail-safe. For one thing, as we shall see with trained controls, learning during a test may be influenced by what participants have learned (or fabricated) during training, or even by the simple fact of having experienced a training session. For another, the extent to which untrained controls' response patterns can be interpreted as reflecting what has likely also happened in an experimental group may depend crucially on whether the researcher has been able to establish similar expectations, motivations, and understandings of the task across groups.

Why might participants' mind-sets differ across groups? To recap, in research employing ALSs, the goal is to determine whether participants develop new response tendencies based on exposure to training materials containing a linguistic target. The test instructions are then supposed to act as a prompt to rely on these newly acquired tendencies, and such behavior is counted as learning. Logistically, though, untrained controls can be problematic because instructions that make sense for experimental participants may seem decontextualized or may even be uninterpretable for control participants who have not experienced a training phase. The instructions may therefore have to be modified, and this can cause the control and experimental conditions to differ in additional ways beyond exposure to the training. If the groups have different expectations or understandings of the nature of the test, then performance differences between them cannot be said to be exclusively due to the experimental group's learning of a target that was present in the training, and the validity of using the controls as a comparative baseline diminishes.

Rebuschat's (2008) aforementioned semi-artificial language study again provides an illustrative example. His inclusion of an untrained control group represented an improvement over using chance as the sole baseline. However, while the experimental participants were told to base their grammaticality judgments on what they had heard in the training phase, the untrained controls, who had not completed the training, could only be asked to perform a grammaticality judgment test on what may have seemed to be scrambled English sentences. Thus, while the experimental group may have known to reject anything too similar to standard English simply by virtue of having heard scrambled sentences earlier and regardless of what they had or had not learned during the training, the controls may have been inclined to answer straightforwardly according to their knowledge of standard English. Notably, some of the "ungrammatical" test sentences in that study actually corresponded to normal English sentences (e.g., *Some time ago John filled the bucket with apples*), and the untrained controls had extremely high endorsement rates for such sentences in three of the four relevant experiments (Experiments 3, 4, and 6). On these types of test items, it is therefore unclear whether the experimental group can truly be said to have acquired the targeted grammaticality knowledge, as Rebuschat also concluded at the time. An alternative explanation may be that different mind-sets led the groups to judge some sentence types, at least in part, according to their similarity to standard English, producing correct answers for the experimental group and incorrect answers for the control group.

A recent series of experiments by Rogers, Révész, and Rebuschat (2015) was able to avoid some of these problems. Their study also employed sentences with English lexis and scrambled word orders (e.g., *Last month the* ***kasu*** *opened Patrick with the key*); however, the target was not syntax, but rather a system of morphological markings on novel (Czech) vocabulary items that were used to replace certain English words. Given that the ALS modified English in multiple ways, and given the novelty of the words to which the morphemes were attached, it may be less likely that the experimental and control groups interpreted the instructions differently and had different expectations about the target morphemes. This does not mean that the groups could be assumed to have the same mind-sets in general, however; after all, by the time the experimental participants were tested, they had already been engaged in processing sentences with novel word orders, whereas for the controls who experienced only the test phase, the novelty of the syntax may have been salient, drawing attention away from morphology. Nonetheless, the instructions, at least, were similarly interpretable across groups.

Achieving similar mind-sets across groups may be an easier task in SLA research where all participants are engaged in learning the target language outside the experiment. In such studies, because all participants already construe themselves as learners of the target language, test instructions are likely to make sense regardless of whether participants have been trained in the experiment. However, even in those cases, interactions with affect and motivation might reduce the comparability of experimental and control groups (Sachs & Weger, 2011). For instance, if experimental participants are eager to demonstrate what they have learned while controls are confused, frustrated, or less motivated to exert effort on a test due to a perceived lack of helpful training, the validity of the comparison is diminished.

Situations such as these highlight that ALS researchers may need to take extra steps to improve the comparability, and therefore the validity, of control groups by devising creative ways of promoting similar expectations among all participants, while also designing linguistic targets, stimuli, instructions, and test sentences to avoid issues of the sort discussed in the preceding text. While the use of untrained controls represents an improvement over chance as a baseline, researchers should be cognizant of their limitations and consider the ways in which a trained control group may offer greater comparability.

