
Proceedings Paper:
Prange, Jakob and Wong, Ivy orcid.org/0000-0002-4774-6147 (2023) Preanalyzing L2 PrepositionLearning with Bayesian Mixed Effects and a Pretrained Language Model. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). , Association for Computational Linguistics, 12722–12736.

https://doi.org/10.18653/v1/2023.acl-long.712

Reanalyzing L2 Preposition Learning with Bayesian Mixed Effects and a Pretrained Language Model

Jakob Prange
Hong Kong Polytechnic University
jakob.prange@polyu.edu.hk

Man Ho Ivy Wong
Hong Kong Shue Yan University
mhwong@hksyu.edu

Abstract

We use both Bayesian and neural models to dissect a data set of Chinese learners' pre- and post-interventional responses to two tests measuring their understanding of English prepositions. The results mostly replicate previous findings from frequentist analyses and reveal new and crucial interactions between student ability, task type, and stimulus sentence. Given the sparsity of the data as well as high diversity among learners, the Bayesian method proves most useful; but we also see potential in using language model probabilities as predictors of grammaticality and learnability.[1]

1 Introduction

Learning a second or third language is hard—not only for NLP models but also for humans! Which linguistic properties and external factors make it so difficult? And how can we improve instruction and testing to help learners accomplish their goals? Here we also ask a third question: How can we best apply different computational models to such behavioral experimental data in order to get intuitive and detailed answers to the first two questions in a practical and efficient way? For example, we are interested in whether language model (LM) probabilities might give a rough estimate of grammaticality and learning difficulty (table 1, right columns).

This work is in part a replication study of Wong (2022), who, in addressing these questions about native Chinese speakers' learning of English prepositions in context (see examples in table 1), mainly focused on instructional intervention and found generally positive effects as well as differences between instruction types, in particular favoring conceptual over rule-based teaching. We pick up where Wong (2022) left off and search for more fine-grained patterns among students' individual differences, linguistic items, stimulus sentences, and their grammaticality. Our main hypothesis is that the full story of the complex interactions among these factors can only be revealed by modeling them holistically. Such a fine-grained holistic analysis is well-aligned with Item Response Theory (IRT; Fischer, 1973; Lord, 1980). IRT allows us to formulate models in terms of predicting whether students provide the intended response to each test item. We consider sparse Bayesian and dense neural versions of this framework. We can then inspect how strongly each model input (the linguistic, experimental, and student-specific factors mentioned above, which are realized as random and fixed effects for the Bayesian model and feature vectors for the neural model) affects the outcome. As a representative of yet another modeling strategy, and also as an additional input to the IRT models, we obtain probability estimates for the target prepositions in each stimulus sentence from a pretrained transformer LM. These probabilities serve as a proxy for contextual formulaicity, as learned distributionally from a large corpus.

While the theoretical advantages of Bayesian over frequentist statistics, as well as the generally strong performance of neuro-distributional models, are often cited as justification for choosing one particular modeling method, both replication studies and side-by-side comparisons of such drastically different modeling strategies for linguistic analysis remain rare (with notable exceptions, e.g., Michaelov et al., 2023; Tack, 2021).

We contribute to this important development by

- designing (§4.1), fitting, and evaluating (§5.1) a Bayesian mixed effects model on Wong's (2022) data (§3), considering more potential linguistic and human factors in preposition learning than previously and finding significant effects for several of them;
- training (§4.2) and evaluating an analogous multilayer perceptron (MLP) model and comparing it with the Bayesian model in terms of

[1] Our experimental code is available at https://github.com/jakpra/L2-Prepositions.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 12722–12736
July 9-14, 2023

#	Usage	Stimuli	Grammatical?	Student judgment pre	Student judgment post	LM p_{tgt}	LM p_{ctx}	
1a	HIR-Spat	The bell hung **over** the baby's cradle and made him smile.	✓	80.65	92.59	**4.15**	**54.66**	As expected
1b		**through**	✗	50.00	46.67	0.03	54.10	
2a	HIR-Abst	The tutors watched **over** the students during the oral presentation.	✓	80.00	96.77	**96.70**	**67.19**	
2b		**on**	✗	35.00	46.15	0.07	50.64	
3a	CVR-Abst	Tremendous fear fell **over** the town after the murder.	✓	71.43	91.67	**18.58**	**65.61**	
3b		**through**	✗	41.67	27.27	0.39	63.91	
4a	CRS-Spat	The painter reached **over** the paint can for a brush.	✓	41.18	63.33	0.05	**49.50**	sign (Δp_{tgt}) ???
4b		**through**	✗	36.11	33.33	**0.78**	45.42	
5a	CRS-Abst	The lawyer jumped **over** a few pages of the contract.	✓	72.73	94.12	6.97	**52.39**	
5b		**to**	✗	42.11	34.78	**20.27**	51.66	
6a	CVR-Abst	Happiness diffused **over** the guests when they see the newly-weds.	✓	44.44	88.46	2.47	**65.27**	
6b		**on**	✗	80.00	35.48	**3.19**	62.82	
7a	CVR-Spat	The canvas stretched **over** a large hole in the road.	✓	44.12	70.37	**17.99**	51.66	sign (Δp_{ctx}) ???
7b		**through**	✗	55.56	60.00	15.52	**52.79**	
8a	CVR-Abst	The tension swept **over** the school when the alarm rang.	✓	66.67	100.00	**3.93**	46.61	
8b		**onto**	✗	37.84	12.50	<0.01	**46.63**	
9a	CRS-Abst	The politicians skipped **over** sensitive topics during the debate.	✓	83.33	94.59	**2.88**	40.80	
9b		**to**	✗	60.98	35.00	0.17	**42.06**	

Table 1: Examples of stimulus sentences for grammatical (✓) and ungrammatical (✗) preposition use. In the **Student judgment** columns we show the percentage of students who judged the example as grammatical at the pretest (including control group) and posttest (treatment groups only) in Wong's study. The **LM** columns show our probed RoBERTa probabilities p_{tgt} and p_{ctx} [in %], which are defined in §4.3 and discussed in §5.3.

both feature ablation and overall prediction accuracy of the outcome, i.e., whether a student will answer a test prompt correctly (§5.2);

- and probing a pretrained LM (§4.3 and §5) for contextual probabilities of target prepositions in order to determine their correlation—and thus, practical usefulness—with human language learning.

Thus, we aim to both better explain L2 preposition learning and compare Bayesian, frequentist, and neural approaches to doing so.

