Robust Estimation of Population-Level Effects in Repeated-Measures NLP Experimental Designs

Alejandro Benito-Santos, Adrián Ghajari and **Víctor Fresno**

UNED NLP&IR Group
Department of Languages and Systems
Escuela Técnica Superior de Ingeniería Informática
Universidad Nacional de Educación a Distancia (UNED)
Juan del Rosal 16, 28040 Madrid, Spain
Correspondence: al.benito@lsi.uned.es

Abstract

NLP research frequently grapples with multiple sources of variability—spanning runs, datasets, annotators, and more—yet conventional analysis methods often neglect these hierarchical structures, threatening the reproducibility of findings. To address this gap, we contribute a case study illustrating how linear mixed-effects models (LMMs) can rigorously capture systematic language-dependent differences (i.e., population-level effects) in a population of monolingual and multilingual language models. In the context of a bilingual hate speech detection task, we demonstrate that LMMs can uncover significant population-level effects—even under low-resource (small-N) experimental designs—while mitigating confounds and random noise. By setting out a transparent blueprint for repeated-measures experimentation, we encourage the NLP community to embrace variability as a feature, rather than a nuisance, in order to advance more robust, reproducible, and ultimately trustworthy results.

1 Introduction

Robust statistical modeling, inference, and reporting lie at the core of scientific progress. In this context, Null-Hypothesis Significance Testing (NHST) is a fundamental tool to ensure that observed improvements in NLP models are not merely coincidental. In recent years, the NLP community has placed greater emphasis on empirical results and model comparisons, increasing the need for sound significance analysis. However, studies have found that significance testing in NLP is often inconsistent or improperly applied. For example, a survey of ACL/TACL 2017 papers showed many works ignored significance tests or used them incorrectly (Dror et al., 2018). This is problematic because without rigorous significance evaluation, we risk mistaking random fluctuations for genuine improvements, leading to a waste of valuable time and computing resources.

A long-standing concern in experimental linguistics is the multilevel, hierarchical, and nested nature of language data, which can undermine naive significance tests. Clark (1973) famously demonstrated the "language-as-fixed-effect" fallacy, showing that treating specific stimuli (e.g. a sample of words or sentences) as fixed can yield misleading conclusions. In his example, two researchers obtained statistically significant yet contradictory results on the same hypothesis simply by using different word samples. The fallacy arises because each treated their 20 words as if they were the entire population, rather than a random sample. The conclusion that can be drawn from this example is that ignoring nested structure – such as items within a dataset, annotator-specific biases, or model parameterizations – can inflate Type I error rates. This insight carries over to NLP experiments: if we evaluate a new algorithm on a single test set or a fixed set of instances, we might overgeneralize significance to all possible texts or contexts.

Far from being resolved, modern NLP systems further accentuate these concerns. For example, it is known that neural network models introduce additional randomness (from weight initialization, data shuffling, etc.), meaning that repeated runs can yield different outcomes. Empirical studies have confirmed a reproducibility crisis in NLP (Belz et al., 2023), which translates into the general observation that many supposedly superior results fail to repeat under minor experimental variation (Xue et al., 2023). For example, Reimers and Gurevych (2017) showed that merely changing the random seed can produce statistically significant differences in model performance ($p < 10^{-4}$). In their study, the authors show that for two state-of-the-art NER systems, a lucky seed can boost the F_1 score by about 1% – enough to turn a mediocre

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 33076–33089
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

model into the perceived leader. Such variability means a single reported score is an unreliable basis for comparison, and improvements might not reliably reproduce. Berg-Kirkpatrick et al. (2012) found that the threshold of improvement required for significance depends on factors like test set size and system similarity. They noted that more similar systems can achieve significance with smaller gains (since their outputs are correlated), and that increasing test set size improves statistical power but with diminishing returns. These issues highlight why hierarchical modeling of repeated-measures experimental data (i.e., data arising from experiments where the same subjects are observed *multiple times* under different conditions), and careful experimental design are crucial to account for the existing variability at multiple levels (runs, datasets, annotators, etc.), making conclusions more robust. In this study, we contribute a case study aimed to illustrate *methodological best practices* for the statistical analysis and reporting of repeated-measures experimental data in NLP. We do so in the context of a cross-lingual evaluation task, and 2) limited resources (small-N experimental designs). Here, we employ linear mixed-effects models (LMMs) to investigate systematic differences in performance favoring evaluation in English, which is an effect documented in the literature (Conneau et al., 2018).

