Behav Res (2013) 45:151–159
DOI 10.3758/s13428-012-0240-y

The past tense inflection project (PTIP): speeded past tense inflections, imageability ratings, and past tense consistency measures for 2,200 verbs

Emily R. Cohen-Shikora · David A. Balota · Abhi Kapuria · Melvin J. Yap

Published online: 7 September 2012

Abstract The processes involved in past tense verb generation have been central to models of inflectional morphology. However, the empirical support for such models has often been based on studies of accuracy in past tense verb formation on a relatively small set of items. We present the first large-scale study of past tense inflection (the Past Tense Inflection Project, or PTIP) that affords response time, accuracy, and error analyses in the generation of the past tense form from the present tense form for over 2,000 verbs. In addition to standard lexical variables (such as word frequency, length, and orthographic and phonological neighborhood), we have also developed new measures of past tense neighborhood consistency and verb imageability for these stimuli, and via regression analyses we demonstrate the utility of these new measures in predicting past tense verb generation. The PTIP can be used to further evaluate existing models, to provide well controlled stimuli for new studies, and to uncover novel theoretical principles in past tense morphology.

Keywords Verb processing · Megastudy · Past tense inflection · Item-level variance · Verb consistency · Verb imageability

A long-standing question in language acquisition and inflectional morphology is how individuals produce the past tense form of a verb. Past tense inflection (PTI), like spelling-to-sound conversion in English, is quasiregular, meaning that a set of generally applicable descriptive rules are useful for most verbs (e.g., add *-ed* to the stem form), but there are also some irregular forms (e.g., *do–did*) and subregularities (as in the *eep–ept* past tense family: *sleep–slept*, *weep–wept*, *keep–kept*, etc.). Indeed, past tense inflection has been a central focus of the debate between parallel distributed models (Rumelhart & McClelland, 1986) and more symbolic, rule-based models (Pinker & Ullman, 2002) of language processing. The models have in large part been assessed in terms of how well they have been able to capture the richness of the past tense morphology empirical literature.

Although there has been extensive theoretical work in the area of past tense verb generation, experimental work examining response times (RTs) has been relatively limited. For example, in the stem inflection task, participants are asked to produce the past tense (real or hypothetical) of a target verb or novel nonword (e.g., Bybee & Slobin, 1982; Cortese, Balota, Sergent-Marshall, Buckner, & Gold, 2006; Woollams, Joanisse, & Patterson, 2009), with the type of response being the critical dependent measure. Other researchers have used acceptability ratings, in which participants are asked to judge experimenter-provided past tenses of real words (as in Prado & Ullman, 2009) and nonwords (as in Berko's, 1958, classic *wug–wugged* study; Albright & Hayes, 2003; Bybee & Moder, 1983; Prasada & Pinker, 1993). Accuracy-focused and content-focused studies of past tense inflection also have often used special populations, such as children (e.g., Berko, 1958; Bybee & Slobin, 1982; Marchman, 1997) or neuropsychological populations, such as patients with anomia, agrammatism, Alzheimer's disease, or Parkinson's disease (for a review, see Pinker & Ullman, 2002). Several studies have also used neuroimaging techniques, such as event-related functional magnetic

Electronic supplementary material The online version of this article (doi:10.3758/s13428-012-0240-y) contains supplementary material, which is available to authorized users.

E. R. Cohen-Shikora (✉) · D. A. Balota · A. Kapuria
Department of Psychology, Washington University in St. Louis, St. Louis, MO, USA
e-mail: ecohensh@wustl.edu

M. J. Yap
Department of Psychology, National University of Singapore, Singapore, Singapore

resonance imaging and positron emission tomography, which attempt to identify the brain regions active during inflection and to report brain region or network activation in lieu of response latencies on a behavioral task (e.g., Dhond, Marinkovic, Dale, Witzel, & Halgen, 2003; Indefrey et al., 1997; Sach, Seitz, & Indefrey, 2004; Ullman, Bergida, & O'Craven, 1997; but see Joanisse & Seidenberg, 2005, for a notable exception).

These studies are certainly informative, but a fuller picture of PTI might be afforded by examining the speed with which nonclinical adults' responses are executed, and not only the content of their responses. Indeed, in another important quasiregular domain with a rich theoretical lineage—isolated word recognition—considerable emphasis is placed on RTs (see Balota, Yap, & Cortese, 2006, for a review), although both RTs and accuracy are typically considered. In this light, it is surprising that more work has not focused on RT as the primary dependent variable in past tense verb generation.