## TRAINED CONTROL GROUPS

The use of trained controls has precedent in research using language-like stimuli devoid of semantics (e.g., Onnis, Waterfall, & Edelman, 2008) but is a more recent development in research employing ALSs (Hamrick, 2012, 2013, 2014a, 2014b; Rebuschat et al., 2013, 2015). Unlike instructed SLA research, where comparison groups are often exposed to the same linguistic input under different instructional conditions, experiments with ALSs expose experimental and trained control participants to training conditions, materials, instructions, and test items that are identical, except that relevant linguistic features have been pseudorandomized in a frequency-balanced way in the trained control group's training materials so as to eliminate the systematic syntactic patterns or form-meaning connections that constitute the target regularity for the experimental group. That is, trained controls are exposed to (a) nontarget aspects of the stimuli that replicate exactly what is seen by the experimental group, and (b) would-be target aspects of the stimuli that purposeful randomization has rendered not only uninformative (and therefore impossible to learn), but also protected from unintended additional similarities with real languages due to their lack of systematicity where the target is concerned. This avoids some of the limitations of untrained controls by allowing researchers to give all groups the same instructions as well as equivalent opportunities to learn nontarget regularities (or fabricate them) during the training phase.

Research employing AGs/ALSs has repeatedly demonstrated that it is not only possible but indeed common for experimental participants to contrive nontarget patterns or internalize other unintended but real (e.g., partial) patterns during training. The development of nontarget "knowledge" sometimes makes no difference in participants' test scores (e.g., Hamrick, 2013), but it can sometimes boost (e.g., Hamrick, 2014a; Knowlton & Squire, 1996) or harm their performance, such as when participants devise misguided rules of thumb that they reportedly use to override their intuitions (e.g., Rebuschat et al., 2015). To illustrate, consider Hamrick's (2013) experiments designed to investigate incidental learning of novel syntactic structures. Experimental participants were exposed to a semi-artificial language with English phrases placed into three structures derived from Persian (e.g., *Yesterday Charlie at the supermarket milk bought*; *Yesterday Charlie milk at the supermarket bought*; *Yesterday bought Charlie at the supermarket milk*), then given a surprise posttest. Although the experimental participants' level of performance at test suggested target learning, subsequent analysis of their postexperimental verbal reports indicated that many focused specifically on verb finality without learning other targeted aspects of the syntax. Because two-thirds of the training structures and grammatical test sentences happened to be verb-final, a tendency to endorse verb-final test items led to high levels of performance regardless of whether experimental participants had learned any other aspects of the system. Thus, Hamrick may have found some evidence of learning fragmentary, partial patterns, but not of learning the whole linguistic target as defined by the researcher.

Fortunately, by including a trained control group, Hamrick was able to discover an apparently independent inclination for participants to focus on verb finality during training irrespective of whether the regularity actually existed in the training materials. Trained control participants were exposed to sentences containing the same words, phrases, and meanings as those shown to the experimental participants, but the orders of the phrases were pseudorandomized in a frequency-balanced way to remove any phrase-order patterns. In other words, the trained controls saw all possible orders (including ungrammatical ones) an equal number of times, and verbs occurred in all positions with equal frequency. Despite these facts, postexperimental verbal reports indicated that trained controls were (overly) sensitive to instances of sentence-final verb placement, which may have been more salient due to a general human bias toward stimulus edges (Endress, Carden, Versace, & Hauser, 2010), combined with the novelty (for L1-English speakers) of occasionally seeing a verb at the end of the sentence. Whatever the reason for the attentional bias, the results indicated that preexisting proclivities can influence test response patterns in a way that suggests learning even when learning was not possible due to the uninformative (randomized) nature of the training materials. Here again, the use of an additional control group simultaneously illuminated and complicated interpretations of the experimental group's results. Such insights would have been missed if Hamrick had compared the experimental group's performance against only chance or untrained controls' performance.

