Studies in Second Language Acquisition (2023), **45**, 1481–1505
doi:10.1017/S0272263123000165

METHODS FORUM

Reevaluating trials to criterion as a measure in second language research

Nick Henry

University of Texas at Austin–Germanic Studies, Austin, TX, United States
Email: nhenry@austin.utexas.edu

(Received 15 September 2022; Revised 23 February 2023; Accepted 03 March 2023)

Abstract

Research on input processing and processing instruction has often employed a scoring method known as trials to criterion to observe the effects of instruction that emerge during training. Despite its common use in this research (see Fernández, 2021) this metric has never been evaluated critically. The present study first discusses several challenges associated with trials to criterion, including issues with its conceptual and methodological implementation. The study then introduces three alternative approaches for analyzing accuracy data collected during training sequences: trials to accuracy threshold, growth curve analysis, and bootstrapped differences of timeseries. For each approach, advantages and disadvantages are discussed and example analyses are presented using data from previous research. This discussion shows how these alternative approaches can supplement current trials-to-criterion-based analyses, expand the methodological choices available to researchers, and permit new and interesting research questions.

Introduction

Within research on processing instruction (PI), many studies have varied aspects of the training to investigate how the components or sequencing of PI affect its outcomes. Much of this work has manipulated the presence or absence of explicit information (EI) to assess whether it is necessary or beneficial in PI trainings (Fernández, 2008; Henry et al., 2009; VanPatten et al., 2013; VanPatten & Oikkenon, 1996), but research has also manipulated other aspects of training, for example, the sequencing of EI (Diaz, 2017), the role and type of feedback (Sanz & Morgan-Short, 2004), or the type of aural stimuli used in training (Henry, Jackson, et al., 2017). Research in this area is interested in not only whether a manipulation results in similar learning outcomes but also whether learners begin processing sentences correctly (i.e., use the targeted cues) at similar points during the training—that is, whether certain manipulations lend an advantage in terms of the speed with which learners begin to process sentences correctly. For this reason, Fernández (2008) introduced the scoring metric *trials to* *criterion* (TTC), which can be used to compare similar training units. This metric has been adopted by a large number of studies since its inception (e.g., Culman et al., 2009; Henry et al., 2009, 2017; VanPatten et al., 2013; Villegas & Morgan-short, 2019), but despite its common use in this research, this metric has never been evaluated critically. Therefore, the purpose of the present study is to review practices and procedures for the TTC scoring method, discuss theoretical and methodological challenges related to TTC, and to introduce three alternative methods that attempt to avoid them. To understand TTC and how it has been applied in the literature, it is first necessary to discuss the construction of processing instruction trainings to which the TTC scoring method is most typically applied.

Background

Processing instruction and structured input

Processing instruction is the pedagogical application of VanPatten's (2004, 2015) input processing model, which consists of several principles that describe learners' processing strategies—that is, how they use linguistic forms to interpret sentences and determine overall sentence meaning. For example, the *first noun principle* states that learners typically rely on word order to assign thematic roles and thus interpret the first noun in the sentence as the agent or subject of that sentence. As a consequence, case markers (e.g., clitic pronouns in Spanish or case-marked articles in German) are often left unprocessed and are difficult to acquire. Processing instruction therefore seeks to reorient learners' processing strategies so that they actively attend to such forms during sentence comprehension.

Practically, PI attempts to reorient learners' processing strategies through *structured input* (SI) activities. Although PI employs two types of structured input activities, most studies that use TTC as a scoring method use forced-choice tasks known as referential activities. In these activities, input is manipulated so that learners must use the target form to interpret a sentence. For example, in Fernández (2008), participants heard a sequence of target object–verb–subject (OVS) sentences intermixed with distractor subject–verb–object (SVO) sentences, as indicated in (1) and (2):

(1) Lo llaman sus padres por teléfono. (OVS)
Him-$_{\text{OBJ}}$ call his parents-$_{\text{SUB}}$ by telephone.
"His parents call him"

(2) El niño llama a sus padres por teléfono. (SVO)
The boy$_{\text{-SUB}}$ calls OBJ- his parents by telephone.
"The boy calls his parents"

After each sentence, participants saw two pictures—for example, a boy calling his parents, or parents calling their son—and selected which picture depicted the sentence they heard. They then received one-word feedback ("Correct!" or "Incorrect") about the accuracy of their response.

Because the target and distractor items are presented in a sequence and intermixed, and because the sentences are tightly controlled for other potential cues, only the target form (in this case, the object markers *Lo/s*, *la/s*, and *a*) is reliable. Thus, participants cannot rely on word order, context, phonological information, or real-world knowledge to determine who does what to whom. Further, because the task requires correct interpretation of that cue, participants' responses are an indication of whether they were applying the correct processing strategy. That is, in this example, because learners typically interpret the first noun of the sentence as the subject, a correct answer to an OVS sentence should indicate that the participant applied the correct processing strategy and accurately interpreted the clitic pronoun as an object marker (see Wong, forthcoming).

Trials to criterion and Fernández (2008)

The trials-to-criterion (TTC) method was first introduced by Fernández (2008) to track learners' processing strategies during training and to measure whether benefits of PI were observable during training itself. This is particularly important because traditional pretest–posttest designs can hide differences between similar trainings if, for example, early advantages of one training paradigm are washed out by extensive practice. As Fernández (2008) explained with reference to prior studies on the role of EI research (Farley, 2004; Sanz & Morgan-Short, 2004; VanPatten & Oikkenon, 1996) in PI, "It may be possible, however, that the effects of EI were hidden due to the offline treatments and the pretest and posttest designs used in these PI studies and that learners actually may have benefited from EI at some point in the instruction. In order to observe the possible role of EI in PI, it is necessary to conduct an online study that tracks learners' behavior while they are engaged in activities designed to promote acquisition" (p. 278). Therefore TTC emerged as a way to extract meaningful data from the learner responses to the training component of a study (i.e., the forced-choice SI task).

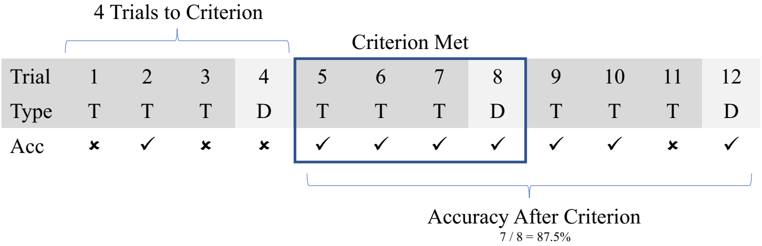
Fernández's (2008) landmark study investigated the effects of EI on the processing and acquisition of clitic object pronouns (Experiment 1) and the subjunctive of doubt (Experiment 2) in Spanish. Two groups of participants completed structured input presented either with or without EI (+/–EI). Participants in both groups completed a 30-item computerized PI training like the example for OVS sentences given above. During training, their accuracy was recorded by the computer so that Fernández could track performance during the task.

To determine the point at which participants began processing sentences accurately, Fernández (2008) needed to establish a criterion that suggested a change in learners' processing behaviors. As Fernández (2021) later explained, "the level of performance was arbitrarily set to correctly processing three target items plus one distractor in a row, which was considered the minimally convincing evidence of learners having achieved appropriate strategies for processing both target items and distractors" (p. 251). For each group she measured the number of participants who met criterion. Then, for each participant, she computed two scores: (1) *trials to criterion*, and (2) *accuracy after criterion* (AAC). TTC is the number of items that a participant heard prior to reaching criterion. As seen in Figure 1, for example, if a participant answered items 5–8 correctly, they received a score of 4 indicating that they saw four items before beginning to process sentences correctly. AAC is measured as a participants' overall accuracy on the items after they reached criterion. In this example, AAC equals the accuracy on items 5–12.[1]

In Experiment 1, which focused on OVS sentences, Fernández (2008) found that there were no differences between the +EI and –EI groups in terms of the number of participants who met criterion, the groups' TTC scores, or their AAC scores. In

[1] If a participant answers the first four items correctly, AAC reflects their total accuracy on all items. If a participant reaches criterion on the final four items of the training set, their AAC score is 100%.

https://doi.org/10.1017/S0272263123000165 Published online by Cambridge University Press



Trial	1	2	3	4	5	6	7	8	9	10	11	12
Type	T	T	T	D	T	T	T	D	T	T	T	D
Acc	✗	✓	✗	✗	✓	✓	✓	✓	✓	✓	✗	✓

Figure 1. Example of TTC and AAC on a hypothetical 12-item training set.

Experiment 2, which focused on the subjunctive, however, she found an advantage for the +EI group in all three measures. Fernández speculated that the difference between the experiments could be related to differences between the training tasks or in the processing strategies implicated for the two forms.

Beyond Fernández (2008): The design of TTC studies

During the last decade, numerous PI studies have used training paradigms similar to those used in Fernández (2008) and followed her methods and rationale for using TTC and AAC. In the following sections, I present a synopsis of the design differences in these studies. A full review of this research is beyond the scope of this paper, as the focus here is on the methodology used in these studies, but Fernández (2021) and Lee (2015) both provide excellent reviews of this research.