Only a few previous studies of past tense verb inflection have used RT as a dependent variable: Joanisse and Seidenberg (2005) and Woollams et al. (2009) collected past tense production latencies on relatively small subsets of verbs (44 and 60 verbs, respectively). Also, two unpublished studies presented at conferences, Seidenberg and Bruck (1990) and Prasada, Pinker, and Snyder (1990), have examined response latencies in a past tense inflection task. These studies diverged with respect to their results and conclusions, and thus do not present a cohesive and empirically driven description of past tense inflection. For example, Joanisse and Seidenberg did not find reliable differences in RTs or accuracy between irregular verbs and "pseudoregular" verbs (verbs that do not take *-ed* endings, but nonetheless show some subregularity, as in the *sleep–slept*, *weep–wept* examples above). In addition, Woollams et al. had their participants generate the past tense given either present tense stems or action pictures. They found that the benefit of regular over irregular verbs that has been demonstrated in previous studies was absent for the picture-inflection condition, and thus described these data as supportive of a single-route model.

Seidenberg and Bruck (1990) provided further support for a connectionist model in their finding of a slowdown in RTs for regular verbs with semiregular phonological neighbors relative to regular verbs with regular neighbors. In contrast to the connectionist perspective, Prasada et al. (1990) argued that their findings of RT performance dissociations between regular and irregular verbs—specifically, word frequency effects for irregular, but not regular, verbs—were more supportive of a symbolic rule-based model. The divergent conclusions from these studies are possibly due in part to different methods and stimuli across the experiments.

One way to address the discrepancies among previous studies, as well as the limitations associated with factorial designs employing relatively few stimuli, is to sample a much larger set of items from the language. Megastudies include stimuli on the order of thousands, rather than 50 to 100, and allow for the effects of variables to be modeled continuously rather than categorically (see Balota, Yap, Hutchison, & Cortese, 2012, for a recent review). The inclusion of larger samples of items also helps reduce item-specific idiosyncrasies, which may influence the conclusions in smaller studies. Another benefit of large-scale megastudies is that once collected, the data can be used to select well-characterized stimuli for empirical studies and to ask targeted theoretical questions about existing data. Indeed, megastudies have become a critical method of model testing since their first use over 20 years ago (Seidenberg & Waters, 1989; Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty, 1995). Perry, Ziegler, and Zorzi (2007, 2010) recently relied heavily on the English Lexicon Project (ELP) megastudy data to constrain the CDP+ and CDP++ models of visual word recognition. Given the success that this megastudy approach has shown within the domain of visual word recognition, and the relative paucity of RT studies in the important area of past tense verb inflection, we developed the Past Tense Inflection Project (PTIP) to fill this void in the literature.

In addition to providing a large database of response latencies and accuracies for past tense verb inflection, we also developed two new measures that are important to consider in past tense inflection, consistency and imageability. Similar to the spelling-to-sound consistency measure that has been well-studied in visual word recognition research (Treiman et al., 1995; Yap & Balota, 2009; Ziegler, Stone, & Jacobs, 1997), consistency was defined as the similarity between the target verb's past tense inflection and its rhyme neighbors' past tense inflections. This measure was computed by taking a present-tense target's number of friends (verbs conjugated similarly) and dividing it by its total number of rhyme neighbors (including enemies—those verbs whose rhyme neighbors are not conjugated in the same way). For example, the word BIND (conjugated BOUND) has seven verb rhyme neighbors: BIND, BLIND, FIND, GRIND, GRIND, MIND, and WIND. GRIND is counted twice because it has two viable past tense inflections (GRINDED and GROUND), and BIND is counted once as a rhyme for itself, which was done to prevent words with no rhyme neighbors from having a consistency value of zero. This procedure yields four friends (BOUND, FOUND, GROUND, and WOUND) and three enemies (BLINDED, GRINDED, and MINDED), producing a consistency value of .57 (4/7) for the word BIND.

The second variable that we measured was imageability. Imageability is a variable that reflects the extent to which

one is able to form a mental image of a word, and indeed many imageability norms are already available (e.g., Cortese & Fugett, 2004; Schock, Cortese, & Khanna, 2012; Stadthagen-Gonzalez & Davis, 2006). However, we know of few imageability norms that specify the grammatical class of the word that is being imaged (although see Prado & Ullman, 2009). Hence, we collected online data through Amazon's Mechanical Turk, a crowd-sourcing marketplace site used for recruiting workers for a paid task (for more information, see Mason & Suri, 2012). In our study, we were careful to ensure that the verb interpretation of each word was used, by specifying in the instructions that all words were verbs, and by presenting each verb in the infinitive with the particle "to" preceding it. Like consistency, imageability has been important to the field of visual word recognition, but it has been included in only one study of past tense inflection (Prado & Ullman, 2009). Providing consistency and imageability values for a large set of verbs is an important first step toward extending the examination of these variables in this important literature.