The possibility that participants will react unexpectedly to features of training stimuli can involve semantics as well as grammar. In Rebuschat et al.'s (2013, 2015) replications of Williams's (2005) study on implicit learning of form-meaning (determiner-animacy) relationships, for instance, there was an unexpected tendency for both experimental and trained control participants to formulate hypotheses about the relevance of animacy in determiner use. Importantly, this was despite the fact that an animacy regularity was present only in the experimental group's training materials; for the trained control group, each determiner was used an equal number of times with animate and inanimate referents. If there was no underlying pattern that could have directed trained control participants' attention to animacy, then what prompted them to formulate such hypotheses? Concurrent and retrospective verbal reports indicated that participants in all groups had negative affective reactions to certain animate entities (e.g., rats, flies) in the training sentences. The inclusion of a trained control group along with triangulation from verbal reports allowed the researchers to discover that some of the experimental group's success may have been due *not* to gradually emerging associations between determiners and animacy, per se, but to the noticing of exemplars that struck participants in both groups as salient and led to extrapolations independently of the existence (or learning) of any pattern. Taken together, these findings should motivate researchers to ask whether their design will allow them to be certain that experimental participants have learned the target from training, or whether there is a possibility that participants exposed to similar stimuli minus the regularity might fabricate or assume it irrespective of actual learning. The fact that the latter has been shown to occur in studies with trained controls means that extra caution is warranted in claiming evidence of learning in studies without this design feature.

In highlighting some advantages of trained control group designs, it is important to point out that they too suffer from several limitations. One, to be discussed in greater depth in the recommendations section, is that without additional process measures (e.g., think-aloud data), it may be difficult to identify the source of systematic response patterns that arise due to exposure independently of (opportunities for) learning. Another is analogous to the mind-set-comparability issues presented in relation to untrained controls. Within the AG literature, it has been demonstrated that, compared to untrained controls, trained controls' performance is less affected by preexisting systematic biases (Reber & Perruchet, 2003); however, the very few ALS experiments in SLA that have included trained controls have found at least some systematicity in participants' responses following pseudorandomized training. Beyond biases of the sorts described in the preceding text, such patterns may also stem from a human tendency to seek coherence when faced with randomness. Indeed, some evidence suggests that exposure to unpatterned stimuli actually increases the likelihood that people will search for patterns and mistakenly report finding them (Whitson & Galinsky, 2009). As such, it is possible that exposing trained controls to pseudorandomized materials might trigger deeper, or more persistent, search processes than are seen in an experimental group. If randomness in the stimuli prompts trained controls to process the materials differently from experimental participants, then their value as a comparison group decreases.

Finally, researchers must bear in mind that scoring methods might unfairly penalize trained controls for learning from pseudorandomized training materials. In Rebuschat et al.'s (2013, 2015) experiments, for instance, we tested all participants on "memory" items that had appeared as exemplars in the experimental group's training as well as on novel "generalization" items. In the interest of one kind of direct comparability, we scored all items consistently across groups according to what the experimental group was supposed to have learned (i.e., the target regularity), not according to the pseudorandomized exemplars the trained controls had seen. Results indicated that the trained controls performed significantly below chance on the memory items, possibly reflecting accurate recall of pseudorandomized training exemplars. Consequently, the finding that experimental participants performed significantly above chance and better than the trained controls on those items cannot, in itself, be interpreted as evidence of learning. To have confidence that apparent evidence of learning by experimental participants is genuine, assessments should, where applicable, give credit for exemplar learning according to what each group has actually seen. If this is done and experimental participants are still found to perform differently from trained controls on memory items, then their different performance might suggest contributions of other knowledge besides memory for exemplars, perhaps complementing information gained from using generalization items to test rule-based knowledge. In any case, researchers must consider whether their particular combination of pseudorandomized training materials, test format, and scoring method will count trained control participants' answers as right or wrong for the appropriate reasons as far as internal validity is concerned.

## SOME RECOMMENDATIONS FOR ESTABLISHING EVIDENCE OF LEARNING

To this point, we have focused primarily on elucidating some of the tacit assumptions and challenges associated with measuring learning in experiments employing ALSs. Fortunately, there are steps that researchers can take to address some of these challenges and minimize the extent to which experiments are based on unwarranted assumptions. There is no one-size-fits-all solution, and space limitations prevent extensive recommendations here, but we can offer some brief suggestions regarding the benefits of triangulation and the potential for more sophisticated statistical approaches to illuminate trends in the data. Essentially, we recommend that researchers seek as much information as feasible through additional controls and measures, then report the whole complex picture without cherry-picking and interpret results cautiously with the limitations we have reviewed in mind.