2 Background

2.1 English Preposition Semantics

Prepositions are among the most frequently used word classes in the English language—they make up between 6 and 10 % of all word tokens depending on text type and other factors (cf. Schneider et al., 2018). This is because English does not have a full-fledged morphological case system and instead often expresses semantic roles via word order and lexical markers like prepositions. At the same time, the inventory of preposition forms is relatively small—a closed set of largely grammaticalized function words covering a wide range of predictive, configurational, and other relational meanings. The resulting many-to-many mapping between word forms and meanings is complex and warrants nuanced linguistic annotation, analysis, and computational modeling in context (O'Hara and Wiebe, 2003; Hovy et al., 2010; Srikumar and Roth, 2013; Schneider et al., 2018; Kim et al., 2019b). Further, considerable cross-linguistic variation in the precise syntax-semantics interactions of prepositions and case has been shown to affect not only machine translation (Hashemi and Hwa, 2014; Weller et al., 2014; Popović, 2017), but also construal in human translation (Hwang et al., 2020; Peng et al., 2020; Prange and Schneider, 2021) and—crucially—learner writing (Littlemore and Low, 2006; Mueller, 2011; Gvarishvili, 2013; Kranzlein et al., 2020).

2.2 Cognitive and Concept-based Instruction

Cognitive linguistics (CogLx) maintains that many aspects of natural language semantics are grounded in extra-linguistic cognition, even (or especially) when they do not directly arise from syntactic composition, or at the lexical level. For example, Brugman (1988), Lakoff (1987), and Tyler and Evans (2003) argue that spatial prepositions evoke a net-

work of interrelated senses, ranging from more prototypical to extended and abstract ones. Incorporating such conceptual connectedness into language instruction has shown some benefits (Tyler, 2012; Boers and Demecheleer, 1998; Lam, 2009).

2.3 Computational Modeling in SLA

Until recently, most studies in applied linguistics and second-language acquisition (SLA)—insofar as they are quantitative—have relied on null-hypothesis testing with frequentist statistical measurements like analysis of variance (ANOVA) (Norouzian et al., 2018). This has the advantage that it is generally unambiguous and interpretable what is being tested (because concrete and specific hypotheses need to be formulated ahead of time) and that conclusions are based directly on data without any potentially confounding modeling mechanisms. At the same time, frequentist analyses are relatively rigid, and thus run into efficiency, sparsity, and reliability issues as interactions of interest grow more complex. Li and Lan (2022) propound a more widespread use of computational modeling and AI in language learning and education research. A promising alternative exists in the form of Bayesian models (e.g., Murakami and Ellis, 2022; Privitera et al., 2022; Guo and Ellis, 2021; Norouzian et al., 2018, 2019), which circumvent sparsity by sampling from latent distributions and offer intuitive measures of uncertainty "for free" in form of the estimated distributions' scale parameters. They can also be made very efficient to train by utilizing stochastic variational inference (SVI).

Bayesian modeling for educational applications goes hand-in-hand with Item Response Theory (IRT; Fischer, 1973; Lord, 1980), which posits that learning outcomes depend on both student aptitude and test item difficulty. This addresses another limitation of frequentist analysis—the focus on aggregate test scores—by modeling each student's response to each item individually. We loosely follow this general paradigm with our model implementations, without committing to any specific theoretical assumptions.

Within NLP, Bayesian and IRT-based approaches have been used to evaluate both human annotators (Rehbein and Ruppenhofer, 2017; Passonneau and Carpenter, 2014) and models (Kwako et al., 2022; Sedoc and Ungar, 2020), to conduct text analysis (Kornilova et al., 2022; Bamman et al., 2014; Wang et al., 2012), and natural language inference (Gantt et al., 2020).

Murakami and Ellis (2022) show that grammar learning can be affected by contextual predictability (or formulaicity). While they used a simple n-gram model, we account for this phenomenon more broadly with a pretrained transformer LM.

3 Original Study and Data

Wong (2022) measured students' pre- and post-interventional understanding of the English prepositions *in*, *at*, and *over*, particularly contrasting CogLx/schematics-based instruction with different flavors of rule-based methods. To this end, intermediate learners of English (all university students) with first languages Mandarin or Cantonese took initial English language tests ('pretest') targeting different usages of prepositions. They were then taught with one of four methods (incl. one control group, who received instruction about definite and indefinite articles instead of prepositions), and subsequently tested two more times. There were two different tests: a grammaticality judgment test (GJT) to measure effects on language processing and a picture elicitation test (PET) to measure effects on production.

While all preposition-focused training was found to enhance learners' understanding of prepositions compared to both the pretest and the control group, schematics-based mediation led to stronger learning results than any of the other methods, especially at the PET (fig. 1) and on spatial usages of prepositions (the interaction between instruction method and spatial usage is not shown in fig. 1 for brevity). These latter findings in particular support our hypothesis that in addition to external factors like task type and instruction method, learning difficulty may also be affected by *inherent linguistic* properties of the prepositions and their usages (just as, e.g., Guo and Ellis (2021) show for distributional properties of grammatical suffixes). In this work we take a second look at Wong's data to directly address this possibility for preposition learning.

3.1 Data Summary

We conduct all of our computational analyses with Wong's data (stimuli and behavioral results) but expand on the original study by explicitly modeling as potential factors several additional dimensions, relating to individual differences and interactions among stimuli, task types, and students (table 2, §3.2 and §3.3). 71 students (after outlier filtering) participated in the study. There are a total of 48

test items (12 senses × 4 contexts) and 22 fillers for the GJT as well as 36 test items (12 senses × 3 contexts) and 15 fillers for the PET. Outlier students and filler items are removed before any analysis/model training, resulting in 17,644 data points overall (GJT: 10,156; PET: 7,488).

3.2 Stimulus Sentences

In the GJT (but not in the PET), students receive a linguistic stimulus to evaluate for grammaticality (see examples in table 1). Intended-grammatical stimuli involve target prepositions used in a sentence context that evokes their intended sense or function (fxn), either literally/spatially or figuratively/abstractly. For each intended-grammatical stimulus, there is an intended-***un***grammatical stimulus, consisting of the same sentence context but replacing the target preposition with another that is meant to fit the context less well.

3.3 Categorical Features

Instruction method. The main goal of Wong's (2022) study was to compare CogLx-based schematic mediation (*SM*) with more traditional rule-and-exemplar (*RM*) and bare-bones correctness-based mediation (*CM*). SM, RM, and CM instruction focused on the same preposition forms and usages students were tested on.

Time of test. Students were tested three times: Two days before instructional intervention (*PRE*test, ◁ in fig. 1), two days after instruction (*POST*test, ○), and again 3 weeks later (*DeLaY*ed posttest, ▷).

Preposition form, function (fxn), and usage. The test cues are made up of 6 pairs of preposition usages across three forms: '*in*' with the CONTAINMENT (CTN) function; '*at*' with the TARGET (TGT) and POINT (PNT) functions; and '*over*' with the HIGHER (HIR), ACROSS (CRS), and COVER (CVR) functions. Each usage pair consists of a spatial (e.g., 'in the box') and a non-spatial cue (e.g., 'in love') sharing the same schematization (in this case, CONTAINMENT). The cues were selected based on the Principled Polysemy Framework (Tyler and Evans, 2003), thereby ruling out overly fine-grained senses and allowing systematic presentation for instruction and testing.