In order to demonstrate the utility of the new variables (consistency and imageability ratings of the verb form of words), we will report initial item-level regression analyses and demonstrate that both of these new variables have predictive utility in past tense verb formation performance. The important issue is whether these new variables predict any unique variance above and beyond the standard variables available in the literature, such as word frequency, regularity, and length. These analyses are primarily reported as initial demonstrations of the new variables' utility and are not intended to be a comprehensive analysis of this rich data set, which is beyond the scope of the present report.

The present study is based on 89 participants' accuracy and RTs for a past tense inflection task with 2,200 verbs. Each participant produced responses to 888 items. For each verb in the PTIP database, we included measures of RT, accuracy, and regularization errors (e.g., saying GRINDED for GRIND), along with the new imageability and consistency measures described above. The PTIP database is useful in examining the specific effects of predictor variables on RT and accuracy and allows for detailed item-level predictions. It is available as supplementary materials with this article for researchers who plan to examine other theoretical questions about past tense inflection, or are hoping to select well-controlled and well-examined stimuli for new studies. These data will serve as both a reference and an impetus for further research in the domain of past tense inflection.

Experiment 1

The first experiment was conducted in order to collect imageability rating norms for the verbs in the PTIP database.

Method

Participants A group of 218 participants were recruited via Amazon's Mechanical Turk (AMT; see Mason & Suri, 2012). The participants were paid $1.50 for their time, and they provided no demographic information. Participants were excluded on the basis of having completed more than one set of item ratings (N = 20) and of particularly low correlations to the item means calculated across all participants (cutoff of r < .10, N = 30). In both cases, it appears that these participants were not following instructions. After these participants were eliminated, 168 remained.

Materials The 2,200 words from the PTIP database (see below), plus another 112 words for use in another study, were divided into eight lists of 289 items each. The eight lists were presented as separate jobs in AMT.

Procedure Each participant completed one list of the rating task, which was presented in Adobe Flash and appeared after a consent screen in the AMT job description. The instructions were the same as those used in Cortese and Fugett (2004), which instructed participants to rate each word on the basis of its ability to call a mental image to mind on a scale from 1 (*very low imageability*) to 7 (*very high imageability*), with an option for "Do not know this word" (0). However, unlike previous imageability studies, and in order to ensure that participants were rating the imageability of the verb reading of the word, each word was preceded with the word TO—for instance, TO BIND. Verbs were presented one at a time in random order, and RTs and ratings were collected for each item.

Results

The ratings were aggregated across participants for each item (excluding "do not know" responses), so that one mean imageability value was calculated for each verb. These values were used in Experiment 2 (see below). The mean rating across all verbs was 4.28 (SD = 0.92), and the mean RT across all verbs was 3,191 ms (SD = 1,695). The overall split-half reliability was r = .80, p < .001.

Experiment 2

Method

Participants

A group of 113 native English-speaking college students from the Washington University subject pool participated in

the study. After eliminating extreme outliers (less than 80 % accuracy overall; four participants) or participants whose data were subject to recording error (missing sound files from which to code accuracy—20 participants), 89 participants contributed to the final database.

Materials

Stimuli were selected from the database that Albright and Hayes (2003) created from the English portion of the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995). All of Albright and Hayes's irregular verbs were selected (N = 187), as well as a subset of 2,013 regular verbs (approximately every third regular verb in the database). *Regularity* in this case was defined as words conjugated with one of the three *-ed* past tense allomorphs (/Id/, /t/, /d/). Each participant completed 888 trials. The regular verb stimuli were divided into three between-subjects lists for counterbalancing purposes, and the full set of irregular items was seen in every list. This procedure was used so that participants would not get in the habit of consistently producing the regular allomorphs for all verbs, regular and irregular. Thus, each list of 888 trials included the 187 irregular forms. The data in the two PTIP supplementary files includes item-level data across all 2,200 words, item-level data for each of the three lists separately, and raw trial-level data for each individual participant.

Procedure

The participants were seated in front of a 16-in. CRT monitor, which displayed present tense verbs on each trial. Participants were instructed to speak aloud the past tense of the verb onscreen as quickly and accurately as possible. Response latencies were recorded by a microphone attached to a voice key, and a separate microphone recorded audio files of the participant's responses for the purpose of rescoring them offline.