Items and distribution of target items in the training sequence. Trainings have differed with respect to the size of the training set. Although some studies—primarily replications of Fernández (2008)—include 30-item trainings (Culman et al., 2009; Henry et al., 2009, Villegas & Morgan-Short 2019), most have followed VanPatten et al. (2013), who used a longer training. Indeed, several of these studies—Henry et al. (2017), Henry (2021), and Henry (2022)—used the same 50-item set as VanPatten et al. (2013).[2] Glimois (2019) used a 48-item training set and Lee & Doherty (2020) used a 60-item set.

Contrary to Fernández (2008), who placed distractors after every 2–3 target items, most studies have used a rigid training sequence that consisted of fixed item sequences, typically a repeating T-T-T-D sequence. This convention was first introduced by Henry et al. (2009) and has been followed in most other studies (e.g., Glimois, 2019; VanPatten et al., 2013), with Villegas & Morgan-Short (2019)—a replication of Fernández (2008)[3]—and Lee & Doherty (2020) as exceptions. This means that the vast majority of studies have had a target-sentence distribution of around 75% (i.e., three targets and one distractor in each 4-item sequence in the training).

TTC scoring procedures. In addition to the standard item sequence, Henry et al. (2009) introduced a now-standard TTC scoring convention that deviated from Fernández (2008). When computing group averages for TTC, Fernández excluded participants who did not meet criterion. Rather than exclude these participants, Henry et al. (2009)

[2] In each of these studies, the training was exactly the same as in VanPatten et al.'s (2013) study, except that the audio had been rerecorded.

[3] Villegas & Morgan-Short used the same training as Fernández, except that audio samples were rerecorded.

gave them a score of 30, the number of items in the training set, reasoning that these participants would have met criterion had the training been extended. This convention has been followed in the majority of studies that succeeded it: Glimois (2019), Henry et al. (2017), Henry (2021), Henry (2022),[4] and VanPatten et al. (2013). Again, Villegas & Morgan-Short (2019) followed Fernández's procedure when replicating this study.[5]

AAC reporting and scoring. Fernández (2008) included AAC as a secondary measure to TTC, which ensures that TTC has not overestimated learners' abilities. Most studies have also included AAC (Henry et al., 2017; Henry 2021, 2022); however, as Henry et al. (2009) notes, AAC is not a meaningful metric when high numbers of participants fail to reach criterion, as many participants are then excluded from the group-level AAC calculation. In such situations, the number of participants meeting criterion is more meaningful, and some studies have therefore decided simply to not report AAC scores (Henry, 2009; VanPatten et al., 2013). Glimois (2019) included AAC but scored this measure differently: whereas most studies include the four-item criterion sequence as indicated in Figure 1, Glimois did not (i.e., most studies would include items 5–12 in Figure 1, but Glimois would only include items 9–12).

Alternative approaches. Three additional studies warrant special mention because of their similarity to the research cited above even though they did not use either TTC or AAC to compare groups. The first of these was actually the first to attempt a replication of Fernández (2008): Culman et al. (2009). This study focused on the role of EI for the acquisition of accusative case markers. Like Fernández, Culman et al. intermixed SVO and OVS sentences; however, because the accusative case is marked ambiguously in two of German's three genders, the training sequence contained three sentence types: OVS targets, SVO distractors, and ambiguous sentences. As a result, target sentences were spaced too wide apart to use TTC (and AAC) effectively. Thus, as an alternative, they probed the participants' accuracy rates at four points in the training (items 1–3, 8–10, 15–17, and 28–30). This essentially provided a snapshot of learner performance as the training progressed.

Lee (2014) used TTC to measure learning rate as an individual difference (not a dependent variable). TTC was defined as "the number of the practice item on which the learners began to answer three items in a row correctly" (p. 154). Notably, this definition differs from the traditional use of TTC in that it did not include the stipulation that learners process both target and distractor sentences correctly. Thus, this definition may not have been appropriate to establish whether learners used the target processing strategies or simply applied a different strategy—for example, a second-noun strategy (see Fernández, 2021). Lee & Doherty (2020) used the same TTC criterion and called this an "accuracy trend." In this study, they looked only at the target sentences and calculated the number of accuracy trends, the mean length of trends, and maximum length of trends.

Challenges in TTC research

Challenges in rationale and interpretation. A review of the studies using TTC points to several challenges associated with the measure. The first set of these challenges are related to the interpretability of the score and what it purports to measure.

[4] TTC is not reported in Henry (2022) but is reported in Henry (2015), the dissertation from which this work stems.

[5] Villegas & Morgan-Short (2019) do not actually report TTC and AAC metrics in the published conference proceedings, but they did report them in the conference presentation.

First and foremost, the interpretation of the score varies across studies. For instance, Henry et al. (2009) and VanPatten et al. (2013) state that TTC indicates when participants begin to process input correctly. However, Fernández (2021) claims that TTC measures when learners had "*achieved appropriate strategies* for processing both target items and distractors" (p. 251, emphasis added). Similarly, Lee (2014) states that TTC indicates "when learners *begin and then maintain* the correct processing strategy" (p. 150, emphasis added). Thus, although some researchers take TTC to indicate the first instance of appropriate processing, others assume that participants will continue to process sentences appropriately. However, both of these interpretations are potentially problematic because TTC scores are susceptible to chance and do not account for accuracy after criterion is met.

Consider first the role of chance in TTC scores. The vast majority of tasks used in this research are forced-choice tasks with only two answer choices. Using the standard definition of criterion (four correct answers in a row), participants have a 6.25% (0.5^4) chance of gaining criterion on any given item. Although participants do not typically guess at random throughout training,[6] they may be more likely to guess when they find that their usual processing strategies are incorrect, when they are confronted with unfamiliar lexical items, or if they do not hear target sentences clearly in aural tasks. Consequently, TTC scores may not reliably estimate a given learner's behavior over the course of the training.

Second, TTC itself does not include any information about the participants' processing strategies after criterion is met. In theory, AAC controls for the possibility that TTC scores are inaccurate: if a participant has a very low TTC score but a very low AAC score, researchers can see that they met criterion very quickly but did not maintain the correct processing strategy. However, AAC is currently only used as a control on group-level data and individual-level AAC scores are not considered. As individual differences research becomes more important within PI research (Villegas & Morgan-Short, 2019; see Henry, forthcoming, for discussion), it is important that this sort of control be applied at the subject level as well so that TTC scores can be confidently used as an individual difference measure that captures both initial use of the processing strategy and its maintenance over time (as is claimed). Unfortunately, AAC seems ill-suited as a control on individual TTC scores because there is no obvious way to adjust TTC scores based on AAC (as TTC and AAC, after all, rely on the same criterion). Further, AAC cannot be used for participants who do not reach criterion at all and thus cannot be used meaningfully in studies like Henry et al. (2009) who found a significant number of participants who did not reach criterion in one group.

Challenges to methodology. In addition to the conceptual issues outlined above, there are multiple methodological challenges that arise in PI studies that use TTC,[7] primarily because of its rigid design requirements.

Sequencing. As Fernández (2021) notes, it is important to use both target *and* distractor items during PI because together they show not only whether learners have

[6] This is evidenced by the high number of participants who never reach criterion in some groups (e.g., in Fernández, 2008; Experiment 1, only about 56% of participants reach criterion).

[7] As an anonymous reviewer pointed out, the reliability of teaching materials and their degree of discrimination should be considered if PI trainings are to be the target of analysis. Although this is an important issue and echoes broader conversations about the validity and reliability of research instruments in SLA (see Chapelle, 2020), I set it aside for now, as the focus in this section is on methodological challenges that arise specifically from the use of TTC.

https://doi.org/10.1017/S0272263123000165 Published online by Cambridge University Press

changed processing strategies but also whether they are able to use the target form to distinguish meaning. As a result, researchers using TTC must pay careful attention to the sequencing of target and distractor items so that both appear in any possible criterion sequence. For example, in Fernández's (2008) original study, criterion was reached when participants processed four items correctly in a row. Thus, she placed distractor items after every two, three, or four target sentences, resulting in sequences such as the following (T= target, D = distractor):

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
T	T	D	T	T	D	T	T	T	D	T	T	D	T	T	T	D

Contrary to Fernández's study, however, most studies (Culman et al., 2009; Henry et al., 2009; Henry, Jackson, et al., 2017; VanPatten et al., 2013) have used repeating T-T-T-D sequences.

In practice, each of these sequencing decisions comes with its own downsides. In studies using a repeating sequence, it is possible that learners cue into the sequence rather than the target form.[8] On the other, hand, although designs like Fernández's (2008) study are less predictable, some four-item sequences contain three targets and some only contain two with two distractors. Thus, in a sense, the definition of criterion shifts throughout the training, and TTC may represent more or less robust application of the appropriate processing strategy depending on when criterion is reached (i.e., in a sequence with two or three targets).

Distribution of items. The sequencing of items also results in rigid requirements about the distribution of target and distractor items. Because TTC studies typically use a repeating T-T-T-D sequence, most studies contain 75% target items. Although this ratio of items may be appropriate for training, this requirement prevents researchers from using a target-distractor ratio that is necessary to meet other research parameters (e.g., if a study's research questions require training input to more closely represent the distribution of natural language). More importantly, this requirement precludes researchers from investigating the effects of the distribution itself.