Test type. In the *GJT*, learners had to decide, for each stimulus sentence containing a preposition, whether the whole sentence is "correct" or "incorrect".[2] We consider as a potential factor on the outcome whether a stimulus is intended-grammatical (*GJT-Y*) or not (*GJT-N*). In the *PET*, learners were shown an illustration of a concrete scenario instantiating one of the cues and were asked to produce a descriptive sentence containing a preposition. Responses were counted as correct if they chose the target preposition.

		W22	Ours
Random Effects			
Feature	**Values**		
Instruction	*SM, RM, CM, CTRL*	✓	✓
Time	*PRE, POST, DLY*	✓	✓
Test	*GJT, PET*	✓	✓
Usage	*Spatial, Abstract*	✓	✓
Answer	*GJT-Y, GJT-N, PET*	✗	✓
Form-Fxn	*in*-CTN, *at*-TGT *at*-PNT, *over*-HIR, *over*-CRS, *over*-CVR	✗	✓
Student	s_1, ..., s_{71}	✗	✓
Fixed Effects			
p_{tgt}—LM probability of target preposition		✗	✓
p_{ctx}—Avg. LM prob. of non-tgt tokens in sent.		✗	✓

Table 2: Features under consideration in Wong (2022) (W22) and our work.

Students. By adding local student identities to the model input (anonymized as, e.g., s_1, s_{23}), we allow fine-grained degrees of freedom w.r.t. individual differences, as is suggested by IRT.

4 Models

Our main point of reference (or quasi-baseline) is Wong's frequentist data analysis, which is summarized in §3. In this work, we newly consider the following different modeling strategies: We train a **Bayesian logistic model** (**BLM**, §4.1) as well as a small **multilayer perceptron** (**MLP**, §4.2) on the same data. With the BLM we can define and interpret the precise structure of how individual features and their interactions affect the outcome. In contrast, the MLP utilizes nonlinear activation functions and multiple iterations/layers of computation, allowing it to pick up on complex interactions among input features without prior specification and thus to potentially achieve higher predictive accuracy, at the cost of interpretability. Both the BLM and MLP are implemented in Python and PyTorch, and are light-weight enough to be trained and run on a laptop CPU within several minutes for training and several seconds for inference. We also query a pretrained **neural language model** (**LM**, namely RoBERTa; Liu et al., 2019b) to obtain contextual probabilities for the stimulus sentences used

[2] The testing prompt did not explicitly highlight or otherwise draw attention to the preposition in question.

in the grammaticality judgment test and add those probabilities to the BLM and MLP's inputs (§4.3).

4.1 Bayesian Logistic Model

We model the posterior likelihood of a correct response (i.e., a given student providing the intended answer to a given stimulus) as a logistic regression conditional on the aforementioned categorical variables. Concretely, responses are sampled from a Bernoulli distribution with log-odds proportional to the weighted sum of the random and fixed effects. As potential factors we consider the features listed in §3.3 and table 2, as well as their mutual interactions. For the *students* feature, to keep model size manageable, we only consider pairwise interactions with usage (spatial/abstract), form-fxn, and answer. Otherwise all n-wise interactions are included. The effects' weight coefficients are sampled from Normal distributions whose means and standard deviations are fitted to the training data via SVI with the AdamW optimizer, AutoNormal guide, and ELBO loss. We use standard-normal priors for means and flat half-normal priors for standard deviations, meaning that, by default, parameter estimates are pulled towards null-effects, and will only get more extreme if there is strong evidence for it. The model is implemented using the Pyro-PPL/BRMP libraries (Bingham et al., 2018).

4.2 Multilayer Perceptron

We train and test a multilayer perceptron (MLP) with depth 3. We mirror the BLM setup by treating student response correctness as the output and optimization objective and the different feature sets as concatenated embedding vectors. Between hidden layers we apply the GELU activation function, and during training additionally dropout with $p = 0.2$ before activation. We also apply dropout with $p = 0.1$ to the input layer. We minimize binary cross-entropy loss using the AdamW optimizer. We train for up to 25 epochs but stop early if dev set accuracy does not increase for 3 consecutive epochs.

4.3 RoBERTa

We feed the GJT stimulus sentences to RoBERTa-base (Liu et al., 2019b, accessed via Huggingface-transformers). RoBERTa a pretrained neural LM based on the transformer architecture (Vaswani et al., 2017) and trained on English literary and Wikipedia texts to optimize the masked-token and next-sentence prediction objectives. For each sentence, we wish to obtain RoBERTa's posterior probability estimates for each observed word token $w_i \in \mathbf{w}_{0:n-1}$, given $\mathbf{w}_{0:n-1} \backslash \{w_i\}$, i.e., all other words in that sentence. Thus we run RoBERTa n times, each time i masking out w_i in the input. From these n sets of probabilities, we extract two measurements of formulaicity we expect to be relevant to our modeling objective of student response correctness:[3] (a) p_{tgt}, the contextual probability of the target or alternate preposition given all other words in the sentence and (b) p_{ctx}, the average contextual probability of all words *except* the preposition.[4] Examples are given in table 1. We standardize these two variables to $\mathcal{N}(0,1)$ and add them to the BLM (as fixed effects, both individually and with interactions) and MLP (as scalar input features).

5 Evaluation

We first analyze the BLM's learned latent coefficients (§5.1). Then we compare different versions of the BLM and MLP w.r.t. their ability to predict unseen student responses using their estimated weighting of linguistic and external features as well as LM probabilities (§5.2). Finally, we manually inspect a small set of stimulus sentences with anomalous LM probabilities w.r.t. their intended grammaticality and observed judgments (§5.3).

5.1 Determining Relevant Input Features

Setup. We fit BLMs on the entire data set (without reserving dev or eval splits). We run SVI for 1000 iterations with a sample size of 100 and a fixed random seed. We compute effect sizes (Cohen's d), and p-values based on 95%-confidence intervals of differences between estimated parameter values (Altman and Bland, 2011).

Replication. As in Wong (2022), we use the features *instruction, time, form-fxn, usage*, and additionally let the model learn individual coefficients for each student. Separate models were trained for GJT and PET. As shown in fig. 1, we mostly replicate similar trends (differences between differences) as found previously, namely:

- *Time:* DLY ≈ POST > PRE;
- *Instruction:* treatment > ctrl; SM > CM ≈ RM;

[3] We also preliminarily experimented with inputting the entire LM hidden state of the last layer to the models but did not find it to be helpful. Kauf et al. (2022) found that alignment with human judgments varies from layer to layer, which presents an interesting avenue for future work.

[4] Note that the preposition token still has the potential to affect the other words' probabilities by occurring in their context condition.

- and we generally see larger learning effects in the PET than in the GJT.

However, many effect sizes are amplified—and thus p-values more significant-looking—in our model. A potential explanation for this could be that the BLM models each individual item response whereas ANOVA only considers overall %-correct. We are thus comparing effects on all students' accuracy at multiple test items in aggregate with effects on each student's accuracy at each test item separately. It seems intuitive that the latter 'micro-effects' are much greater on average than the former 'macro-effects', which are themselves effects on the average performance metric. Another reason could be that because the Bayesian effect sizes stem from simulated data points, they are only indirectly related to the real underlying data via SVI. The estimated distribution these samples are drawn from only approximates the real data and thus the effect size estimations may be over-confident. See §6.1 for a discussion of advantages and disadvantages.