2 Methods

Drawing from other fields of science, recently NLP researchers have turned to hierarchical models and multi-level experimental designs to expand the field. At the core of these approaches are linear mixed-effects models (LMMs), a class of regression models that include both fixed effects (systematic influences of interest) and random effects (group-specific variations) (Bates et al., 2015). The general form of an LMM can be expressed as:

$$y_{ij} = X_{ij}\beta + Z_{ij}b_j + \epsilon_{ij}, \tag{1}$$

where y_{ij} is the observed dependent variable for the i-th observation of the j-th grouping factor, X_{ij} is a matrix of fixed effect predictors with associated coefficient vector β, Z_{ij} is a matrix of random effect predictors with corresponding random effects b_j, and ϵ_{ij} represents the residual error term. Both the random effects b_j and the residuals ϵ_{ij} are assumed to follow a normal distribution $b_j \sim \mathcal{N}(0, \Psi)$, and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

By modeling random effects, LMMs capture the idea that our data points are often grouped – e.g. multiple sentences drawn from the same document, or multiple judgments from the same annotator – and that these groups constitute samples from a larger population. As we illustrate in this paper, LMMs offer significant advantages over simpler statistical tests such as paired t-tests or repeated-measures ANOVA (RM-ANOVA) in the context of NLP research where their adoption is still limited —although notable exceptions to this norm exist, especially emerging from the human evaluation subfield— (Pavlick and Kwiatkowski, 2019; Howcroft and Rieser, 2021). In a different vein, our work expands on previous studies that employ LMMs to measure systematic variation in model outputs (Søgaard, 2013; Gantt et al., 2020) and apply it under a cross-lingual evaluation paradigm. Following this trend, in more recent developments Sanchez Carmona et al. (2025) have also endorsed the use of mixed-effects modeling within the NLP domain. Nevertheless, our methodology diverges considerably in its scope, motivation, and application. Their study utilizes cross-classified models to analyze test score variability across various models, datasets, and fine-tuning configurations within the context of extensive biomedical benchmarking. Conversely, our research concentrates on repeated-measures designs typical of small-N NLP experiments, highlighting the robust estimation and interpretation of factorial interactions and population-level effects through the application of linear mixed-effects models (LMMs). Rather than isolating nuisance effects, we demonstrate how structured variability can be leveraged as a signal to derive reliable inferences within constrained experimental settings. These two contributions are complementary, each addressing distinct methodological challenges with the overarching objective of enhancing statistical rigor in NLP evaluation. Furthermore, our work covers the existing gap in the NLP literature on the interpretation and reporting of regression models displaying strong interaction effects, a circumstance that can severely complicate the analysis task. Within this framework, we demonstrate that LMMs provide a robust mechanism for unraveling the complex influences inherent in research inquiries associated with a *population of language models*. This represents a pivotal theoretical aspect of our study, which, to the best of our knowledge, is unprecedented in its scope and approach.

3 Case Study

3.1 Motivation

To illustrate the ideas presented above, we set up a case study aiming to measure a relevant effect documented in the NLP literature in a population of encoder-only language models [1]: the language-dependent effectiveness gap; this is, the performance decrease that arises when using language models in other languages different than English (Conneau et al., 2018; Ranasinghe and Zampieri, 2020). Particularly, we want to investigate whether this theoretical effect can be observed in a population of encoder-only models operating in a set of well-defined evaluation tasks. Particularly, we decided to focus on the English-Spanish pair. Although both languages are amongst the most popular in the world, NLP systems generally yield more robust outcomes on English, largely due to the more extensive availability of high-quality annotated corpora and evaluation benchmarks for English relative to Spanish (Agerri and Agirre, 2023). To ensure that we could effectively measure this effect, it was necessary to find a dataset with the following desirable characteristics:

The dataset was selected to ensure comparability between its English and Spanish sections while avoiding biases introduced by automatic annotation translation (Jalota et al., 2023). Additionally, to mitigate data contamination—where pre-training on test data biases model evaluations (Riddell et al., 2024; Jacovi et al., 2023; Ranaldi et al., 2024)—a dataset with strict test-data privacy was chosen [2]. Furthermore, a hierarchical structure was required to facilitate a crossed repeated-measures design, allowing for multiple observations per model and item while enabling robust statistical analyses that account for inter-model variation and cross-linguistic differences (Barr et al., 2013; Conneau et al., 2018). Based on these criteria, the sEXism Identification in Social neTworks (EXIST) shared task dataset was ultimately selected.