Randomization. Similarly, the rigid sequencing of items has contributed, at least partly, to the lack of randomization of training items. Although randomization is not always considered necessary for instructional trainings, it is common, if not typical in lab-based SLA research, and it would be advantageous in many PI studies given the need to control stimuli for biasing information. For instance, in word order studies, randomization can guard against plausibility or probability biases for specific items. It can also guard against biases in pronunciation or prosody that might be difficult to control or are not readily apparent to researchers. Further, it can prevent participants from learning a repeated sequence and guard against effects that arise from using any particular sequence. TTC studies, however, have not broadly implemented any randomization, even though this is technically feasible in computer-administrated trainings (e.g., using nested experiment structures in E-prime where the type of item—target or distractor—is specified as a sequence but actual items are randomized).

[8] To guard against this possibility, studies by Henry et al. (2009) and Henry et al. (2017) have asked participants what strategies they used to do the task and screened out the few participants who have reported an awareness of this sequence.

The present study

As described in the previous section, there are multiple challenges to the TTC scoring measure, which relate to the restrictive methods required and the interpretability of the data. Alternative approaches to TTC should therefore seek to overcome these issues. First, in terms of the interpretability of data, alternative measures should minimize the role of chance while clearly defining whether the maintenance of a processing strategy is required. Second, in terms of methodology, alternative measures should allow flexibility of item sequences during training, the randomization of sequences, and the distribution of target and distractor items.

The primary purpose of the present study is thus to present three alternative approaches to the TTC scoring method that provide clearly interpretable results and allow for flexibility in study design as operationalized here: (1) trials to accuracy threshold (TTAT), (2) growth-curve analysis (Mirman, 2014), and (3) bootstrapped differences of time series (BDOTS; Seedorff et al., 2018). TTAT is an alternative scoring method, which is conceptually similar to TTC and produces a single score based on participants' accuracy data during training. GCA and BDOTS, on the other hand, are two statistical methods that may be used to analyze time-series data and model accuracy rates over the course of the training. In each section that follows, I provide an overview of the approach, an illustrative data set, and a summary that highlights advantages and disadvantages of the approaches.

Before proceeding it should be noted that these are not the only approaches that researchers may find useful. In particular, researchers may consider other approaches to statistical modeling such as generalized additive mixed models (GAMM), cluster-based permutation analysis (CBA), or divergent point analysis (DPA; see Ito & Knoeferle, 2022). However, as will be discussed in the remainder of this paper, the approaches selected here provide useful avenues to overcome methodological challenges and add to current analyses.

Trials to accuracy threshold

Overview of the approach

As discussed above, TTC's use of sequence-based definitions for criterion (e.g., four items in a row) requires rigid sequencing and a set distribution of target and distractor items, and it limits the number of sentence types that can be used (as in Culman et al., 2009). Each of these challenges can be alleviated if a percentage-based criterion score is used. Specifically, I propose here a method for calculating trials to criterion, trials to accuracy threshold (TTAT), which reflects the point after which participants reach a predefined percentage-based accuracy threshold. In some respects, this method provides a way to combine standard TTC and AAC, which not only offers methodological advantages but also addresses the conceptual issues discussed previously.

The method for calculating trials to criterion with a TTAT score involves several steps and can be completed using the supplemental materials provided at https://doi.org/10.18738/T8/T9SFAG. First, a participant's responses are arranged in order and scored for their accuracy. For each item, the overall accuracy percentage for *the response and all of the subsequent responses* is then evaluated for each item. This is referred to here as the %After score. Take for example the 12-item response set seen in Table 1. At Item 1, the %After score represents the participant's accuracy on Items 1–12. In this case, the participant scored 8/12 = 66%. At Item 2, as shown by the box, it represents their accuracy on Items 2–12 (i.e., Item 1 is discarded from the accuracy

Table 1. Illustration of trials to accuracy threshold (TTAT)

Trial	1	2	3	4	5	6	7	8	9	10	11	12
Type	T	T	T	D	T	T	T	D	T	T	T	D
Accuracy	×	✓	×	×	✓	✓	✓	✓	✓	✓	×	✓
%After-1	66											
%After-2		72										
%After-3			70									
%After-4				77								

Table 2. Illustration of trials to accuracy threshold by condition (TTATxc)

Trial	1	2	3	4	5	6	7	8	9	10	11	12
Type	T	T	T	D	T	T	T	D	T	T	T	D
Accuracy	×	✓	×	×	✓	✓	✓	✓	✓	✓	×	✓
%AfterT	67	75	71	83	83	80	75	67	67	50	66	---
%AfterD	67	67	67	67	100	100	100	100	100	100	100	100

Note. %AfterT = the %After score for targets. %AfterD = the %After score for distractors. The %AfterT score for Item 12 is unavailable, as Item 12 is the last item in the series and is a distractor (i.e., there are no targets remaining in the series).

score). Here the participants scored 8/11 = 72%. At Item 3, it is accuracy on Items 3–12 (7/10 = 70%), at Item 4, it is accuracy on Items 4–12 (7/9 = 77%), and so on. For this example, criterion is defined as meeting a 75% accuracy threshold, which the participant first exceeds at Item 4. Thus, their TTAT score is 3.

In using this method, it is important to consider the appropriate accuracy threshold. I suggest that for most processing instruction studies, this will be equal to the distribution of target items in the training sequence (e.g., 75% as in the preceding example), lending reasonable confidence that the learner is accurate both on target *and* on distractor items.

However, this method can also be extended so that accuracy across conditions is computed separately in each condition. This provides greater confidence that participants are accurate with multiple sentence types and lends flexibility both in scoring and in study design. For example, researchers may want to ensure that participants maintain a high accuracy rate on distractor items while allowing a comparatively lower percentage on (the more difficult) target items. Similarly, researchers may want to include several different sentence types. This is illustrated in Table 2, which represents the same set of hypothetical data as in Table 1. For this example, criterion is defined as 75% accuracy on both target and distractor items. As seen, although the participant meets the 75% accuracy threshold for targets at Items 2 and 4, they do not meet the threshold for distractors until Item 5. Thus, their trials to criterion score would be 4.

A comparison with standard TTC scores

To investigate how TTAT compares with the standard sequence-based TTC scoring method, I collected data from nine of the studies (eight published papers and one unpublished dissertation) reviewed previously.[9] Together, these studies represented a

[9] Only two of the learner groups from VanPatten et al. (2013) were available. As these data were also reported in VanPatten & Borst (2012a, 2012b), these citations are used for clarity.

https://doi.org/10.1017/S0272263123000165 Published online by Cambridge University Press

Table 3. Reported and rescored TTC and AAC values for nine PI studies

Study	Group	Reported TTC	Rescored TTC	Reported AAC	Rescored AAC
Culman et al. (2009)	1st Sem., PI	X	X	X	X
	1st Sem., SI	X	X	X	X
	3rd Sem., PI	X	X	X	X
	3rd Sem., SI	X	X	X	X
Henry et al. (2009)	PI	12.47	11.37	X	0.77
	SI	22.05	17.95	X	0.64
VanPatten & Borst (2012a)	PI	5.25	5.25	X	0.76
	SI	23.96	23.96	X	0.76
VanPatten & Borst (2012b)	PI	18.35	18.35	X	0.75
	SI	16.63	16.63	X	0.71
Henry et al. (2017)	PI	4.60	5.43	0.82	0.82
	SI	11.10	13.10	0.74	0.75
	PI+P	4.90	4.59	0.80	0.80
	SI+P	15.10	16.48	0.74	0.75
Glimois (2019)	IFPI-Mono	0.44	0.44	0.98	0.98
	IFPI-Bi	3.18	3.18	0.93	0.93
	IFSI-Mono	15.55	15.55	0.68	0.76
	IFSI-Bi	12.42	12.42	0.77	0.81
	VOCPI-Mono	0.22	0.22	0.98	0.99
	VOCPI-Bi	0.50	0.50	0.98	0.98
	VOCSI-Mono	14.26	14.26	0.76	0.84
	VOCSI-Bi	10.74	10.74	0.72	0.81
Villegas & Morgan-Short (2019)	PI	X	10.95	0.77	0.78
	SI	X	16.46	0.59	0.61
	Control+	X	0.57	X	0.94
Henry (2021)	Blocking	12.57	12.57	0.76	0.78
Henry (2022)	PI	2.46	2.46	0.75	0.75
	PI+P	7.89	7.89	0.72	0.72

total of 28 distinct learner groups (see Table 3). Prior to comparison with the TTAT method, the data were rescored for both TTC and AAC to account for (a) incomplete data sets acquired from these studies,[10] (b) errors in the reported scoring,[11] (c) inconsistencies in scoring between studies,[12] and (d) inconsistencies in reported data.[13] The reported and rescored values for each study are found in Table 3 (note that some small discrepancies in the table below arise due to differences in rounding between the reported and rescored data).

Data were then scored for both TTAT (following the example in Table 1) and TTATxC (following the example in Table 2). Average scores for each group in the data set are found in Table 4. Note that the percentage-based criterion also enables a score to be generated for Culman et al. (2009), where this was impossible with TTC.

[10] Data from Henry et al. (2017) were missing for one participant.

[11] Upon review, data from Henry et al., (2009) had a scoring error that affected several participants' TTC scores, and Henry (2021) had an error that affected two participants' AAC scores.