It is not uncommon for SLA researchers to recommend triangulation as a means of addressing the limitations of individual methods used in isolation. In light of the problems we have outlined concerning different baselines for establishing evidence of learning, we would strongly recommend that, when possible, researchers include multiple control groups to take advantage of their different benefits and gain a richer understanding of the kinds of learning that are (or are not) occurring in their studies. Because resources are often scarce, however, it may make sense to prioritize the inclusion of a trained control group. In combination with verbal reports, trained controls' performance can often provide unanticipated and illuminating information about how participants (and their biases) interact with the training stimuli. In the event that trained controls' performance is difficult to interpret (e.g., due to mind-set-related confounds or scoring problems), researchers may benefit from adding an untrained control group to allow for comparisons unaffected by these issues—bearing in mind, however, that untrained controls are subject to logistical issues and mind-set differences of other sorts.

One might argue that if researchers carefully consider potential biases in advance, they can avoid certain confounds and do not need to use resource-consuming methods to check for them directly. Indeed, it may be possible to predict certain types of biases, such as well-established L1-based processing tendencies. However, a wide variety of other sources of bias exist, from basic perceptual tendencies to knowledge gained in linguistics and foreign language classes. All of these may interact with features of an ALS in unexpected ways, even when the same ALS is used in multiple studies. For instance, Williams's (2005) implicit learning study and a series of replications by Rebuschat et al. (2013, 2015) produced results that differed from replications by Hama and Leow (2010) and Faretta-Stutenberg and Morgan-Short (2011). Notably, the former set of studies included linguistics students as participants whereas the latter purposely excluded them—a non-L1-related difference that may help to explain the differing results. In verbal reports, many of Rebuschat et al.'s (2015) experimental and trained control participants demonstrated metalinguistic knowledge of phenomena such as semantic features and morphophonological alternations, which they mentioned had come to mind while processing the stimuli. Because concurrent verbalization (i.e., think-alouds) can be reactive (e.g., Rebuschat et al., 2015), it is crucial to include a silent control group as well when they are used. However, we have found the follow-up analyses they allow to be an invaluable source of insights well worth the additional time and effort. Without qualitative data on participants' approaches to the tasks and reactions to certain aspects of the stimuli and test items, we might have come to very different conclusions or found certain results to be uninterpretable.

In addition to the triangulation of research methods and data-collection instruments, there are some statistical procedures that researchers may find informative. To illustrate, imagine a scenario similar to Hamrick's (2014a) study, described earlier, where chance performance on a grammaticality judgment test is 50% and an experimental group performs significantly above chance (60%), as does a control group (57%), with no statistically significant difference between them. Limited to basic statistical analyses, a researcher might conclude that (a) no learning occurred in either group (because the controls were not exposed to the target regularity and therefore could not have learned it) or (b) the same pseudo-learning effect somehow occurred in both groups (perhaps during the test phase, or due to an unintentionally informative cue in the stimuli). However, there is a third possibility: namely, that the groups performed above chance for different reasons. Perhaps the experimental group did (partially) learn the target pattern, while the controls seized upon some other aspect of the stimuli that happened to be positively correlated with the target.

How can researchers explore these possibilities statistically? One option might be to use predictive multilevel models, which would allow researchers to analyze test results on a trial-by-trial basis. For instance, with Hamrick's (2014a) data, test item number could be used as a predictor to assess whether learning occurred during the test phase and, if so, whether the amount differed across groups. Analyzing further, a researcher could check for interactions between potentially relevant test-item features (e.g., verb finality, grammaticality) and group membership, to determine whether there were different bases for the experimental and control groups' performances (Baayen, Davidson, & Bates, 2008). Importantly, though, because this would require the researcher to specify covariates of potential or known interest beforehand, it might be unhelpful in situations in which the problem is precisely that unexpected and unknown factors might be driving the groups' performances. While such a statistical procedure offers potential solutions in principle, and while its use would be consistent with broader methodological reforms taking place in applied linguistics (Cunnings, 2012), to our knowledge no one has utilized it for these purposes in SLA research employing ALSs. As such, further elaboration of these possibilities and specifications of appropriate procedures will be necessary to move the field forward along these lines.