Although our model estimates spatial usages as generally more difficult than abstract ones, we do not replicate Wong's finding of an *interaction* between abstractness and instruction or time. Still, our Bayesian quasi-IRT approach allows us to find additional interesting patterns that could not have been captured by a frequentist analysis[5] as they involve student-level and item-level interactions:

Answer type and individual differences. We trained a single combined model on both GJT and PET data. As can be expected, in addition to the overall trends (fig. 1), we also find a strong effect for expected answer type (fig. 2): the *receptive* task of accepting grammatical items (GJT-Y) is much easier than the *productive* task of choosing the right preposition when describing a picture (PET). Interestingly, ruling out ungrammatical items (GJT-N) is equally as difficult as the PET. In addition, outcomes are affected by individual differences between students, and student aptitude heavily depends on answer type (fig. 3) as well as on preposition form/function (fig. 5 in appendix A). There is some (negative) correlation between individual aptitudes at GJT-N and GJT-Y and some (positive) correlation between GJT-N and PET. Still, both correlations are weak (R^2 = 0.23 and 0.20).

In sum, not only do receptive vs. productive task

[5] Or only very tediously so.

[6] Where W22 does not report Cohen's d, we show their reported partial-eta-squared η_p^2 instead.

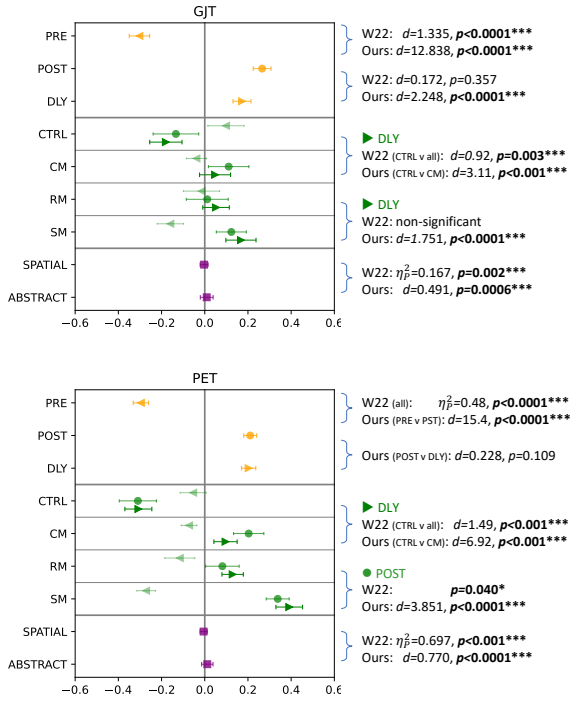

Figure 1: Summary of our Bayesian effect estimations (marginal means and standard deviations over model parameters) for selected features. Coefficient values (x-axis) indicate the extent to which the feature value (y-axis) contributes to a correct (positive) or incorrect (negative) student response. On the right we compare effect sizes (Cohen's d) and statistical significance to Wong's (2022) frequentist analysis.[6]

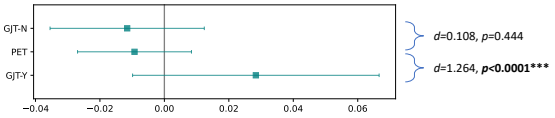

Figure 2: Estimated effects for different answer types.

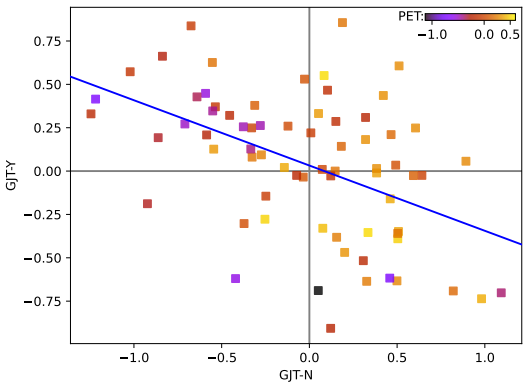

Figure 3: Effect estimation means of individual students (points) in interaction with answer type (x=GJT-N, y=GJT-Y, color=PET). There is a weak negative correlation between being good at GJT-N and GJT-Y answers (blue line, R^2=0.23) and a weak positive correlation between GJT-N and PET skills (R^2=0.20).

types vary in their overall difficulty (fig. 2), but the wide spread in individual student differences (fig. 3) suggests that the skill sets required (let's call them "sensitivity" and "specificity") are somewhat complementary to each other and tend to be distributed unevenly among students. Each student has a unique combination of them. We discuss this further in §6.2.

LM probabilities. The model was trained on GJT data only. Recall from §3.3 that GJT testing prompts did not explicate the target preposition or even mention the word 'preposition'. All else equal, it is thus conceivable that, despite the preposition-focused training, students evaluate the sentences' grammaticality for reasons unrelated to the target preposition. However, we can with high probability rule out this option as our model estimates strong effects for numerous features directly related to the preposition, namely: p_{tgt} by itself (d=4.57; p<0.0001***); interaction p_{tgt}:p_{ctx} (d=14.92; p<0.0001***);[7] and spatial vs. abstract usage of each preposition form and function (fig. 1, fig. 6 in appendix A). Furthermore, due to the heavy interaction between LM probabilities and categorical cue properties,[8] the singular random effect of spatial vs. abstract usage decreases when the model considers the LM-based fixed effects (d=0.372; p=0.0093**) compared to when it does not (d=0.491; p=0.0006***, fig. 1).

5.2 Predicting Student Responses

Setup. We train the BLM and MLP using a training:evaluation:development data split ratio of 84:15:1, placing less weight on the dev set since it is only used to determine early-stopping during MLP training. Experiments are run 10 times with random data splits and model initializations.