Each trial began with a fixation cross presented for 1,000 ms. Next, the present tense verb was presented and remained onscreen until the voice key detected input, after which the experimenter designated the response as correct or as a regularization (*eated* instead of *ate*), other error (such as providing the past participle—*eaten* instead of *ate*), or microphone error/dysfluency. Finally, a 1,000-ms blank black screen was presented after the experimenter had coded each trial. The stimuli were presented in white, 14-point font against a black background. The experiment was controlled using the E-Prime software, Version 2.0 (Psychology Software Tools, Pittsburgh, PA).

Summary of data available in the PTIP database

Table 1 provides an example of five regular and five irregular verbs, along with the descriptive information available in the PTIP database for each verb.

The means and standard deviations for each of the variables in the PTIP database, along with the standard predictor variables listed below, which are available from the ELP, are displayed in Table 2. Table 3 displays the correlation matrix among the following predictor variables and behavioral measures.

Regularity As discussed above, regular verbs were defined as taking one of the three *-ed* past tense allomorphs (/Id/, /t/, /d/), and irregular verbs were those taking any other past tense ending. Verbs that take more than one acceptable past tense (such as *kneel–kneeled/knelt*) were included as stimuli (N = 36), and either inflection was considered to be correct.

Length Length in letters (length), syllables (N syll), and morphemes (N morph).

Log frequency Log frequency is based on the Hyperspace Analogue to Language (HAL) frequency norms (Lund & Burgess, 1996).

Orthographic N/phonological N Orthographic N (Ortho N) and phonological N (Phono N) are measures of how many other words (neighbors) can be made by changing one of the letters or phonemes in the original (see, e.g., Coltheart, Davelaar, Jonasson, & Besner, 1977).

Consistency The past tense consistency measure was calculated in the following manner: First, rhyme neighbors (with the same number of syllables) for the present tense version of each verb were obtained from a rhyming website (www.rhymezone.com) based on the Carnegie Mellon Pronouncing Dictionary (Weide, 1995). Rhymes with the same past tense conjugation (including the target word itself) were considered friends, and those with different past tense conjugations were considered enemies. Past tense words that were classified as archaic, rare, or British were not included as neighbors, on the basis of Dictionary.com (Dictionary.com LLC, n.d.). Rhymes and targets with more than one viable past tense were included as two separate entries. Consistency was calculated by dividing the number of friends by the total number of rhyme neighbors [Consistency=Friends / (Friends +Enemies)]. The friends and enemies of each verb are available upon request.

Imageability See the results of Experiment 1.

Table 1 Sample information from the supplementary file

Verb	Past Tense	Regularity	Consistency	Imageability	PTI RT	PTI Z	PTI Acc	Reg Err
breach	breached	regular	0.83	3.47	790	.15	0.90	–
equip	equipped	regular	0.83	4.56	747	.08	1.00	–
mumble	mumbled	regular	1.00	5.07	685	–.41	0.90	–
pioneer	pioneered	regular	0.94	2.62	697	–.24	0.97	–
remember	remembered	regular	1.00	3.39	687	–.33	1.00	–
forgive	forgave	irregular	0.50	3.81	711	–.21	0.88	.03
mistake	mistook	irregular	0.18	3.81	758	.07	.76	.06
find	found	irregular	0.57	4.88	803	.23	.90	.00
inlay	inlaid	irregular	1.00	3.52	748	–.03	.92	.00
slink	slunk	irregular	1.00	3.93	753	.06	.64	.10

The columns marked "PTI" reflect mean item performance on the past tense inflection task: RT is in raw response time units, Z is response time converted to within-participants standard deviations, Acc is proportion correct, and Reg Err is the proportion of regularization error responses

Results and discussion

In addition to providing a database including past tense verb generation performance, new imageability ratings, and new consistency measures for the full set of 2,200 verbs, we report here initial regression analyses of the past tense verb performance. The primary goal of these analyses was to determine whether the new consistency and imageability measures capture unique variance.

The preliminary analyses reported below were conducted at the item level, by aggregating valid responses for an item across participants. For simplicity, the following analyses were conducted after eliminating 36 verbs that have more than one valid past tense inflection (such as *hang*, for which both *hung* and *hanged* are valid). Only correct responses with latencies between 200 and 3,000 ms were included, which eliminated 10.6 % of the original 75,633 trials. Further individual screening then eliminated responses greater than three standard deviations from each participant's mean, accounting for an additional 1.3 % of trials. Here, and in the database, we report both mean raw RTs and *z*-score-transformed RTs (based on within-subjects means and standard deviations). The *z*-score-transformed RTs control for differences in overall RT and variance between participants (see Faust, Balota, Spieler, & Ferraro, 1999). Based on the above procedures, for regular verbs, approximately 30 participants contributed to each mean, and for irregular verbs, approximately 90 participants contributed to each mean (since the irregular verbs were presented across all three lists). Importantly, listwise data are also available in the supplementary materials, which include approximately 30 observations for both regular and irregular items.