## CONCLUSION

We have reviewed some of the assumptions, limitations, and challenges associated with three commonly used baselines against which learning is measured in research employing ALSs. We have argued that relying exclusively on chance as a comparison entails unwarranted assumptions that threaten a study's validity by not acknowledging, for instance, the potential for participants to show preexisting biases or to learn during the test phase. Untrained controls provide a substantial improvement over chance as a baseline, but it is paramount to take steps to ensure that experimental and control participants have similar expectations, mind-sets, and interpretations of the test instructions. Trained controls have been used for this purpose and have the added benefit of allowing for similar kinds of nontarget learning to occur during the exposure phase. However, simply including such groups may not be sufficient; additional triangulation can contribute vitally to researchers' abilities to make valid and insightful claims about learning, and advanced statistical procedures, while requiring further elaboration, hold promise for enabling further progress in this area. Meanwhile, by carefully weighing the assumptions behind different types of baselines and controls, researchers can make more informed methodological choices and draw more appropriate conclusions about whether they have truly found evidence of the development of target knowledge in SLA experiments employing ALSs.

*Received 30 November 2016*
*Accepted 1 December 2016*

### NOTES

[1] Within the context of *ab-initio* language learning experiments, simplified versions of natural languages may be functionally equivalent to ALSs in important respects. From the perspective of participants who have never been exposed to the language before, mini-Latin (Stafford, Bowden, & Sanz, 2012) might as well be eXperanto (de Graaff, 1997),

for example. While it is relevant to ask how complex and natural any mini-language/ALS is, and while participants' motivations about learning "real" versus "fake" languages may influence their response tendencies, the internal-validity concerns we raise must be addressed regardless of whether a richer version of the linguistic system exists in the real world. We thank an anonymous reviewer for the question.

2. Participants might also respond systematically using pedagogical rules from foreign language classes that seem potentially relevant.

## REFERENCES

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, *28*, 369–382.

de Graaff, R. (1997). The eXperanto experiment. *Studies in Second Language Acquisition*, *19*, 249–276.

Dienes, Z., & Altmann, G. T. A. (2003). Measuring learning using an untrained control group: Comment on R. Reber and P. Perruchet. *The Quarterly Journal of Experimental Psychology*, *56*, 117–123.

Endress, A. D., Carden, S., Versace, E., & Hauser, M. D. (2010). The apes' edge: Positional learning in chimpanzees and humans. *Animal Cognition*, *13*, 483–495.

Ettlinger, M., Morgan-Short, K., Faretta-Stutenberg, M., & Wong, P. C. M. (2015). The relationship between artificial and second language learning. *Cognitive Science*, *40*, 822–847.

Faretta-Stutenberg, M., & Morgan-Short, K. (2011). Learning without awareness reconsidered: A replication of Williams (2005). In G. Granena, J. Koeth, S. Lee-Ellis, A. Lukyanchenko, G. Prieto Botana, & E. Rhoades (Eds.), *Selected proceedings of the 2010 Second Language Research Forum: Reconsidering SLA research, dimensions, and directions* (pp. 18–28). Somerville, MA: Cascadilla Proceedings Project.

Field, A. (2009). *Discovering statistics using SPSS*. London, UK: Sage.

Grey, S., Williams, J. N., & Rebuschat, P. (2014). Incidental exposure and L3 learning of morphosyntax. *Studies in Second Language Acquisition*, *36*, 611–645.

Hama, M., & Leow, R. (2010). Learning without awareness revisited: Extending Williams (2005). *Studies in Second Language Acquisition*, *32*, 465–491.

Hamrick, P. (2012). *Associative chunk learning supports early phases of adult L2 syntactic development.* Poster presented at the Second Language Research Forum, Pittsburgh, PA.

Hamrick, P. (2013). *Development of conscious knowledge during early incidental learning of L2 syntax* (Unpublished doctoral dissertation). Georgetown University, Washington, DC. Retrieved from ProQuest Dissertations and Theses database (3558525).

Hamrick, P. (2014a). A role for chunk formation in statistical learning of second language syntax. *Language Learning*, *64*, 247–278.

Hamrick, P. (2014b). Recognition memory for novel syntactic structures. *Canadian Journal of Experimental Psychology*, *68*, 2–7.

Hamrick, P. (2015). Declarative and procedural memory abilities as individual differences in incidental language learning. *Learning and Individual Differences*, *44*, 9–15.

Hamrick, P., & Rebuschat, P. (2012). How implicit is statistical learning? In P. Rebuschat & J. N. Williams (Eds.), *Statistical learning and language acquisition*. Berlin: Mouton de Gruyter.