3.2 Dataset Description

Developed as part of the CLEF 2023 conference, this dataset is a significant resource for detection of sexism and hate speech in social media. The dataset comprises 10,034 annotated tweets in both English (5,307) and Spanish (4,727). Each language portion is split into training (70%), development (10%), and test (20%) sets to ensure a robust evaluation framework. A key feature of the dataset is its diverse annotations: 1,065 annotators from 45 countries contributed to the labeling process, providing a wide range of perspectives on sexism. The dataset construction also aimed to mitigate different kinds of biases (i.e., seed, terminology, and temporal) through a careful selection of seed terms in both languages, random sampling across different time periods, and inclusion of diverse content sources. EXIST 2023 consists of three classification sub-tasks: binary (`t1`), ternary (`t2`), and hierarchical multilabel (`t3`). Each task supports evaluation in two distinct `evaluation modes` (hereafter, simply `modes`): `hard` and `soft`. In the `hard` `mode`, the gold standard is comprised of majority-voted labels, whereas the `soft` `mode` utilizes labels that represent the empirical class probabilities derived from human annotations, which allows for more nuanced model training and evaluation under the "learning with disagreements" (LeWiDi) paradigm (Uma et al., 2021). Specific details on each of these tasks can be found in Appendix A.

3.3 Experiments

To investigate our research question —*"Do language models evaluated in English perform better than their Spanish counterparts in the context of the EXIST 2023 shared task?"*—, we selected a sample of three popular encoder-only architectures (i.e., BERT, RoBERTa and DistilBERT) from the literature. For each architecture, we chose representatives that were pre-trained either in a) English, b) Spanish, or c) multilingual data, yielding a total of nine unique models: six monolingual (three in English and three in Spanish) and three multilingual (details on each of these models are shown in Appendix D). Monolingual models were evaluated on their relevant language portion of the dataset (English or Spanish), whereas multilingual models were tested on both. Therefore, our experimental setup involved fine-tuning each model on every possible combination of `language`, `task` and evaluation `mode` using the corresponding test data for each model. This resulted in six unique conditions that were *seen* by each model (t1-hard, t1-soft, t2-hard, etc.). We employed a simple grid search to identify per-model best configurations for fair comparison across conditions: the grid search was

[1] We focused on encoder-only models due to their prevalence in text classification and the documented presence of language-specific effects in this architecture class.

[2] Test data was kindly provided upon request by the task organizers.

performed on a hyperparameter space of 12 elements: batch size (16 and 32), learning rate (1e-5, 3e-5, 5e-5), and weight decay (0.1 and 0.01). During model training, we employed two distinct loss functions, depending on the nature of the classification task. For single-label classification tasks (task 1 and task 2), we used cross-entropy loss, and for multi-label, binary cross-entropy loss. In the soft case, we leveraged binary cross-entropy with logits loss to optimize over probability distributions rather than discrete labels. Inference was achieved employing the argmax for single-label classification and thresholding ($p > 0.5$) of sigmoid outputs for multi-label classification. Training was performed on a single NVIDIA 4090 RTX GPU, investing a total of 144 GPU hours.

System performance was evaluated using the Information Contrast Metric (ICM) [3] (Amigó et al., 2020), considering both its hard and soft variants (Plaza et al., 2023). ICM is an information-theoretic similarity measure that generalizes both pointwise mutual information (PMI) and Tversky's linear contrast model which has demonstrated strong performance across a wide range of classification tasks (Amigo and Delgado, 2022). Using this metric, for each model, the optimal configuration was selected based on the parameter set that maximized performance across all possible combinations of task, mode, and language, as measured on the evaluation set. Finally, the best configuration was evaluated on the test set, yielding a raw ICM system score s_{ICM}.

Since ICM quantifies information, its values are defined over the range $(-\infty, \infty)$. Accordingly, all scores were normalized within the interval $(-\mathcal{G}\text{ICM}^{en/es}, \mathcal{G}\text{ICM})$, where $\mathcal{G}_{\text{ICM}}^{en/es}$ represents the gold standard scores for the English or Spanish test sets (i.e., the oracle classifier's scores). The corresponding gold standard scores are provided in Table 4.