Results. As shown in table 3, both models easily outperform simple baselines, and the two models' overall accuracies are roughly on par (within each other's stdevs) with a slight advantage for the BLM. For predicting GJT outcomes only, the aforementioned interaction between students and answer types is most crucial, followed by information about the target preposition (BLM) and instruction (MLP), respectively. The LM-based features p_{tgt} and p_{ctx} are useful for both models,

	GJT + PET		GJT only	
	BLM	**MLP**	**BLM**	**MLP**
Uniform BL	49.7 ±1.1		49.7 ±1.2	
BLM prior BL	49.7 ±2.1		48.2 ±1.4	
Majority BL	64.2 ±0.9		68.1 ±0.7	
Full model	<u>72.6</u> ±1.1	71.5 ±0.6	<u>72.5</u> ±0.8	71.3 ±0.9
– students	−2.2 ±0.6	−0.9 ±0.7	**−2.6** ±0.9	**−2.0** ±0.8
– answer	**−5.6** ±0.8	**−4.6** ±0.6	−2.4 ±0.8	**−2.0** ±0.8
– fxn & usage	−5.4 ±1.0	**−4.6** ±1.0	−1.5 ±0.4	−0.8 ±1.3
– instr & time	−2.1 ±0.9	−1.8 ±0.9	−0.4 ±0.7	−1.4 ±0.9
$-p_{tgt}$ & p_{ctx}	n/a	n/a	−0.9 ±0.9	−0.4 ±0.9

Table 3: Baselines (BL), BLM and MLP prediction performance, and feature ablation (student response correctness prediction accuracy in %). Means and standard deviations over 10 random seeds, which affect not only model initialization but also data splitting and shuffling. Best full model results on each data split are underlined; highest-impact features in each column are bolded.

but less so than the categorical ones. This is somewhat unexpected based on their strong effect sizes (§5.1) and the overwhelmingly high performance of LMs on other tasks. A potential reason is the contrast between the LM reflecting a gross average of language use—which indeed correlates with grammaticality (R^2 = 0.48, fig. 4)—and the unreliability of student judgments, especially at the pretest and in the control group (fig. 1 top). The lack of stimulus sentences (and thus LM probabilities) in the PET further increases the importance of the answer, form-function, and usage features in the GJT+PET condition. We also see a larger ablation effect of the instruction and time features, which is consistent with the larger interaction effect estimates for the PET (fig. 1 bottom).

5.3 Qualitative Analysis of Stimuli

We take a closer look at individual stimuli in fig. 4. From the y-axis distribution in the center and right panels we can clearly see the learning development among students undergoing preposition-focused training. At the pretest (center), aggregate students' grammaticality judgment is less decisive (mostly vertically centered around $50\% \pm \approx 20pp$. At the posttest (right), the spread is much more decisive, ranging from almost 0% to 100%. At both points in time, there is a slight bias towards positive judgment, i.e., students are generally more willing to accept ungrammatical stimuli as grammatical than to reject grammatical ones. In contrast, LM probabilities (x-axis) tend to err on the conservative side, i.e., the LM has higher recall on recognizing ungrammatical items, whereas students have higher

[7] While p_{ctx} by itself is only very weakly correlated with either grammaticality or student response, it does become a useful predictor in interaction with p_{tgt} (cf. fig. 4 left).

[8] Linguistic categories may to some extent be encoded in the LM's distributed representations (Jawahar et al., 2019).

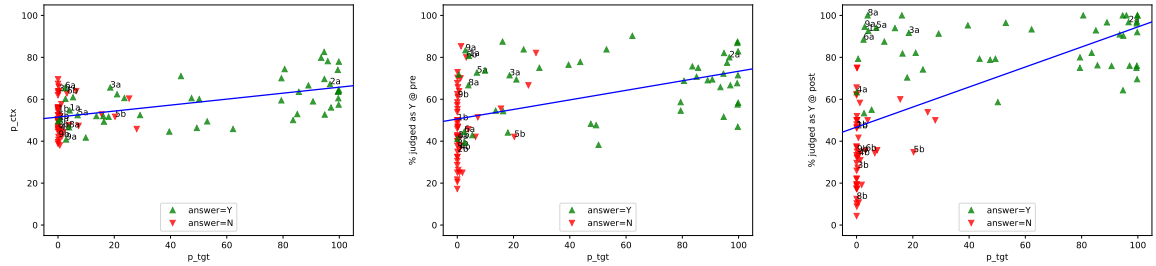

Figure 4: Correlations of LM probabilities, student grammaticality judgment %, and intended answer (color/shape) for individual stimuli (points). **Left:** p_{tgt} (x) and p_{ctx} (y); **Center:** p_{tgt} (x) and pretest judgment (y); **Right:** p_{tgt} (x) and posttest judgment of non-control groups only (y). R^2(answer, judge@post)=0.72; R^2(p_{tgt}, answer)=0.48; R^2(p_{tgt}, judge@post, blue line right)=0.41; R^2(p_{tgt}, p_{ctx}, blue line left)=0.30; R^2(answer, judge@pre)=0.28; R^2(p_{tgt}, judge@pre, blue line center)=0.22; R^2(p_{ctx}, answer)=0.15; R^2(p_{ctx}, judge@post)=0.13; R^2(p_{ctx}, judge@pre)=0.04. Data points/sentences listed in table 1 and discussed in §5.3 are labeled.

recall on recognizing grammatical items, each at the cost of precision.[9]

We expect that intended-grammatical (✓) usages generally receive higher LM probabilities (Δp) than intended-ungrammatical (✗) usages. This is the case most of the time (for 41/48 stimulus pairs total), except for 7 cases, 6 of which involve the preposition '*over*' as the target. We present these sentences in table 1, along with 3 examples where both Δp's are as expected.

What makes ex. 4 – 9 special? A potential explanation is that the verb+preposition+object constructions in ex. 1 – 3 seem to be more clearly distinguishable as either grammatical or ungrammatical than the rest. In contrast, the ✗ sentences in ex. 4 – 6 are not *truly* ungrammatical. The scenarios they describe are unlikely but possible, and the unlikeliness mostly arises through the full-sentence context rather than the prepositional construction alone. In fact, each alternative preposition in 4b, 5b, and 6b might in isolation be a *more* expected collocation with the verb than '*over*', which would explain the p_{tgt} trend. Ex. 7 – 9 (both ✗ and ✓) describe much more rare (i.e., unlikely as far as the distributional LM is concerned) scenes, which may lead to the overall lower p_{ctx} values.[10]

[9]Note that LM probabilities are not based on a binary grammaticality decision but on a selection decision over the entire vocabulary, and also that gradient linguistic judgments in general cannot be said to *only* revolve around grammaticality (cf. Lau et al., 2017). We could address this by looking at the ratio between the probabilities for each pair, but that would in turn establish a dependency among stimuli within each pair which is not present in the human experiment—each stimulus is presented in isolation, in randomized order. Thus, for transparency, we stick with the plain probability and elaborate qualitatively on the expected behavior below.

[10]A second tendency may lie in the concreteness and perceived simplicity (both in terms of semantics and register) of the preposition-governing *verbs*: '*hang, watch, fall*' are all fairly concrete, unambiguous, and colloquial, whereas '*reach, diffuse, stretch, sweep*' have more specialized meanings and are somewhat higher register.