Hamrick, P., & Rebuschat, P. (2014). Frequency effects, learning conditions, and the development of implicit and explicit lexical knowledge. In J. Connor-Linton & L. Amoroso (Eds.), *Measured language: Quantitative approaches to acquisition, assessment, processing and variation* (pp. 125–139). Washington, DC: Georgetown University Press.

Knowlton, B. J., & Squire, L. R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 169–181.

MacWhinney, B., Bates, E., & Kliegl, R. (1984). Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning and Verbal Behavior*, *23*, 127–150.

Morgan-Short, K., Steinhauer, K., Sanz, C., & Ullman, M. T. (2012). Explicit and implicit second language training differentially affect the achievement of native-like brain activation patterns. *Journal of Cognitive Neuroscience*, *24*, 933–947.

Onnis, L., & Thiessen, E. (2013). Language experience changes subsequent learning. *Cognition*, *126*, 268–284.

Onnis, L., Waterfall, H., & Edelman, S. (2008). Learn locally, act globally: Learning language with variation set cues. *Cognition*, *109*, 423–430.

Perruchet, P., & Reber, R. (2003). Why untrained control groups provide invalid baselines: A reply to Dienes and Altmann. *The Quarterly Journal of Experimental Psychology*, *56*, 125–130.

Plonsky, L. (2015). *Demographics in SLA: A systematic review of sampling practices in L2 research*. Paper presented at the Second Language Research Forum, Atlanta, GA.

Reber, A. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*, 855–863.

Reber, R., & Perruchet, P. (2003). The use of control groups in artificial grammar learning. *The Quarterly Journal of Experimental Psychology*, *56*, 97–115.

Rebuschat, P. (2008). *Implicit learning of natural language syntax* (Unpublished doctoral dissertation). University of Cambridge, Cambridge, UK.

Rebuschat, P., & Williams, J. N. (2012). Implicit and explicit knowledge in second language acquisition. *Applied Psycholinguistics*, *33*, 829–856.

Rebuschat, P., Hamrick, P., Riestenberg, K., Sachs, R., & Ziegler, N. (2015). Triangulating measures of awareness: A contribution to the debate on learning without awareness. *Studies in Second Language Acquisition*, *37*, 299–334.

Rebuschat, P., Hamrick, P., Sachs, R., Riestenberg, K., & Ziegler, N. (2013). Implicit and explicit knowledge of form-meaning connections: Evidence from subjective measures of awareness. In J. M. Bergsleithner, S. N. Frota, & J. K. Yoshioka (Eds.), *Noticing and second language acquisition: Studies in honor of Richard Schmidt* (pp. 255–275). Honolulu: University of Hawai'i at Manoa, NFLRC.

Redington, M., & Chater, N. (1996). Transfer in artificial grammar learning: A reevaluation. *Journal of Experimental Psychology: General*, *125*, 123–138.

Rogers, J., Révész, A., & Rebuschat, P. (2015). Challenges in implicit learning research: Validating a novel artificial language. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 275–300). Philadelphia, PA: John Benjamins.

Sachs, R., & Weger, H. D. (2011). *Motivation and effort in the comparison group: A factor in studies of L2 feedback?* Paper presented at the Second Language Research Forum, Ames, IA.

Stafford, C. A., Bowden, H. W., & Sanz, C. (2012). Optimizing language instruction: Matters of explicitness, practice, and cue learning. *Language Learning*, *62*, 741–768.

Tagarelli, K. M., Borges Mota, M., & Rebuschat, P. (2015). Working memory, learning context, and the acquisition of L2 syntax. In W. Zhisheng, M. Borges Mota, & A. McNeill (Eds.), *Working memory in second language acquisition and processing: Theory, research, and commentary*. Bristol, UK: Multilingual Matters.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632–1634.

Whitson, J. A., & Galinsky, A. D. (2009). Lacking control increases illusory pattern perception. *Science*, *322*, 115–117.

Williams, J. N. (2005). Learning without awareness. *Studies in Second Language Acquisition*, *27*, 269–304.

Williams, J. N., & Kuribara, C. (2008). Comparing a nativist and emergentist approach to the initial stage of SLA: An investigation of Japanese scrambling. *Lingua*, *118*, 522–553.