[12] As discussed earlier, Fernández (2008) and Henry et al., (2009) introduced two methods of scoring TTC. Scoring here follows Henry et al., (2009). AAC scores for Glimois (2019) were rescored using the traditional metric used in the other studies.

[13] Henry (2022) collected TTC and AAC data but did not report these, but they were reported in Henry (2015). Villegas & Morgan-Short (2019) presented only a subset of TTC and AAC data, reported in Fernández (2021).

Table 4. Total accuracy, rescored TTC, TTAT, and TTATxC values for nine studies

		Total	Rescored		
Study	Group	Accuracy	TTC	TTAT	TTATxC
Culman et al. (2009)	1st Sem., PI	0.66	X	13.81	18.75
	1st Sem., SI	0.53	X	28.60	27.40
	3rd Sem., PI	0.69	X	13.86	19.50
	3rd Sem., SI	0.54	X	27.14	25.71
Henry et al. (2009)	+EI	0.66	11.37	15.74	20.00
	–EI	0.49	17.95	25.53	26.74
VanPatten & Borst (2012a)	+EI	0.73	5.25	19.17	23.17
	–EI	0.58	23.95	31.05	36.09
VanPatten & Borst (2012b)	+EI	0.61	18.35	29.48	32.43
	–EI	0.59	16.63	28.37	31.32
Henry et al. (2017)	–P+EI	0.79	5.43	12.38	18.48
	–P–EI	0.63	13.10	27.75	32.70
	+P+EI	0.77	4.59	13.59	21.06
	+P–EI	0.63	16.48	27.33	32.57
Glimois (2019)	IFPI-Mono	0.98	0.44	0.00	0.00
	IFPI-Bi	0.94	3.18	2.50	2.86
	IFSI-Mono	0.69	15.55	26.55	27.65
	IFSI-Bi	0.74	12.42	18.23	24.50
	VOCPI-Mono	0.98	0.22	0.00	0.00
	VOCPI-Bi	0.97	0.50	0.00	1.81
	VOCSI-Mono	0.70	14.26	16.05	22.47
	VOCSI-Bi	0.73	10.74	19.84	21.74
Villegas & Morgan-Short (2019)	EXP	0.66	10.95	14.37	20.21
	IMP	0.50	16.46	26.92	27.54
	Control+	0.93	0.57	0.04	0.78
Henry (2021)	Blocking (+BP)	0.71	12.57	19.52	23.00
Henry (2022)	PI	0.73	2.46	18.96	21.04
	PI+P	0.68	7.89	28.93	31.89

As seen in Table 4 scores for TTAT and TTATxC are generally higher than the TTC scores, but the relationship between groups—especially within studies—is relatively stable. Indeed, as seen in Figure 2, Traditional TTC and TTAT scores are highly correlated, as are the TTC and TTATxC scores (Table 5). This correlation remains strong when these group-level data are disaggregated and tested individually for each experiment (all $r > .429$, all $p < .001$), suggesting that all three scores indicate when participants begin to process sentences correctly. The individual data (top panels in Figure 2) do show numerous participants who scored very well on TTC but never met criterion for TTAT or TTATxC; however, participants almost never scored very well on TTAT and TTATxC while failing to meet the sequence-based criterion.

Because TTC scores should represent whether participants maintain correct processing strategies, one would expect low TTC scores to correlate with high accuracy scores (i.e., if a participant "gets it" early in the experiment their overall accuracy should be higher than one who "gets it" comparatively late). Thus, one can test TTC's overall performance by exploring the relationship between TTC scores and total accuracy. As seen in Figure 3, TTC and both TTAT methods correlate with total accuracy. This was true when data were scored at the group level, (Table 5) or disaggregated (all $r > .597$, all $p < .001$). However, an examination of the individual, disaggregated data (top panels) also shows a wide spread of total accuracy scores when TTC scores are low (0–10). Comparatively, participants display less variability

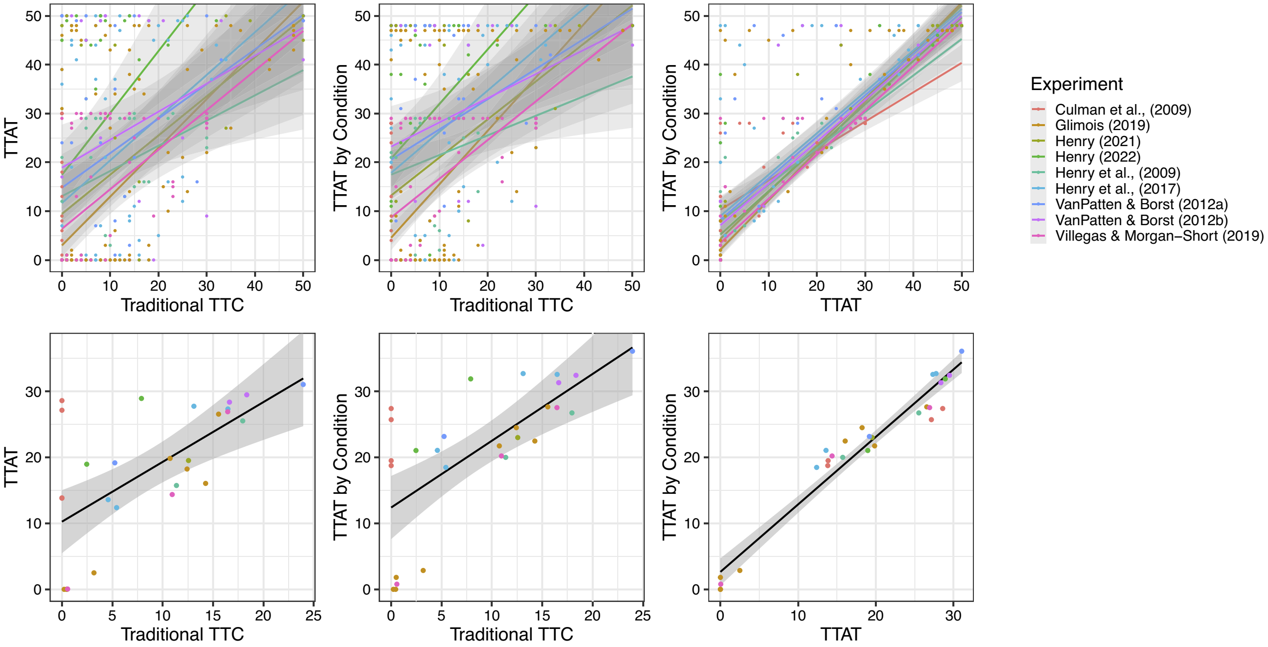


Figure 2. Correlations between traditional TTC scores and TTAT scores for individual (top) and group-level (bottom) data.

Table 5. Means, standard deviations, and correlations for group-level data with confidence intervals

Variable	M	SD	1	2	3
1. TTC	8.62	7.26			
2. TTAT	18.06	10.12	.65** [.37, .82]		
3. TTATxC	21.12	10.65	.69** [.43, .85]	.97** [.94, .99]	
4. Total accuracy	0.71	0.14	−.56** [−.77, −.23]	−.90** [−.95, −.80]	−.88** [−.94, −.75]

Note. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). **p* < .05. ***p* < .01.

in total accuracy when they have low TTAT and TTATxC scores. For all three metrics, participants who never reach criterion display a relatively wide range of total accuracy scores, though it is notable that fewer participants overall reach criterion in TTAT and TTATxC scores.

Summary

The proposed scoring metrics using percentage-based criteria seem to perform similarly to TTC and capture the point at which participants begin to process sentences correctly. However, they also provide several important advantages relative to traditional TTC scoring methods. First, because the score is based on accuracy thresholds and a maintenance of correct processing strategies, the score is less influenced by chance performance on any given item; further, because percentage-based criteria essentially combine TTC and AAC type scores, TTAT scores are more clearly defined (i.e., by definition, participants must maintain use of a processing strategy) and there is no need to use multiple measures as a control. Taken together, TTAT scores are thus more reliable and interpretable than TTC scores. This can provide a single score for use in individual difference research like Lee (2014). Perhaps most importantly, percentage-based criteria provide methodological flexibility, allowing researchers to sequence training items (pseudo)randomly, use multiple sentence types (like in Culman et al., 2009), and manipulate the distribution of target and distractor items more freely.

There are, however, several disadvantages to the method. First, although the score provides more flexibility in the definition of criterion, this introduces a certain degree of freedom for researchers and they must therefore provide clear justification for the criterion threshold. Second, percentage-based criteria are likely prone to error when the switch point between correct and incorrect processing is sudden and dramatic. Third, because scores are computed on the remaining items in a sequence, the role of chance increases as the number of items decreases.