6 Discussion

6.1 Which model type is most appropriate?

For the purpose of our study, the Bayesian logistic model of student responses has clear advantages over both the previous frequentist analysis of score aggregates (complexity of interactions, intuitiveness; §5.1) and the neural response classifier (higher interpretability with roughly equal prediction accuracy; §5.2). However, while this observation is in line with both our expectations and recent literature in SLA (e.g., Norouzian et al., 2018, 2019), we still recommend testing model practicability on a case-by-case basis. For example, if much more training data is available, a neural classifier is likely to outperform a sparse model at prediction accuracy. Whenever the BLM and ANOVA agree on a feature's significance (and they usually—but not always—do), the BLM's estimates are relatively amplified (§5.1). This can be useful for *identifying* potentially relevant effects and interactions, but should also be taken with a grain of salt as it sometimes may construe results too optimistically. Where do these divergences come from? We hesitate to make any strong statements about broad philosophical differences between Bayesian and frequentist statistics in the abstract. Rather, we suspect that it mostly comes down to practical considerations like framing model and data around individual item responses vs. aggregate score, as well as varying degrees of commitment to latent sampling and optimization. Item response prediction accuracy and ablation analyses give some in-

sight into how individual features affect models' estimates of the outcome variable and is consistent with statistical analyses (§5.2). This is particularly useful for discriminative neural models such as our MLP classifier, and is, of course, common practice in NLP classification studies. However, it is also much more costly, less precise, and less reliable than Bayesian and frequentist approaches.

6.2 Implications for SLA

Our analysis of answer types and student aptitudes (§5.1 and §5.2) confirms Wong's (2022) and others' findings about differences between productive and receptive knowledge. We support Wong's argument that the type of assessment should align with both instruction type and and intended learning outcome. We further observe that even within the generally *receptive* task of grammaticality judgment, the subtask of ruling out ungrammatical items (GJT-N) requires higher specificity than accepting grammatical ones (GJT-Y) and is thus more closely aligned with *productive* tasks (e.g., PET). Interestingly, students who are better than average at productive tests tend to be slightly weaker than average at receptive ones and vice versa. A potential future use case of explicitly modeling students' individual differences w.r.t. different task types and linguistic items is that educational applications can be tailored to their *weaknesses*, which is expected to increase learning effectiveness and efficiency.[11] Outside of directly deploying learning technology to end users, our findings can inform educators and SLA researchers. For example, unexpected patterns in LM probabilities (§5.3) may point to suboptimally designed stimulus pairs. Thus, LM probing could be a useful tool in cue selection and stimulus design of similar studies in the future.

6.3 Implications for NLP

In this work, we primarily analyze human learner behavior *using* different machine learning models, while in NLP-at-large it is much more common to *analyze* machine learning models w.r.t. a human ground truth. At the same time, our observations that different senses and usages even of the same preposition form heavily affect human learnability are somewhat analogous to previous results in automatic preposition disambiguation (varying model performance for extended vs. lexicalized senses; Schneider et al., 2018; Liu et al., 2019a). Liu et al. also found that LM pretraining improves disambiguation performance, while Kim et al. (2019a) drew attention to differences among various NLP tasks as 'instruction methods'. This is not to say that current LM training practices are necessarily plausible models of human language learning and teaching, but even these high-level similarities in behavioral patterns invite further investigation.

[11] In practice, such a process should ideally be decentralized by training separate models for each student on the client side, to uphold privacy and other ethical standards.

7 Conclusion

Much quantitative research in many areas of linguistics, including SLA, has been relying on the frequentist method for a long time—and for good reasons: It enables strong conclusions about clear hypotheses, closely following the observed data.

Here we compared several alternative approaches to estimating a multitude of potential effects more holistically, namely via IRT-inspired Bayesian sparse models of explicit interactions among facts, neural classifiers of student responses and feature ablation, as well as contextual probabilities of the experimental stimuli obtained from a pretrained language model (§4).

Overall, we were able to replicate previous frequentist findings regarding the difficulty of acquiring the preposition system in English as a second language and the benefits of concept-based instruction (§5.1). Our computational analysis emphasized the increased flexibility and occasionally stronger effect size estimates of IRT and Bayesian models, as well as their natural interpretability compared to neural models with equal predictive power.

We also found novel interactions among task and subtask type, student individual differences, preposition cue and LM contextualization (§5), and discussed them in the broader contexts of both NLP and SLA, hoping to build bridges between the two research communities (§6). As a final takeaway for both fields, the differences between the LM's and students' overall tendencies to accept or reject stimuli (§5.3 and fig. 4 right) could potentially be exploited in both directions: The aggregate distributional grammatical knowledge of an LM could be used to teach students the most accepted usages of prepositions and other function words across a large population of speakers (i.e., improve their specificity), while LMs could learn to be more creative and to utilize humans' intuitive cross-lingual meaning mappings by learning from second-language learner data.

Limitations

Our study and findings are limited to the specific L1–L2 pair of Chinese (Mandarin and Cantonese)–English. Further, the experimental setting we draw our data from is highly controlled, with carefully-chosen lexical items and carefully-designed (length- and distractor-matched) stimulus sentences. While this enables strong statistical conclusions about the data itself, it poses a sparsity problem for most state-of-the-art NLP models, as can be seen even in the small and simple multilayer perceptron we test.

While it would also be interesting to know whether students respond differently to the same instruction type or vice versa, the between-subjects experimental design underlying our data does not allow such a measurement.

We inspect several model types representing a selection of extreme areas of a vast continuum of computational analysis methodologies. Naturally, this means that we cannot go into a lot of depth regarding model engineering and detailed comparison among similar implementations of each type.

Ethics Statement

Student identities are completely anonymized in our analyses and in the data we feed to our models. By locally distinguishing individual students, we do not wish to single out, over-interpret, or judge any individual student's behavior or aptitude, but rather to fit the models to our data as best we can and also to control for spurious patterns that might have been missed during initial outlier-filtering.

Acknowledgments

We thank the anonymous reviewers for their insightful questions and feedback. This work has been supported by Hong Kong PolyU grant 1-YWBW, awarded to the first author, and grant EDB(LE)/P&R/EL/203 of the Hong Kong Standing Committee on Language Education and Research (SCOLAR), awarded to the second author.

References

Douglas G Altman and J Martin Bland. 2011. How to obtain the p value from a confidence interval. *BMJ*, 343.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. In *Proc. of ACL*, pages 370–379, Baltimore, Maryland.

Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. 2018. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*.

Frank Boers and Murielle Demecheleer. 1998. A cognitive semantic approach to teaching prepositions. *ELT Journal*, 52(3):197–204.

Claudia Marlea Brugman. 1988. *The story of over: Polysemy, semantics, and the structure of the lexicon*. Taylor & Francis.

Gerhard H. Fischer. 1973. The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37(6):359–374.

William Gantt, Benjamin Kane, and Aaron Steven White. 2020. Natural language inference with mixed effects. In *Proc. of *SEM*, pages 81–87, Barcelona, Spain (Online).

Rundi Guo and Nick C. Ellis. 2021. Language usage and second language morphosyntax: Effects of availability, reliability, and formulaicity. *Frontiers in Psychology*, 12.

Zeinab Gvarishvili. 2013. Interference of L1 prepositional knowledge in acquiring of prepositional usage in English. *Procedia-Social and Behavioral Sciences*, 70:1565–1573.

Homa B Hashemi and Rebecca Hwa. 2014. A comparison of MT errors and ESL errors. In *Proc. of LREC*, pages 2696–2700.