Table 2 Descriptives and behavioral results for the stimuli (all verbs, $N = 2{,}169$)

	Mean	*SD*
Length	6.21	2.0
Log HAL frequency	7.31	2.4
Number of syllables	1.87	0.9
Number of morphemes	1.38	0.6
Ortho *N*	3.99	5.7
Phono *N*	9.26	12.9
Imageability	4.20	1.2
Consistency	0.91	0.2
Mean latency—Raw	735	70
Mean latency—Standardized	–.07	.36
Proportion correct	.92	.10
Proportion regularization errors	.01	.03
Proportion microphone errors	.05	.06

Length, log HAL frequency, number of syllables, number of morphemes, ortho *N*, and phono *N* are taken from the English Lexicon Project (Balota et al., 2007). *Log HAL frequency* log of the Hyperspace Analogue to Language frequency (Lund & Burgess, 1996), *Ortho N* orthographic *N*, and *Phono N* phonological *N* (Coltheart et al., 1977)

The item-level results are provided in Table 4. The data were analyzed using a hierarchical regression procedure. The first step of the regressions included all of the phonological onset characteristics. Specifically, we dichotomously (as 1 or 0) classified each phoneme onset with respect to the following phonological features for each of 13 categories (see Treiman et al., 1995), where 1 denotes the presence of the feature and 0 denotes its absence: affricative, alveolar, bilabial, dental, fricative, glottal, labiodental, liquid, nasal, palatal, stop, velar, and voiced. This classification should capture sensitivity to voice key biases and may also be sensitive to the ease of implementation of different phonological codes during production. Spieler and Balota (1997) noted considerable predictive power of onsets for single-syllable words. After onsets were entered, lexical variables—including length, number of syllables, number of morphemes, log HAL frequency, ortho *N*, phono *N*, and regularity—were entered

Table 3 Correlation matrix, predictor variables, and dependent variables

Predictor/Dependent Variable	1	2	3	4	5	6	7	8	9	10	11
1. Regularity	–										
2. Length	–.13***	–									
3. Log frequency	.16***	–.42***	–								
4. # Syllables	–.12***	.87***	–.39***	–							
5. # Morphemes	–.02	.66***	–.33***	.68***	–						
6. Ortho *N*	.15***	–.67***	.36***	–.57***	–.40***	–					
7. Phono *N*	.17***	–.64***	.36***	–.60***	–.42***	.79***	–				
8. Imageability	.09***	–.29***	.17***	–.34***	–.27***	.21***	.21***	–			
9. Consistency	–.77***	.23***	–.22***	.23***	.09***	–.21***	–.26***	–.10***	–		
10. Raw RT	.21***	.08*	–.24***	.08***	.12***	.08***	.11***	–.19***	–.25***	–	
11. *Z* scores	.20***	.09***	–.25***	.09***	.13***	.08***	.11***	–.21***	–.24***	.98***	
12. Proportion correct	–.35***	.07**	.09***	.05*	.02	–.08***	–.12***	.08***	.34***	–.52***	–.55***

$^{+}p < .10$. $^{*}p < .05$. $^{***}p < .001$. Length, log HAL frequency, number of syllables, number of morphemes, ortho *N*, and phono *N* are taken from the English Lexicon Project (Balota et al., 2007). *Log HAL frequency* log of the Hyperspace Analogue to Language frequency (Lund & Burgess, 1996), *Ortho N* orthographic *N*, and *Phono N* phonological *N* (Coltheart et al., 1977)

in the second step. The third step included the imageability measure, and the fourth step included the consistency measure, to determine whether these novel metrics added any unique predictive power after the other variables had been entered into the regression model.