Finally, it should also be noted that, because TTAT produces a single score like TTC, data can be analyzed using the same statistical methods used in the TTC literature. This research has traditionally focused on group-level comparisons using parametric (e.g. *t*-tests and analysis of variance [ANOVA]) or nonparametric tests (Mann–Whitney and Kruskal–Wallis tests). As the field grows more statistically sophisticated and research incorporates more complicated designs (e.g., in individual differences research), researchers will need to consider how best to use such scores and whether traditional approaches or more sophisticated approaches are more appropriate or provide more detail.

https://doi.org/10.1017/S0272263123000165 Published online by Cambridge University Press

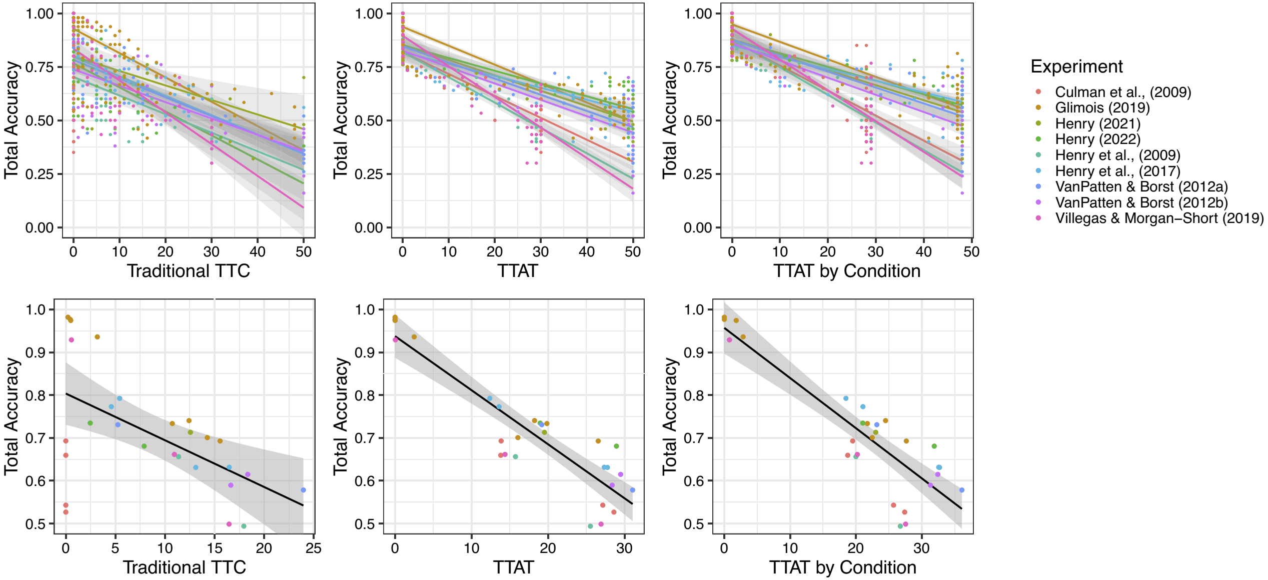


Figure 3. Correlations between total accuracy and traditional TTC, TTAT, and TTATxC scores for individual (top) and group-level (bottom) data.

Growth curve analysis

Overview of the approach

The fundamental question associated with PI studies concerns how processing and learning develops over time. Although criterion-based scoring methods attempt to capture group differences in learning by estimating the point at which sufficient learning has occurred, more sophisticated statistical approaches can provide a detailed view of learning curves and estimate differences between them. One such approach is growth curve analysis (GCA), a type of multilevel regression analysis that was designed to model change over time while also accounting for multiple sources of variation (Mirman et al., 2008). GCA was developed to overcome several challenges associated with the analysis of nested time-course data, which cannot be analyzed appropriately using *t*-tests or (repeated measures) ANOVAs (see Mirman, 2014).

GCA has been adopted in a range of studies in psychology and linguistics, perhaps most notably for eye tracking and the visual world paradigm (e.g., Henry et al., 2022; Henry, Hopp, et al., 2017; Pozzan et al., 2016; see Godfroid, 2019). However, GCA has several properties that make it suitable for analyzing accuracy data gathered during PI (or other trainings). First, because data are typically fit to polynomial functions, the noise in the data are reduced, and researchers can visualize the general learning trends more easily. To illustrate, Figure 4 shows GCA models for response accuracy in each of the nine studies listed in Table 3. Second, GCAs can employ orthogonal power polynomials, which are mutually independent and can be interpreted separately. Thus, with a relatively simple model (e.g., a model with linear, quadratic, and cubic time terms), it is possible to estimate not only whether groups differed in their overall accuracy but also whether groups differed in their learning rate. Third, GCA models can include individual differences variables like working memory (WM) whether they are treated as categorical (e.g., high WM, low WM) or continuous variables. Finally, as mentioned above, GCA overcomes statistical hurdles that make the analyses of these data with traditional metrics otherwise problematic—for example, in accounting for random effects from multiple sources. A fuller mathematical treatment of GCA is beyond the scope of this article but can be found in Mirman (2014).

A case study (from Henry et al., 2017)

To illustrate how the GCA approach can be used to analyze training data from PI studies, I present here a reanalysis of data from Henry et al. (2017), which investigated the effects of PI with and without explicit information (+/–EI) and with and without contrastive prosodic cues (+/–P). The data and R code for this analysis can be found at https://doi.org/10.18738/T8/1O8MIL. For simplicity, I focus on the accuracy of training responses given by three groups: the –P–EI group, the –P +EI group, and the +P–EI group. The research question concerns whether EI or contrastive prosodic cues lend advantages during PI for the accusative case in German. As such, the analysis focuses on the comparisons with the –P–EI group and does not consider the comparison between the –P+EI and +P–EI groups.[14],[15]

[14] GCAs do provide avenues for comparing categorical variables with more than two levels, but these two comparisons are sufficient to illustrate how GCA can be used to answer this research question. See Mirman (2014) for a technical guide on GCA and multiple comparisons.

[15] Note that Henry et al. (2017) was chosen to illustrate the use of GCA and BDOTS, in part, because they both become significantly more complicated when additional groups (comparisons) are included. The

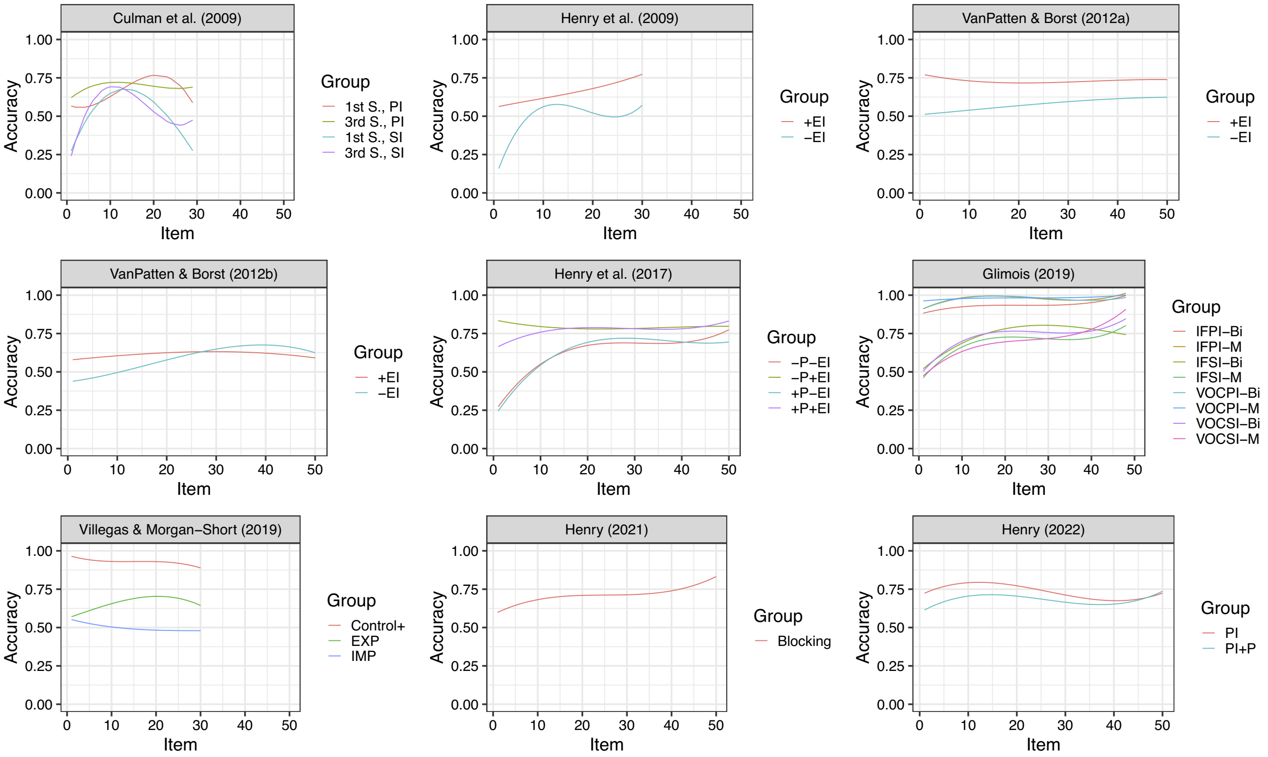


Figure 4. Growth curves for nine processing instruction studies.