3.4 Cross-Lingual Comparison

Since the EXIST task is purely bilingual (i.e., examples on each language portion are unique to that language and not simply translations of their counterparts in the other language), we needed to account for potential differences in difficulty to be able to establish a fair cross-lingual comparison of system performance. To this end, we computed non-linguistic baselines for each task-mode via a simple neural network baseline, which allowed us to isolate the impact of linguistic knowledge from task-specific challenges and measure real performance gains that are not dependent on the particular difficulty of the task in each language. The baselines were identically generated for English and Spanish in PyTorch using a simple neuronal network consisting of two fully connected layers with 128 hidden units and a ReLU activation function. The input textual data was lightly cleaned, tokenized, and ultimately vectorized using TF-IDF and 10,000 features in both cases [4]. The networks were trained for 20 epochs employing the same loss functions as in the LM fine-tuning stage described earlier. To ensure robustness, the final baselines were calculated by averaging over 10 runs and normalized using the intervals depicted on the right of Table 4. The stability of the baselines was high, with a mean standard deviation across task and modes of 0.0038. To make system performances comparable in a scale-free format, we adopted a "fraction of headroom" measure, which gauges how much of the remaining possible improvement each system achieves, thus making comparisons more equitable between languages and tasks whose baseline levels may diverge. Consequently, even if English or Spanish begins at a relatively stronger baseline, this measure preserves the notion of the fraction of "what is left to gain" achieved by a system in a given language-task-mode triplet. Formally, we define the fraction of headroom Δ^{h} as:

$$\Delta^{\text{h}} = \frac{s_{\text{ICM}}^{norm} - b_{\text{ICM}}^{norm}}{1 - b_{\text{ICM}}^{norm}} \tag{2}$$

where s_{ICM}^{norm} is the normalized system ICM score on a given task-mode pair and b_{ICM}^{norm} is the normalized baseline ICM score on the same pair.

3.5 Results

After completing the grid search, we collected 36 unique data points (3 architectures × 2 languages × 3 tasks × 2 modes) on each evaluation scenario (either monolingual or multilingual) from six $N_{mono} = 6$ different subjects (i.e. model pretrains) in the first case, and three $N_{multi} = 3$ in the second one. As it can be seen in the charts of Figure 1, both scenarios demonstrate important interaction effects between the different levels of the three experimental conditions.

[3] ICM was used as it is the official competition metric; however, our modeling framework generalizes to any continuous score.

[4] We employed NLTK 3.8.1, PyTorch 2.2.0 and scikit-learn 1.4.1 to build the baselines.

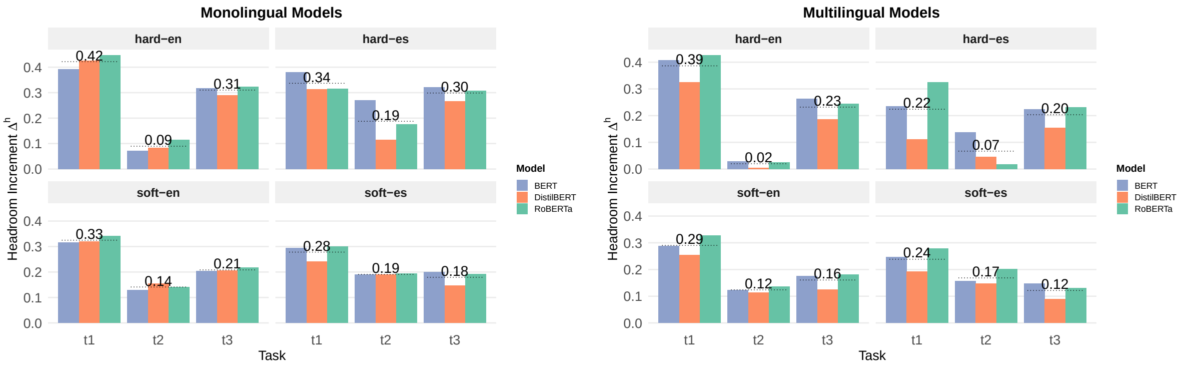

Figure 1: Fraction of headroom Δ^{h} of tested models for each condition in our design. Dotted black lines indicate mean values. Although RoBERTa and BERT seem to perform systematically better than DistilBERT in the multilingual case, the same does not occur in the monolingual portion. This denotes factor interactions that we sought to capture in the model specifications.