First, consider the predictive power of onsets. As is shown in Table 4, the onset coding scheme reliably predicted both RT and accuracy performance; however, the total variance accounted for in this first step was at best only 4.0 % for the *z*-scored data. In order to further explore this variable, we selected 2,091 items that overlapped with the words in the ELP and found that the same onsets predicted 7.8 % of the variance in speeded pronunciation data in the ELP. Although one must be cautious about comparisons across these samples, it is possible that the more complex past tense verb inflections may increase variability, as compared to simple pronunciations, and may decrease sensitivity to the onset coding schema. This is consistent with observations made by Yap

Table 4 Item-level results

Predictor	Raw RT		*Z* Scores		Proportion Correct	
	ΔR^2	β	ΔR^2	β	ΔR^2	β
Step 1	.032***		.040***		.019***	
Onsets						
Step 2	.166***		.170***		.153***	
Regularity		.028		.009		–.242***
Length		.126**		.151**		.065
Log freq		–.315***		–.323***		.202***
N Syll		.013		.010		–.046
N Morph		.047†		.040		.035
Ortho *N*		.107**		.121***		.016
Phono *N*		.182***		.189***		–.091**
Step 3	.034***		.038***		.015***	
Imageability		–.192***		–.203***		.127***
Step 4	.029***		.031***		.011***	
Consistency		–.279***		–.286***		.171***
Total R^2	.260		.279		.198	

$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$. Log HAL frequency, number of syllables, number of morphemes, ortho *N*, and phono *N* are taken from the English Lexicon Project (Balota et al., 2007). *Log freq* log of the Hyperspace Analogue to Language frequency (Lund & Burgess, 1996), *N Syll* number of syllables, *N Morph* number of morphemes, *Ortho N* orthographic *N*, and *Phono N* phonological *N* (Coltheart et al., 1977)

and Balota (2009) when comparing the predictive power of onsets in speeded pronunciation for single-syllable, as compared to multisyllable, words.

Turning to the lexical variables, as is shown in Table 4, the raw and *z*-score RT data indicate that length, ortho *N*, and phono *N* are all reliable predictors in the positive direction, and log frequency, imageability, and consistency are all significant predictors in the negative direction. Thus, long words and words with many orthographic and phonological neighbors are conjugated more slowly, relative to words with fewer letters and more sparsely populated orthographic and phonological neighborhoods. Importantly, low-frequency, low-imageability, and low-consistency verbs are also produced more slowly relative to their higher-frequency, higher-imageability, and higher-consistency counterparts. The accuracy data paint a similar picture: Regularity, frequency, phonological *N*, imageability, and consistency are all moderate predictors in the expected directions. Taken together, the accuracy data suggest that regular verbs, higher-frequency verbs, verbs with fewer phonological neighbors, and higher-imageability verbs are all associated with relatively higher accuracy.

Importantly, the results from the regression analyses clearly indicate that the new imageability and consistency measures both reliably predict unique variance in RTs and accuracy. With respect to the consistency measure, this suggests that a rather straightforward measure of consistency (based on a type estimate) accounts for unique variance, which highlights the importance of this measure in past tense verb inflection above and beyond regularity. In addition, the imageability measure also nicely captures unique variance in past tense verb generation and suggests that semantics plays a role in this syntactic operation (see Prado & Ullman, 2009, for some discussion of frequency and imageability effects in past tense inflection).

A few additional results are of note from the first-pass regression analyses on these data. First, the effect of frequency in raw response latencies, $\beta = -.315$, $p < .001$, and *z* scores, $\beta = -.323$, $p < .001$, appears relatively smaller than is typical in speeded pronunciation or in lexical decision performance. In order to directly test this, the same regression analyses with the same predictor set were run on overlapping items from the ELP speeded pronunciation and lexical decision latencies. Indeed, the results from the analyses including pronunciation as the dependent measure on the same words indicated that the regression coefficients for word frequency were slightly larger than our past tense verb regression coefficients, with raw pronunciation RT $\beta = -.382$, $p < .001$, and *z*-score $\beta = -.384$, $p < .001$. Turning to ELP lexical decision RTs, the regression analyses yielded raw RT $\beta = -.566$, $p < .001$, and *z*-score $\beta = -.626$, $p < .001$. Thus, it appears that the role of word frequency in past tense verb inflection is relatively similar to, albeit slightly smaller than, its role in speeded word pronunciation of the same items, and considerably smaller than its role in lexical decision.