The GCA analysis was used to analyze the accuracy of learner responses over the 50 training items and was conducted with a third-order (cubic) orthogonal polynomial using the empirical logit transformation. This procedure avoids potential issues arising from the binary nature of the structured input task in this study (participants selected one of two pictures during training). In addition, the use of the empirical logit transformation accounts for inaccuracies that could arise in logistic GCAs due to the low number of trials and when dealing with perfect scores (of 0 or 1; Mirman, 2014). The initial analysis included the between-subjects factors Treatment (–P–EI, –P+EI, +P–EI) as a fixed effect on all within-subjects time terms (Time1, Time2, and Time3). The –P–EI group was the focus of analyses and is treated as the baseline (reference) group. Participant was included as a random effect on all time terms. However, initial analyses resulted in singular fit. It was therefore necessary to reduce the random effects. The final model removed random effects for participant on the quadratic and cubic time terms and was given by the equation: ElogAccuracy ~ (Time1 + Time2 + Time3) × Treatment + (Time1 | participant). Model fit was evaluated through an additive statistical approach, confirming that the cubic model was the best fit for these data, Statistical significance was determined using the normal approximation, treating the *t* value as a *z* value. Analyses were conducted with R version 3.5.2 (R Core Team, 2018) using the lme4 package version 1.1-29 (Bates et al., 2015).

For this cubic model, the statistical output produces four types of effects: Main effects indicate a difference in the proportion of fixations over the entire training. Effects shown on the linear term (Time 1) indicate a sharper increase (or decrease) in accuracy over the training (i.e., a steeper slope). Effects on the quadratic (Time 2) and cubic (Time 3) terms indicate a difference in the curvature of the model. Note that in interpreting data from this model, it is important to consider the statistical output together with the visual representation of the data.

Figure 5 shows the estimated model fit (solid lines) and raw data (dashed lines) for each of the three groups under investigation, and Table 6 shows the results of the GCA. The results indicate that the –P–EI and +P–EI groups were similar in their overall accuracy (main effect) and the rate of learning (linear term), and the overall curvature of the model (quadratic and cubic terms). In contrast, the –P–EI and +P+EI groups

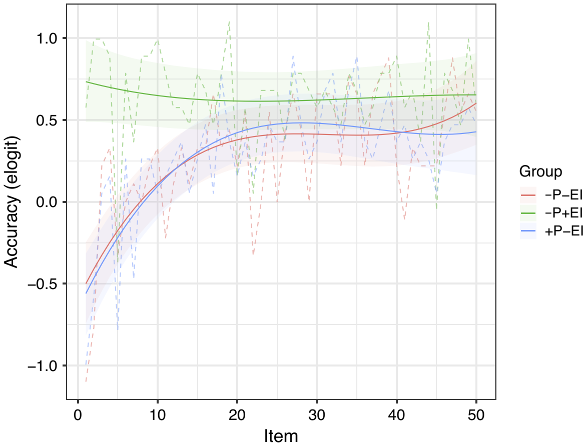


Figure 5. GCA model fit (solid lines) and transformed data (dashed lines) for three groups from Henry et al. (2017).

https://doi.org/10.1017/S0272263123000165 Published online by Cambridge University Press

Table 6. Results from the growth curve analysis

–P–EI (Baseline) vs. –P+EI				
Factor	Estimate	*SD*	*t*	*p*
Main effect	0.35	0.09	4.04	<.001***
Linear term	–1.58	0.33	–4.82	<.001***
Quadratic term	0.96	0.30	3.25	.001***
Cubic term	–0.64	0.30	–2.15	.03*

–P–EI (Baseline) vs. +P–EI				
Factor	Estimate	*SD*	*t*	*p*
Main effect	0.00	0.09	0.01	.99
Linear term	–0.07	0.33	–0.20	.84
Quadratic term	–0.40	0.30	–1.34	.18
Cubic term	–0.13	0.30	–0.42	.67

differed significantly in their overall accuracy, the rate of learning, and the curvature of the model.

Taken together, results suggest that EI had a significant effect on learner performance during the training but that prosody did not. More specifically, visual inspection of the curves suggests that EI provided a substantial initial boost to learners but that the subsequent training did not result in more accurate use over time. In contrast, the two groups that did not receive EI were quite inaccurate at the beginning of training but do learn from the training over time. These results are reflected in the statistical output, which shows that the P+EI group had higher accuracy overall (main effect), fewer learning gains (linear term), and a flatter learning trajectory (quadratic and cubic terms) than the –P–EI group. These results mirror the TTC-based analysis in Henry et al. (2017; see Table 3 for TTC scores from this study), as the –P–EI and +P–EI groups displayed similar TTC scores, whereas the –P+EI group reached criterion much sooner. Indeed, the TTAT and TTATxC scores also suggest an advantage for the –P+EI group.

Summary

As illustrated in the analysis above, GCA can be used to model changes in accuracy over time, capturing effects that are typically revealed by TTC scores and providing a more detailed account of development. That is, whereas TTC analyses are concerned with the first point of correct processing, GCA considers the whole picture. Thus, GCA can reveal differences that are not obvious from TTC scores. For example, if one compares the GCA curves for Henry et al. (2017) with Villegas & Morgan-Short (2019; Figure 4), one sees that the effect of EI in the former is immediate, whereas it is somewhat delayed for the latter. In addition, GCA provides several methodological and statistical advantages to TTC analyses. First, the use of GCA is not bound to sequencing or

research questions and design of this study allowed a comparison of three groups, which was sufficient to illustrate the advantages and disadvantages of these methods. Further, preliminary analyses suggested that one of the two planned comparisons would yield significant differences, whereas the other would not. This was viewed as advantageous to illustrate these two approaches.

distributional constraints as TTC is. Second, it allows analyses that account for multiple random effects. Finally, it provides robust controls for multiple comparisons when compared with *t* tests, ANOVA, and nonparametric tests, which have typically been employed to analyze TTC scores.

However, GCA does have some disadvantages as well. First, GCAs present several analytical challenges, even for researchers with working knowledge of mixed-effects models and statistical modeling using R. One particular difficulty that arises in this approach is in selecting models that capture the shape of the data without overfitting. As Mirman (2014) notes, there are different approaches to selecting a model, including both statistical and theoretical approaches. The approach taken here was to probe model fit through statistical testing and visual inspection and to choose the lowest order function that most clearly captures the general shape of the data, improves model fit, and is necessary to answer research questions. This is acts as a guard against both overfitting and overinterpretation, considering that higher order polynomials are much more difficult to interpret. However, researchers will surely approach this differently, as no two data sets are alike and research designs will differ from study to study. For those interested in implementing these models, Mirman's (2014) book on GCA in R provides an excellent technical reference for researchers interested GCA, including full discussion of this issue.

More importantly, GCAs, at least as presented here, are concerned with total accuracy over time and do not consider accuracy on distractors or other sentence types. Consequently, researchers need to impose external controls to ensure that accuracy curves represent accurate processing of all sentence types. Practically, this could be achieved in two ways: (1) by investigating curves for distractors separately or (2) by removing participants with low accuracy on distractor sentences.[16] Finally, although GCA models capture differences between groups over time and permit inferences about why learning curves differ, results of GCAs are expressed as differences over the entire training set. However, it is quite possible that two groups could, for example, differ at the beginning of training but not at the end. Indeed, this issue is essentially what prompted Fernández's (2008) original TTC study. Although researchers could slice the analysis into various time windows to investigate differences, this approach is not optimal because it introduces a degree of freedom for researchers and creates an opportunity for *p*-hacking, especially if this approach is not theoretically motivated and clearly defined at the outset of analysis.

Bootstrapped differences of time series

Overview of the approach

As discussed in the preceding section, one of the primary drawbacks of GCA is that it captures differences between learning curves over an entire time window and does not indicate when in the time series differences between the curves are different. One solution to this problem lies in using bootstrapped differences of time series (BDOTS) to estimate the precise items for which the two curves differed (see Oleson et al., 2017;

[16] Each approach was undertaken to verify the analysis of Henry et al. (2017) data. The GCA of distractor sentences (approach 1) showed that the –P+EI group was overall more accurate than the –P–EI group and that there were no differences between the –P–EI group and +P–EI groups. There were no differences in the learning rate or curvature of the model. A subset analysis with participants who scored above 50% on distractor items (approach 2) revealed no differences from the main analysis.

Seedorff et al., 2018). On its most basic level, BDOTS simply provides a method for comparing accuracy from two groups at each point in the training series, but it also accounts for statistical issues that arise from repeated testing of raw data (e.g., familywise Type I error).

As described in Seedorff et al. (2018), BDOTS estimates differences in four stages. First, a curve is fit for each participant individually to smooth the data and minimize idiosyncratic patterns of significance (i.e., minimize variation due to subjects).[17] Second, a bootstrapping procedure is used to estimate the standard error at each point. Third, the standard errors of the function are used to conduct two-sample *t*-tests at every point. Finally, BDOTS identifies time windows of significance using a modified Bonferroni-corrected significance level, which accounts for autocorrelation that arises from the lack of true independence between points (see Oleson et al. 2017).

BDOTS has been used primarily in recent visual world eye-tracking studies (Hendrickson et al., 2020, 2022; Kapnoula & Samuel, 2019; McMurray, Ellis, et al., 2019; McMurray, Klein-Packard, et al., 2019), in part because the R package *bdots* (Nolte et al., 2022) makes it relatively easy to fit curves that are typical for these data (i.e., Gaussian and four-parameter logistic). However, the *bdots* also allows researchers to fit other curve types using the polynomial() function during the participant fitting stage. Thus, researchers can, for example, fit a third-order polynomial function as was done for the GCA analysis above. One caveat for using BDOTS to analyze accuracy data in training studies is that the binomial nature of the data (i.e., each subject has a value of 0 or 1 at each point) could affect the standard error because BDOTS currently assumes normality, though this concern should be mitigated by the bootstrapping and estimation processes built into the approach (Oleson, personal communication). Additionally, as with any nonlinear curve-fitting procedure, researchers must be careful to select functions that capture overall trends in the data without overfitting. Because data from these studies do not necessarily follow expected curves (unlike eye-tracking data), researchers must be more cautious when analyzing these types of data with either GCA or BDOTS.