Dirk Hovy, Stephen Tratz, and Eduard Hovy. 2010. What's in a preposition? Dimensions of sense disambiguation for an interesting word class. In *Proc. of COLING*, pages 454–462, Beijing, China.

Jena D. Hwang, Hanwool Choe, Na-Rae Han, and Nathan Schneider. 2020. K-SNACS: Annotating Korean adposition semantics. In *Proc. of DMR@COLING*, pages 53–66, Barcelona, Spain (online).

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proc. of ACL*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Carina Kauf, Anna A. Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan S. She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2022. Event knowledge in large language models: the gap between the impossible and the unlikely. Preprint arXiv:2212.01488.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019a. Probing what different NLP tasks teach machines about function word comprehension. In *Proc. of *SEM*, pages 235–249, Minneapolis, Minnesota, USA.

Najoung Kim, Kyle Rawlins, Benjamin Van Durme, and Paul Smolensky. 2019b. Predicting the argumenthood of English prepositional phrases. In *Proc. of AAAI*, volume 33, pages 6578–6585.

Anastassia Kornilova, Vladimir Eidelman, and Daniel Douglass. 2022. An Item Response Theory framework for persuasion. In *Findings of NAACL*, pages 77–86, Seattle, Washington, USA.

Michael Kranzlein, Emma Manning, Siyao Peng, Shira Wein, Aryaman Arora, and Nathan Schneider. 2020. PASTRIE: A corpus of prepositions annotated with supersense tags in Reddit international English. In *Proc. of LAW@COLING*, pages 105–116, Barcelona, Spain.

Alexander Kwako, Yixin Wan, Jieyu Zhao, Kai-Wei Chang, Li Cai, and Mark Hansen. 2022. Using Item Response Theory to measure gender and racial bias of a BERT-based automated English speech assessment system. In *Proc. of BEA@NAACL-HLT*, pages 1–7, Seattle, Washington, USA.

George Lakoff. 1987. Women, fire, and dangerous things: What categories reveal about the mind. *Chicago: University of Chicago*.

Yvonne Lam. 2009. Applying cognitive linguistics to teaching the Spanish prepositions por and para. *Language Awareness*, 18(1):2–18.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Ping Li and Yu-Ju Lan. 2022. Digital language learning (DLL): Insights from behavior, cognition, and the brain. *Bilingualism: Language and Cognition*, 25(3):361–378.

Jeannette Littlemore and Graham D Low. 2006. *Figurative thinking and foreign language learning*. Springer.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *Proc. of NAACL-HLT*, pages 1073–1094, Minneapolis, Minnesota.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. Preprint arXiv:1907.11692.

Frederic M Lord. 1980. *Applications of item response theory to practical testing problems*. Routledge.

James A. Michaelov, Seana Coulson, and Benjamin K Bergen. 2023. Can peanuts fall in love with distributional semantics? In *Proc. of CogSci*. Preprint arXiv:2301.08731.

Charles M. Mueller. 2011. English learners' knowledge of prepositions: Collocational knowledge or knowledge based on meaning? *System*, 39(4):480–490.

Akira Murakami and Nick C Ellis. 2022. Effects of availability, contingency, and formulaicity on the accuracy of English grammatical morphemes in second language writing. *Language Learning*.

Reza Norouzian, Michael de Miranda, and Luke Plonsky. 2018. The Bayesian revolution in second language research: An applied approach. *Language Learning*, 68(4):1032–1075.

Reza Norouzian, Michael de Miranda, and Luke Plonsky. 2019. A Bayesian approach to measuring evidence in L2 research: An empirical investigation. *The Modern Language Journal*, 103(1):248–261.

Tom O'Hara and Janyce Wiebe. 2003. Preposition semantic classification via Treebank and FrameNet. In *Proc. of CoNLL*, pages 79–86, Edmonton, Canada.

Rebecca J. Passonneau and Bob Carpenter. 2014. The Benefits of a Model of Annotation. *Transactions of the ACL*, 2:311–326.

Siyao Peng, Yang Liu, Yilun Zhu, Austin Blodgett, Yushi Zhao, and Nathan Schneider. 2020. A corpus of adpositional supersenses for Mandarin Chinese. In *Proc. of LREC*, pages 5986–5994, Marseille, France.

Maja Popović. 2017. Comparing language related issues for NMT and PBMT between German and English. *The Prague Bulletin of Mathematical Linguistics*, 108(1):209.

Jakob Prange and Nathan Schneider. 2021. Draw mir a sheep: A supersense-based analysis of German case and adposition semantics. *Künstliche Intelligenz*, 35(3):291–306.

Adam John Privitera, Mohammad Momenian, and Brendan Weekes. 2022. Graded bilingual effects on attentional network function in chinese high school students. *Bilingualism: Language and Cognition*, page 1–11.

Ines Rehbein and Josef Ruppenhofer. 2017. Detecting annotation noise in automatically labelled data. In *Proc. of ACL*, pages 1160–1170, Vancouver, Canada.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Aviram Stern, Adi Bitan, and Omri Abend. 2018. Comprehensive supersense disambiguation of English prepositions and possessives. In *Proc. of ACL*, pages 185–196, Melbourne, Australia.

João Sedoc and Lyle Ungar. 2020. Item Response Theory for efficient human evaluation of chatbots. In *Proc. of Eval4NLP@EMNLP*, pages 21–33, Online.

Vivek Srikumar and Dan Roth. 2013. Modeling semantic relations expressed by prepositions. *Transactions of the ACL*, 1:231–242.

Anaïs Tack. 2021. *Mark my words! On the automated prediction of lexical difficulty for foreign language readers*. Ph.D. thesis, KU Leuven.

Andrea Tyler. 2012. *Cognitive linguistics and second language learning: Theoretical basics and experimental evidence*. Routledge.

Andrea Tyler and Vyvyan Evans. 2003. *The semantics of English prepositions: Spatial scenes, embodied meaning, and cognition*. Cambridge University Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NeurIPS*, pages 5998–6008, Long Beach, CA, USA.

William Yang Wang, Elijah Mayfield, Suresh Naidu, and Jeremiah Dittmar. 2012. Historical analysis of legal opinions with a sparse mixed-effects latent variable model. In *Proc. of ACL*, pages 740–749, Jeju Island, Korea.

Marion Weller, Sabine Schulte im Walde, and Alexander Fraser. 2014. Using noun class information to model selectional preferences for translating prepositions in SMT. In *Proc. of AMTA*, pages 275–287, Vancouver, Canada.

Man Ho Ivy Wong. 2022. Fostering conceptual understanding through computer-based animated schematic diagrams and cue contrast. *TESOL Quarterly*.

A Effects of Preposition Cues

In the main text, for brevity, we omitted a detailed analysis of the effects of specific combinations of preposition form, function, and usage on student performance. Here we take a closer look at the six types of cues: *in* with the CONTAINMENT function, *at* with the TARGET and POINT functions, and *over* with the HIGHER,COVER, and CROSS functions.