Finally, we also explored whether the new past tense verb imageability ratings were indeed more related to past tense inflection than were previous imageability norms that are available, which have not emphasized the verb interpretation of words. Fortunately, a relatively large subset of the words (N = 1,394) overlapped with existing imageability measures of monosyllabic (Cortese & Fugett, 2004) and disyllabic (Schock et al., 2012) words, although in these previous studies, the verb status of the stimuli was not stressed in the instructions. The correlation between the verb-based imageability ratings and these earlier norms was moderate, $r = .43$, $p < .001$. More importantly, the verb imageability ratings were indeed more related to the past tense inflection task than were the standard imageability ratings. Specifically, the new verb imageability ratings predicted more variance in past tense RTs (after phonological onsets and lexical variables were partialed out), $\Delta R^2 = .021$, $p < .001$, $\beta = -.154$, $p < .001$, than the existing norms did, $\Delta R^2 = .013$, $p < .001$, $\beta = -.118$, $p < .001$. As predicted, this pattern was reversed with speeded word pronunciation as the dependent measure; the new verb imageability ratings predicted less variance in ELP naming RTs, $\Delta R^2 = .007$, $p < .001$, $\beta = -.089$, $p < .001$, than the existing norms did, $\Delta R^2 = .017$, $p < .001$, $\beta = -.137$, $p < .001$, for the same set of items. This pattern was also reversed when considering lexical decision performance; the new verb imageability ratings predicted less variance in ELP lexical decision RTs, $\Delta R^2 = .027$, $p < .001$, $\beta = -.174$, $p < .001$, than the existing norms did, $\Delta R^2 = .066$, $p < .001$, $\beta = -.268$, $p < .001$. This is an important observation, indicating that a standard normative procedure (e.g., imageability ratings) can produce different patterns, depending on the context of the rating (focusing on verb interpretation vs. no instruction) and the task demands (speeded pronunciation or past tense inflection).

An additional question that was addressed is whether something is aberrant about participants' performance in a megastudy, in which they conjugate over 800 verbs in one sitting, as opposed to performance in a more traditional factorial experiment, in which they conjugate closer to 100 verbs. One of the ways that this question has been addressed in the visual word recognition literature (as in Balota & Spieler, 1998) is by comparing megastudy item means with item means from an existing smaller study. To this end, we examined 39 PTIP items that overlapped with Woollams et al. (2009), and we found a moderate correlation in raw item-level RTs from the two studies, $r = .385$, $p = .014$; with a clear outlier removed, that correlation increased, $r = .493$, $p = .001$. Therefore, even with such a small and selective set of items, there seems to be some evidence in the present megastudy that the data are convergent with smaller studies of past tense inflection.

Conclusions

In summary, the present study provides a large database of speeded past tense verb production on 2,200 present tense verbs, along with new measures of past tense verb consistency and verb imageability ratings. The results indicated that this corpus of data is stable enough to produce reliable effects from standard predictor variables and that two new measures (consistency of the past tense inflection and verb imageability ratings) captured unique variance in these data. Indeed the variance captured by the lexical predictor variables is quite comparable to the predictive power for the best lexical variables observed in the ELP for speeded pronunciation of monosyllabic words (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004). Hence, the PTIP affords a unique and stable database to investigate the semiregular domain of past tense verb inflection. It appears that RT (along with accuracy) has the potential to be as informative within this semiregular domain as it has been in the domains of speeded pronunciation and lexical decision performance.

Author Note Thanks are extended to Jonathan Jackson and three anonymous reviewers for their constructive comments on an earlier version of the manuscript, Adam Albright for providing the list of verbs to be tested, Anna Woollams for providing the mean item-level RTs from her study, and Dung Bui and Michael Cortese for advice with Experiment 1. This project was supported by NIA Grant No. T32 AG00030.

References

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition, 90,* 119–161. doi:10.1016/S0010-0277(03)00146-X

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (Release 2, CD-ROM)*. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133,* 283–316. doi:10.1037/0096-3445.133.2.283

Balota, D. A., & Spieler, D. H. (1998). The utility of item-level analyses in model evaluation: A reply to Seidenberg and Plaut. *Psychological Science, 9,* 238–240. doi:10.1111/1467-9280.00047

Balota, D. A., Yap, M. J., & Cortese, M. J. (2006). Visual word recognition: The journey from features to meaning (a travel update). In M. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd ed., pp. 285–375). Amsterdam, The Netherlands: Academic Press.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39,* 445–459. doi:10.3758/BF03193014

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: Large scale analysis of lexical processes. In J. S. Adelman (Ed.), *Visual word recognition: Vol. 1. Models and methods, orthography, and phonology* (pp. 90–115). London: Psychology Press.

Berko, J. (1958). The child's learning of English morphology. *Word, 14,* 150–177.

Bybee, J. L., & Moder, C. L. (1983). Morphological classes as natural categories. *Language, 59,* 251–270. doi:10.2307/413574

Bybee, J. L., & Slobin, D. I. (1982). Rules and schemas in the development and use of the English past tense. *Language, 58,* 265–289. doi:10.2307/414099

Coltheart, M., Davelaar, E., Jonasson, J., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555). Hillsdale: Erlbaum.