A case study from Henry, et al., (2017)

To illustrate how BDOTS adds to the GCA approach, I expand here on the previous analysis, focusing on the same three groups from Henry et al. (2017): –P–E, –P+EI, and +P–EI. The data and R code for this analysis can be found at https://doi.org/10.18738/T8/HCIQMM. Again, the research question focuses on whether EI or contrastive prosodic cues lend an advantage during training and focuses on the comparisons with the –P–EI group. Like the GCA analysis, this analysis focused on response accuracy over time and was conducted with a third-order polynomial. Because the *bdots* package cannot output multiple comparisons, the bootstrapping phase of the analysis was conducted in two stages to complete the three-way comparison of the between-subjects factor treatment: first for the –P–EI versus +P–EI comparison and then for the –P–EI versus –P+EI comparison. This analysis was completed using the *bdots* package (Nolte et al., 2022) in R (R Studio Team, 2021).[18]

[17] Although the first stage of the BDOTS analysis accounts for variance due to subjects, Seedorff et al. note that BDOTS cannot yet account for crossed random effects and instead suggest separate item and subjects analyses if necessary.

[18] Plots were created using data extracted using the writeCSV() function, which was updated on June 29, 2022. Users can install updated versions of the package with devtools::install_github("collinn/bdots").

https://doi.org/10.1017/S0272263123000165 Published online by Cambridge University Press

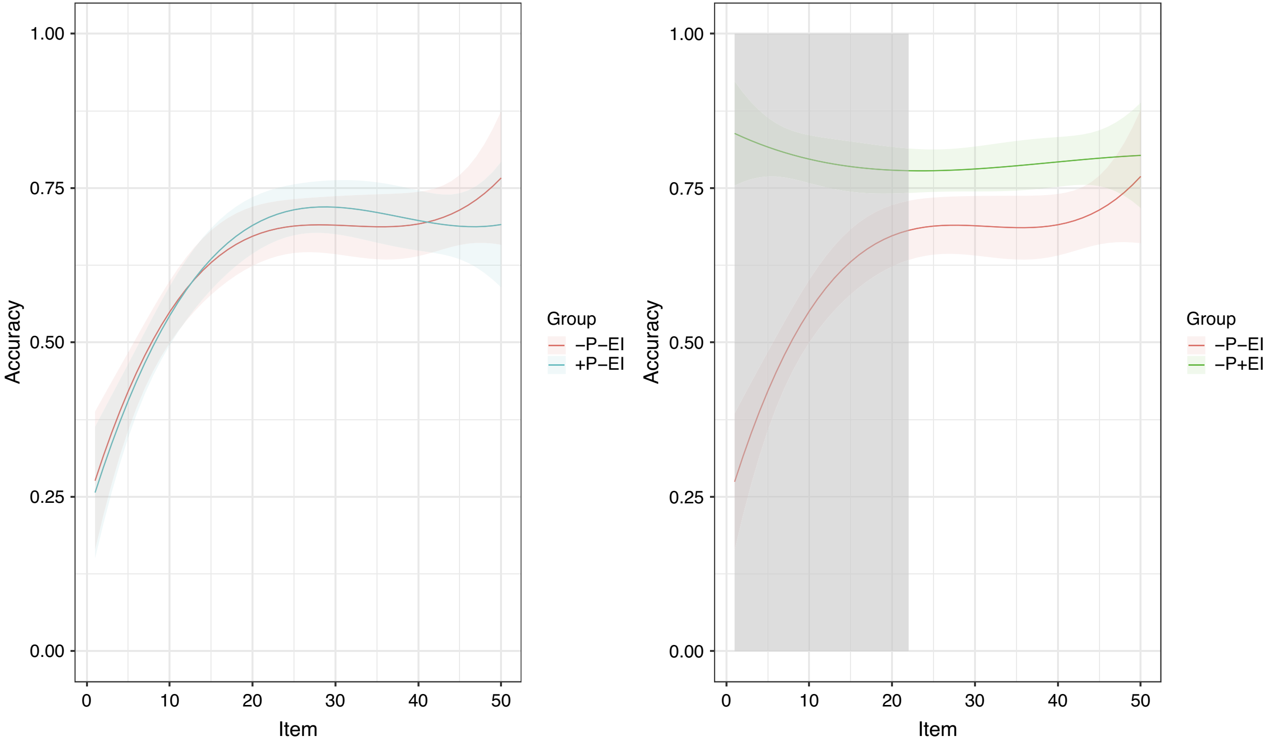


Figure 6. Results of the bootstrapped comparisons of groups from Henry et al. (2017).

In total, 62 subjects were fit using a third-degree polynomial function. For the bootstrapped comparison of the –P–EI and +P–EI groups, the autocorrelation of the *t* statistics was 0.99 and the adjusted alpha was .019. The analysis identified no regions of significance (Figure 6, left panel). For the bootstrapped comparison of the –P–EI and –P+EI groups, the autocorrelation of the *t* statistics was 0.81 and the adjusted alpha was .002. The analysis identified regions of significance between Items 1 and 23 (Figure 6, right panel). In all cases, the –P+EI group exceeded the –P–EI group (average difference = .24).

As indicated in the GCA and TTC analyses of these data, these results suggest that EI had a significant effect on learner performance during the training but that prosody did not. Moreover, these analyses confirm that EI provided an initial boost to learners, essentially giving them a 20–25-item head start. Although the advantage did not persist through the end of training (as also indicated by the immediate posttest in Henry et al., 2017), such an advantage could be important to learners who may not have much exposure to the target form.

Summary

As illustrated in this example analysis, BDOTS provides many of the same advantages as GCA, with the added benefit that it indicates where groups differ. Therefore, it can be used to analyze data from trainings designed in many different ways without the constraints that accompany TTC-based analysis. However, BDOTS also comes with several of the same disadvantages as GCA, including that researchers must control for accuracy in distractor items. Added to these disadvantages are the caveats mentioned previously, in particular that the *bdots* package does not currently have built-in options for binomial data, but this could change in the future. Additionally, as seen in the example analysis, the *bdots* package does not currently complete bootstrapping for all comparisons at once, which can make multiple comparisons tedious, especially when the research design calls for many different groups (e.g., as in Glimois, 2019). These limitations may make such analyses difficult and decrease confidence in their reliability. In such cases, researchers may wish to explore one of the alternative approaches mentioned earlier: in particular, GAMM and DPA may provide good alternatives to BDOTS, each with their own advantages and disadvantages as discussed by Ito and Knoerferle (2022). However, despite its limitations, BDOTS remains a promising avenue for analyzing data from training studies.

Conclusions

Trials to criterion has been an extremely useful method for analyzing the effects of different training paradigms, especially in the PI literature, where it helped researchers distinguish between differences that emerge during training and after training. However, there are several challenges associated with TTC in terms not only of its conceptual implementation but also the methodological restrictions that it places on study design by limiting the sequence, distribution, and randomization of items. The present study sought to shed light on these shortcomings and investigate three alternative approaches for analyzing accuracy data collected during training sequences: TTAT, GCA, and BDOTS.

The analysis here shows first and foremost that although TTC does have some shortcomings, the three alternative methods discussed here are not likely to upend

https://doi.org/10.1017/S0272263123000165 Published online by Cambridge University Press

any of the established findings in TTC research. Rather, these methods provide promising alternatives for the future, as they add detail to TTC analyses and permit researchers to ask new and interesting questions. Each of these approaches has disadvantages, and so, as always, researchers must take care in selecting the approach(es) best suited to answer the research questions and perhaps explore other alternatives in the future as well. Where appropriate, researchers may find it advantageous to use multiple approaches in their analyses, as doing so lends confidence and detail to their conclusions.

Acknowledgments. I would like to thank all of the researchers who kindly made their data available to me for this project (and several others who were willing though they were ultimately unable to share). I would also like to thank Jacob Oleson and Collin Nolte for their statistical and technical advice. Finally, I would like to thank Melisa Dracos, Hyoun-A Joo, and the members of the Second Language Studies research group at UT Austin, all of whom offered valuable feedback on the present study.

Data availability statement The experiment in this article earned an Open Materials badge for transparent practices. The materials are available at https://doi.org/10.18738/T8/T9SFAG, https://doi.org/10.18738/T8/T9SFAGI, and https://osf.io/z9psd/?

Supplementary Materials. To view supplementary material for this article, please visit http://doi.org/10.1017/S0272263123000165.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Culman, H., Henry, N., & VanPatten, B. (2009). The role of explicit information in instructed SLA: An on-line study with processing instruction and German accusative case inflections. *Die Unterrichtspraxis/Teaching German*, *42*, 19–31. https://doi.org/10.1017/S0272263109990027

Chapelle, C. A. (2020). Validity in language assessment. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 11–20). Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. https://doi.org/10.1177/0956797613504966

Diaz, E. M. (2017). The order of explicit info in processing instruction. *Applied Language Learning*, *27*, 41–72.