For example, while generally evaluations in the `soft` mode yielded lower scores in tasks 1 and 3 than in the `hard` mode, this effect was reversed in `task` 2. Also, evaluations in `tasks` on the English portion seemed to obtain higher scores than their counterparts in the Spanish portion, regardless of the employed evaluation `mode`. Beyond that, it can also be noted that each archiecture (BERT, RoBERTa, DistilBERT) exhibits a relatively stable, consistent effect on performance across tasks and modes. Specifically, while the relative ranking of these classes can shift slightly depending on the language (e.g., for Spanish, BERT and RoBERTa seem closer, whereas in English RoBERTa shows a clearer advantage), overall each class maintains a characteristic offset in performance. To more effectively capture the complexity of these multilevel interactions, we employed two distinct hierarchical (mixed-effects) models: one dedicated to analyzing the monolingual architectures (top of Figure 1) and another focused on the multilingual architectures (bottom).

4 Data Analysis

In this section, we present the methodology that was adopted to analyze the data resulting from the repeated-measures configuration described previously [5]. In face of the hierarchical structure of the data at hand, one common, though ultimately inadequate approach consists in repeatedly applying the paired t-test. Although the paired t-test is a popular and simple method for within-subject comparisons (e.g., English vs. Spanish), it is generally inadequate for repeated-measures data. It assumes independent pairs—a condition often violated in complex designs with repeated evaluations across multiple tasks and modes. Even if independence is forced by averaging observations, this process removes legitimate variability and yields underpowered analyses, a common issue in NLP research (Card et al., 2020). For example, the hierarchical design presented in this paper would require six tests, which would inflate the family-wise error rate to approximately a 27% at the standard significance level $\alpha = 0.05$: $(P = 1 - \alpha^N = 1 - (0.95)^6 \approx 0.27)$. Although corrections such as Bonferroni, Holm, or FDR exist (Dror et al., 2017), *mechanically* applying them in a *mechanical* manner can lead to systematic misinferences (e.g., inflated type II errors) (Nakagawa, 2004), so they should not be relied upon as a universal solution to dependency issues. Furthermore, in small-N designs such as the demonstrated in this paper, these partial tests, having only 2 degrees of freedom, would be markedly underpowered and thus would render the detection of even large effects virtually unfeasible. As we explain in the following sections, other more advanced statistical tests beyond t-tests exist (e.g., repeated-measures ANOVA). However, as we demonstrate in section section 5, such tests are less flexible than LMMs, and are thus better replaced by the latter approach.

4.1 Model Fitting

Building on the methodology described in previous sections, we fit two full-factorial models using the fraction-of-headroom measure Δ^h as the response variable. These models captured all two-way and three-way interactions among the focal

[5] R code to replicate the analyses and figures in this paper is provided as supplemental materials.

factor (language) and the remaining experimental conditions (task and mode). In addition, we specified a nested random effect (1 | architecture / model) to account for the variability of models nested within their respective architectures (i.e., models pre-trained in either English or Spanish). We implemented this modeling approach in R using the lme4 package (Bates et al., 2015), employing the following specification for the monolingual scenario:

```
delta ~
  language * task * mode
  + (1 | architecture / model)
```

which expands to:

$$M_{\texttt{mono}}: \quad \Delta^{ijk} = \beta_0 + \beta_1(\ell) + \beta_2(t) + \beta_3(m) \\ + \beta_{12}(\ell \times t) + \beta_{13}(\ell \times m) + \\ + \beta_{23}(t \times m) + \beta_{123}(\ell \times t \times m) \\ + u_{\texttt{arch:model}} + \epsilon_{ijk}. \tag{3}$$

with $u_{\texttt{arch}} \sim \mathcal{N}(0, \sigma_u^2)$ and $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$

We employed the same specification in the multilingual scenario; however, the nesting term within the random effects structure was omitted due to the presence of only a single representative for each architecture in the experimental data ($N_{multi} = 3$). Consequently, the specification of the random effects was reduced to (1 | architecture). Both models were successfully fitted using restricted maximum likelihood (REML). Subsequent to the initial fitting, a simulation-based evaluation of the models' residuals was undertaken (Hartig, 2024) to ascertain the absence of significant deviations from LMM assumptions of normality and homoscedasticity [6]. For the models fitted under the two scenarios, Q-Q plots and residual variance plots are presented in Appendix B. A summary of each model's estimated coefficients for the fixed effects is shown in Table 1. In the monolingual model, the marginal and conditional R^2 (Bartoń, 2024) values were 0.894 and 0.948, respectively, and in the multilingual model they were 0.800 and 0.899, indicating that both fixed and random effects explained a substantial proportion of the variance in our data.