In fig. 5, we see that there is a wide spread among students for each of the cue types, especially at the PET. The fact that these effects are estimated as interactions in addition to the student-level intercepts suggests, again, that students' skill sets are unique, depending on the preposition cue, which is also illustrated for 5 randomly chosen students.

In fig. 6, we see that the difficulty of these six cues varies greatly, depending on both spatial/abstract use and task type. In fact, the difficulty ranking is largely reversed between GJT and PET. As a striking example of this, *at*-TARGET-Abstract and *in*-CONTAIN-Abstract are the easiest cues to judge correctly in the GJT but most difficult to produce in the PET. There exceptions to this trend, too. E.g., *at*-POINT-Abstract is relatively difficult in both GJT and PET. Another interesting observation is that, in the PET, both usages of *over*-HIGHER are much easier to produce than any other cue.

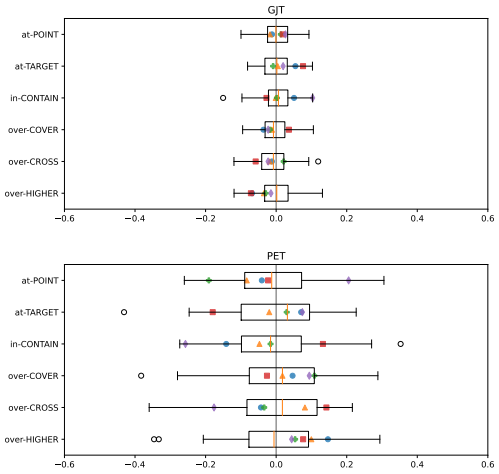

Figure 5: Spread among student effect means (x-axis) in interaction with preposition form/function. 5 randomly chosen students are shown exemplarily (filled shapes; empty circles are outliers). Note that, while in our other figures the error bars denote standard deviations over models' marginal parameter distributions, here they describe the distribution over students of estimated mean interaction effects.

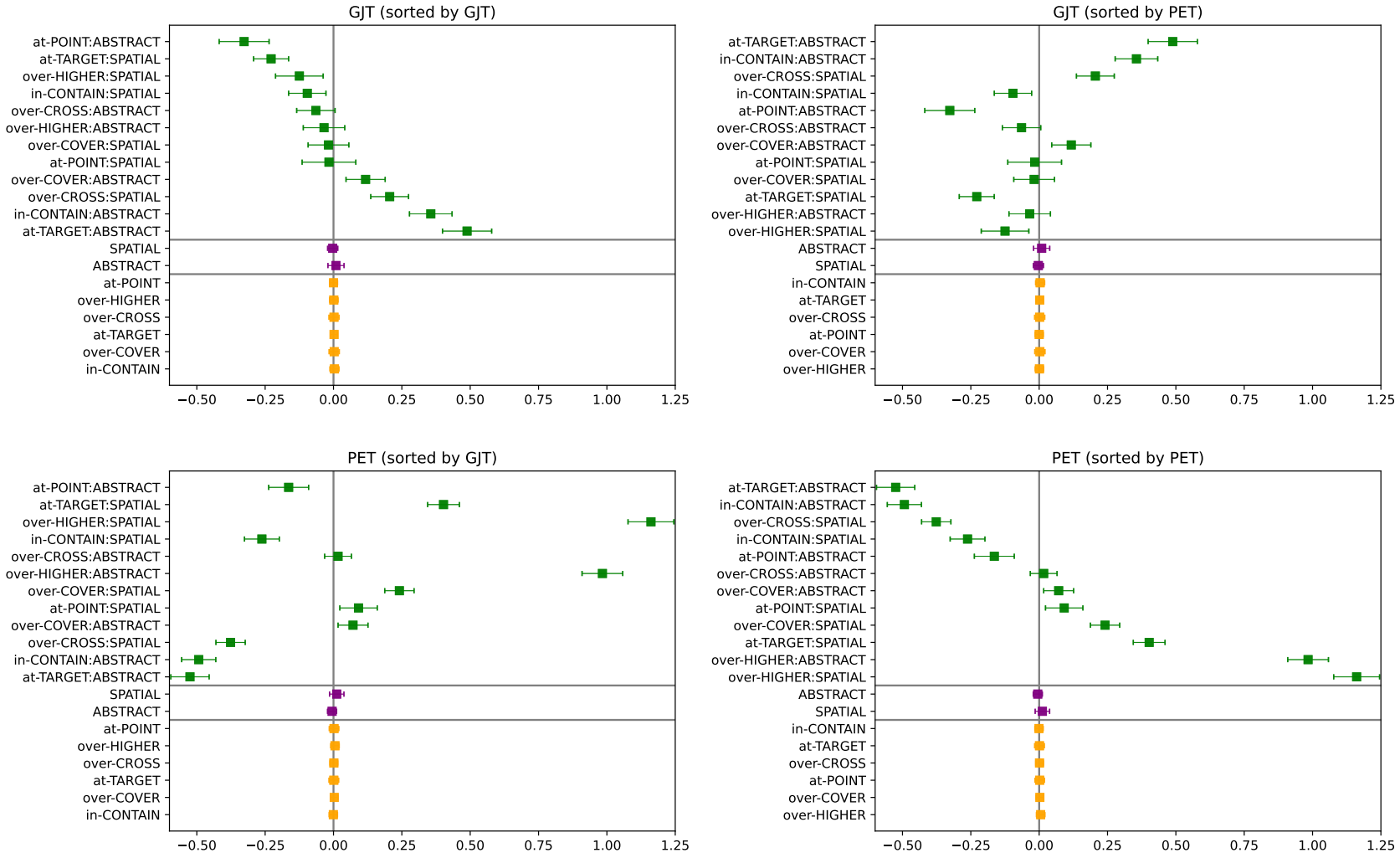

Figure 6: Effect estimates for interactions between preposition form/fxns and spatial vs. abstract usage.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6 Discussion and Conclusions; Limitations Statement

☒ A2. Did you discuss any potential risks of your work?
We conduct a small-scale research and replication study. We will release our experimental software code, but do not deploy any end-user applications.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1 Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

We used experimental data (stimuli and behavioral results) from Wong (2022). This is explained and described in section 3 Original Study and Data.

☑ B1. Did you cite the creators of artifacts you used?
section 3 Original Study and Data and throughout the paper

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The data are currently not publicly available. They were shared with us by the author of the original study, who is also a co-author on this paper.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
section 3 Original Study and Data; Ethics Statement

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
section 3 Original Study and Data

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
section 3 Original Study and Data; section 5 Evaluation

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

section 4 Models; section 5 Evaluation

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
section 4 Models

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. We did not perform extensive hyperparameter search and are not proposing a state-of-the-art model configuration.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
section 5 Evaluation

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
section 4 Models

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

section 3 Original Study and Data; section 5 Evaluation

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
We computationally replicate data analysis and outcome prediction of data that was collected by other researchers. We cite and discuss the relevant publication, which provides detailed information about participants and procedures.

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
See above.

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
See above.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
See above.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
section 3 Original Study and Data