Farley, A. P. (2004). Processing instruction and the Spanish subjunctive: Is explicit information needed? In B. VanPatten (Ed.), *Processing instruction: Theory, research, and commentary* (pp. 231–243). Lawrence Erlbaum. https://doi.org/10.4324/9781410610195

Fernández, C. (2008). Reexamining the role of explicit information in processing instruction. *Studies in Second Language Acquisition*, *30*, 277–305. https://doi.org/10.1017/S0272263108080467

Fernández, C. (2021). Trials-to-criterion as a methodological option to measure language processing in processing instruction. In Michael J. Leeser, Gregory D. Keatin, & Wynne Wong (Eds.), *Research on second language processing and processing instruction: Studies in honor of Bill VanPatten* (236–259). https://doi.org/10.1075/sibil.62.08fer

Glimois, L. (2019). *The effects of input flood, structured input, explicit information, and language background on beginner learners' acquisition of a target structure in Mandarin Chinese* [Unpublished doctoral dissertation]. The Ohio State University.

Godfroid, A. (2019). *Eye tracking in second language acquisition and bilingualism.* Routledge. https://doi.org/10.4324/9781315775616

Hendrickson, K., Apfelbaum, K., Goodwin, C., Blomquist, C., Klein, K., & McMurray, B. (2022). The profile of real-time competition in spoken and written word recognition: More similar than different. *Quarterly Journal of Experimental Psychology*, *75*, 1653–1673. https://doi.org/10.1177/17470218211056842

Hendrickson, K., Spinelli, J., & Walker, E. (2020). Cognitive processes underlying spoken word recognition during soft speech. *Cognition*, *198*, Article 104196. https://doi.org/10.1016/j.cognition.2020.104196

Henry, N. (2015). Morphosyntactic processing, cue interaction, and the effects of instruction: an investigation of processing instruction and the acquisition of case markings in L2 German *[Unpublished Doctoral Dissertation]*. The Pennsylvania State University.

Henry, N. (2021). The use of blocking and inhibition training in processing instruction. *International Review of Applied Linguistics in Language Teaching*. Advance online publication. https://doi.org/10.1515/iral-2021-0068

Henry, N. (2022). The additive use of prosody and morphosyntax in L2 German. *Studies in Second Language Acquisition*. Advance online publication. https://doi.org/10.1017/S0272263122000092

Henry, N. (forthcoming). Explicit information, input processing, and SLA. In W. Wong & J. Barcroft (Eds.), *The Routledge handbook of second language acquisition and input processing*. Routledge.

Henry, N., Culman, H., & VanPatten, B. (2009). More on the effects of explicit information in instructed SLA. *Studies in Second Language Acquisition*, *31*, 559–575. https://doi.org/10.1017/S0272263109990027

Henry, N., Hopp, H., & Jackson, C. N. (2017). Cue additivity in predictive processing of word order in German. *Language, Cognition and Neuroscience*, *32*, 1229–1249. https://doi.org/10.1080/23273798.2017.1327080

Henry, N., Jackson, C. N., & DiMidio, J. (2017). The role of prosody and explicit instruction in processing instruction. *Modern Language Journal*, *101*, 1–21.

Henry, N., Jackson, C. N., & Hopp, H. (2022). Cue coalitions and additivity in predictive processing: The interaction between case and prosody in L2 German. *Second Language Research*, *38*, 397–422. https://doi.org/10.1177/0267658320963151

Ito, A., & Knoeferle, P. (2022). Analysing data from the psycholinguistic visual-world paradigm: Comparison of different analysis methods. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-022-01969-3

Kapnoula, E. C., & Samuel, A. G. (2019). Voices in the mental lexicon: Words carry indexical information that can affect access to their meaning. *Journal of Memory and Language*, *107*, 111–127. https://doi.org/10.1016/j.jml.2019.05.001

Lee, J. F. (2014). The relationship between learning rate and learning outcome for processing instruction on the Spanish passive voice. In A. G. Benati, C. Laval, & M. J. Arche (Eds.), *The grammar dimension in instructed second language learning* (pp. 148–163). Bloomsbury.

Lee, J. F. (2015). Processing instruction on the Spanish passive with transfer-of-training effects to anaphoric and cataphoric reference contexts. *International Review of Applied Linguistics in Language Teaching*, *53*, 203–223. https://doi.org/10.1515/iral-2015-0010

Lee, J. F., & Doherty, S. (2020). Prior knowledge and other individual differences in the development of accuracy over the time-course of a pre-test/treatment/post-test study of instructed second language acquisition. *Instructed Second Language Acquisition*, *4*, 124–157. https://doi.org/10.1558/isla.40608

McMurray, B., Ellis, T. P., & Apfelbaum, K. S. (2019). How do you deal with uncertainty? Cochlear implant users differ in the dynamics of lexical processing of noncanonical inputs. *Ear & Hearing*, *40*, 961–980. https://doi.org/10.1097/AUD.0000000000000681

McMurray, B., Klein-Packard, J., & Tomblin, J. B. (2019). A real-time mechanism underlying lexical deficits in developmental language disorder: Between-word inhibition. *Cognition*, *191*, Article 104000. https://doi.org/10.1016/j.cognition.2019.06.012

Mirman, D. (2014). *Growth curve analysis and visualization using R*. CRC Press.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*, 475–494. https://doi.org/10.1016/j.jml.2007.11.006

Nolte, C., Seedorff, M., Oleson, J., Brown, G., Cavanaugh, J., & McMurray, B. (2022). *bdots: Bootstrapped Differences of Time Series* (R package version 1.1.0). https://github.com/collinn/bdots

Oleson, J. J., Cavanaugh, J. E., McMurray, B., & Brown, G. (2017). Detecting time-specific differences between temporal nonlinear curves: Analyzing data from the visual world paradigm. *Statistical Methods in Medical Research*, *26*, 2708–2725. https://doi.org/10.1177/0962280215607411

Pozzan, L., Gleitman, L. R., & Trueswell, J. C. (2016). Semantic ambiguity and syntactic bootstrapping: The case of conjoined-subject intransitive sentences. *Language Learning and Development*, *12*, 14–41. https://doi.org/10.1080/15475441.2015.1018420

R Studio Team. (2021). *RStudio: Integreted development environment for R.* (1.4.1103). RStudio, PBC.

https://doi.org/10.1017/S0272263123000165 Published online by Cambridge University Press

Sanz, C., & Morgan-Short, K. (2004). Positive evidence versus explicit rule presentation and explicit negative feedback: A computer-assisted study. *Language Learning*, *54*, 35–78. https://doi.org/10.1111/j.1467-9922.2004.00248.x

Seedorff, M., Oleson, J., & McMurray, B. (2018). Detecting when timeseries differ: Using the bootstrapped differences of timeseries (BDOTS) to analyze visual world paradigm data (and more). *Journal of Memory and Language*, *102*, 55–67. https://doi.org/10.1016/j.jml.2018.05.004

VanPatten, B. (2004). Input Processing in second language acquisition. In B. VanPatten (Ed.), *Processing instruction: Theory, research, and commentary* (pp. 5–31). Lawrence Erlbaum.

VanPatten, B. (2015). Foundations of processing instruction. *International Review of Applied Linguistics in Language Teaching*, *53*, 91–109. https://doi.org/10.1515/iral-2015-0005

VanPatten, B., & Borst, S. (2012a). The roles of explicit information and grammatical sensitivity in processing instruction: Nominative-accusative case marking and word order in German L2. *Foreign Language Annals*, *45*, 92–109. https://doi.org/10.111/j.1944-9720.2012.01169.x.FOREIGN

VanPatten, B., & Borst, S. (2012b). The roles of explicit information and grammatical sensitivity in the processing of clitic direct object pronouns and word order in Spanish L2. *Hispania*, *95*, 270–284. https://doi.org/10.1353/hpn.2012.0059

VanPatten, B., Collopy, E., Price, J. E., Borst, S., & Qualin, A. (2013). Explicit information, grammatical sensitivity, and the first-noun principle: A cross-linguistic study in processing instruction. *The Modern Language Journal*, *97*, 506–527. https://doi.org/10.1111/j.1540-4781.2013.12007.x

VanPatten, B., & Oikkenon, S. (1996). Explanation versus structured input in processing instruction. *Studies in Second Language Acquisition*, *18*, 495–510. http://journals.cambridge.org/abstract_S0272263100015394

Villegas, B., & Morgan-Short, K. (2019). The effect of training condition and working memory on second language development of a complex form: The Spanish subjunctive. In H. Wilson, N. King, E. J. Park, & K. Childress (Eds.), *Selected proceedings of the 2017 Second Language Research Forum* (pp. 185–199). Cascadilla Press

Wong, W. (forthcoming). Processing instruction and structured input. In W. Wong & J. Barcroft (Eds.), *The Routledge handbook of second language acquisition and input processing*. Routledge.

Cite this article: Henry, N. (2023). Reevaluating trials to criterion as a measure in second language research. *Studies in Second Language Acquisition,* 45: 1481–1505. https://doi.org/10.1017/S0272263123000165