Cortese, M. J., Balota, D. A., Sergent-Marshall, S. D., Buckner, R. L., & Gold, B. T. (2006). Consistency and regularity in past-tense verb generation in healthy ageing, Alzheimer's disease, and semantic dementia. *Cognitive Neuropsychology, 23,* 856–76. doi:10.1080/02643290500483124

Cortese, M. J., & Fugett, A. (2004). Imageability ratings for 3,000 monosyllabic words. *Behavior Research Methods, Instruments, & Computers, 36,* 384–387. doi:10.3758/BF03195585

Dhond, R. P., Marinkovic, K., Dale, A. M., Witzel, T., & Halgen, E. (2003). Spatiotemporal maps of past-tense verb inflection. *NeuroImage, 19,* 91–100. doi:10.1016/S1053-8119(03)00047-8

Dictionary.com LLC. (n.d.). *Dictionary.com Unabridged*. Retrieved from http://dictionary.reference.com

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin, 125,* 777–799. doi:10.1037/0033-2909. 125.6.777

Indefrey, P., Brown, C., Hagoort, P., Herzog, H., Sach, M., & Seitz, R. J. (1997). A PET study of cerebral activation patterns induced by verb inflection. *NeuroImage, 5,* S548.

Joanisse, M. F., & Seidenberg, M. S. (2005). Imaging the past: Neural activation in frontal and temporal regions during regular and irregular past-tense processing. *Cognitive, Affective, & Behavioral Neuroscience, 5,* 282–296. doi:10.3758/CABN.5.3.282

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28,* 203–208. doi:10.3758/BF03204766

Marchman, V. A. (1997). Children's productivity in the English past tense: The role of frequency, phonology, and neighborhood structure. *Cognitive Science, 21,* 283–304.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44,* 1–23. doi:10.3758/s13428-011-0124-6

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review, 114,* 273–315. doi:10.1037/0033-295X.114.2.273

Perry, C., Ziegler, J. C., & Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the connectionist dual process (CDP++) model. *Cognitive Psychology, 61,* 106–151. doi:10.1016/j.cogpsych.2010.04.001

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences, 6,* 456–463. doi:10.1016/S1364-6613(02)01990-3

Prado, E. L., & Ullman, M. T. (2009). Can imageability help us draw the line between storage and composition? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 849–866. doi:10.1037/a0015286

Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes, 8,* 1–56. doi:10.1080/01690969308406948

Prasada, S., Pinker, S., & Snyder, W. (1990, November). *Some evidence that irregular forms are retrieved from memory but regular forms are rule generated*. Paper presented at the meeting of the Psychonomic Society, New Orleans.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructures of cognition (Vol. 2)* (pp. 216–271). Cambridge: MIT Press.

Sach, M., Seitz, R. J., & Indefrey, P. (2004). Unified inflectional processing of regular and irregular verbs: A PET study. *NeuroReport, 15,* 533–537.

Schock, J., Cortese, M. J., & Khanna, M. M. (2012). Imageability estimates for 3,000 disyllabic words. *Behavior Research Methods, 44,* 374–379. doi:10.3758/s13428-012-0209-x

Seidenberg, M., & Bruck, M. (1990). *Consistency effects in the generation of past tense morphology.* Paper presented at the meeting of the Psychonomic Society, New Orleans.

Seidenberg, M. S., & Waters, G. S. (1989). Word recognition and naming: A mega study. *Bulletin of the Psychonomic Society, 27,* 489.

Spieler, D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science, 8,* 411–416. doi:10.1111/j.1467-9280.1997.tb00453.x

Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods, 38,* 598–605. doi:10.3758/BF03193891

Treiman, R., Mullennix, J., Bijeljac-Babic, R., & Richmond-Welty, E. D. (1995). The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology: General, 124,* 107–136. doi:10.1037/0096-3445.124.2.107

Ullman, M. T., Bergida, R., & O'Craven, K. (1997). Distinct fMRI activation patterns for regular and irregular past tense. *NeuroImage, 5,* S549.

Weide, R. L. (1995). Carnegie Mellon Pronouncing Dictionary. (0.4) Retrieved from www.speech.cs.cmu.edu/cgi-bin/cmudict

Woollams, A. M., Joanisse, M., & Patterson, K. (2009). Past-tense generation from form versus meaning: Behavioural data and simulation evidence. *Journal of Memory and Language, 61,* 55–76. doi:10.1016/j.jml.2009.02.002

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language, 60,* 502–529. doi:10.1016/j.jml.2009.02.001

Ziegler, J. C., Stone, G. O., & Jacobs, A. M. (1997). What is the pronunciation for -ough and the spelling for /u/? A database for computing feedforward and feedback consistency in English. *Behavior Research Methods, Instruments, & Computers, 29,* 600–618. doi:10.3758/BF03210615