In the table, it can be seen that in both scenarios that the model fitted negative coefficient estimates for the effect of Spanish language on system performance. Here, the coefficients for ($\ell = \text{es}$) indicate that for the baseline condition ($t = \text{t1}, m = \text{soft}$), Spanish yields a lower fraction of headroom than English. However, positive language–task interactions (especially for task2 and task3) partially compensate for this gap in those tasks. Likewise, a negative coefficient for ($m = \text{soft}$) reflects a performance drop from hard to soft in the reference condition, but in other task–language combinations, this penalty is modulated by the relevant interaction terms. In both cases, other interactions are found to be significant, revealing the multilevel character of the experiment. Note that here, an inexperienced analyst could mistakenly interpret the effect of language ($\ell = \text{es}$) on the observed variable as precisely -0.16 or -0.08 due to a superficial reading of the coefficient estimates. However, this interpretation would be incorrect due to the significant involvement of language in significant higher-order interactions (e.g., Language × Task × Mode). Therefore, the main effect of language on the population is not directly interpretable from the models' regression coefficient β_1 in either scenario. To estimate a global effect, it is necessary to compute the associated marginal means of the effect (i.e., the predicted means of the outcome variable at each level of language, while averaging over all levels of the other factors) along with their associated confidence intervals.

4.2 Estimating Population-Level Effects

To obtain a more holistic view of the influence of language beyond the baseline condition, we estimate the population-level (marginal) effect of language—i.e., the effect of language on Δ^h for the *average* model in our sample. To achieve this, we employ estimated marginal means (EMMs), which provide model-adjusted predictions by averaging over the levels of other categorical factors. This approach ensures that the estimated effect of language accounts for variability introduced by task and mode, rather than being confounded by specific level comparisons. The emmeans R package (Lenth, 2024) implements this estimation by extracting fixed-effect coefficients from the fitted model and computing means over the relevant factor levels, using the variance-covariance structure to quantify uncertainty and produce an estimate [7]. For a given level ℓ of language, the estimated

[6] Facilitated by the R DHARMa package. See https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html for a full description of the package's main functionalities.

[7] Specifically we employed Satterthwaite's method to derive degrees of freedom as in (Howcroft and Rieser, 2021)

Variable	**Multilingual**		**Monolingual**	
	Coeff. (β)	**SE**	**Coeff.** (β)	**SE**
Intercept (β_0)	0.39^{***}	0.03	0.42^{***}	0.02
Language $(\ell = \text{es})$ (β_1)	-0.16^{***}	0.03	-0.08^{**}	0.03
Task $(t = \text{t2})$ (β_2)	-0.37^{***}	0.03	-0.33^{***}	0.02
Task $(t = \text{t3})$ (β_2)	-0.15^{***}	0.03	-0.11^{***}	0.02
Mode $(m = \text{soft})$ (β_3)	-0.10^{**}	0.03	-0.10^{***}	0.02
Language × Task $(\ell = \text{es}, t = \text{t2})$ (β_{12})	0.21^{***}	0.04	0.18^{***}	0.03
Language × Task $(\ell = \text{es}, t = \text{t3})$ (β_{12})	0.13^{**}	0.04	0.07^{*}	0.03
Language × Mode $(\ell = \text{es}, m = \text{soft})$ (β_{14})	0.11^{*}	0.04	0.04	0.03
Task × Mode $(t = \text{t2}, m = \text{soft})$ (β_{23})	0.20^{***}	0.04	0.15^{***}	0.03
Task × Mode $(t = \text{t3}, m = \text{soft})$ (β_{23})	0.03	0.04	-0.01	0.03
Language × Task × Mode $(\ell = \text{es}, t = \text{t2}, m = \text{soft})$ (β_{123})	$-0.11^{\cdot}$	0.06	-0.09^{*}	0.04
Language × Task × Mode $(\ell = \text{es}, t = \text{t3}, m = \text{soft})$ (β_{123})	-0.12^{*}	0.06	-0.06	0.04

Table 1: Fixed effects for the model comparing Multilingual and Monolingual settings, with notation aligned to the model equation. Significance codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$.

marginal mean $\bar{y}_\ell$ represents the expected value of Δ^h when task and mode are balanced across conditions, ensuring an interpretable population-level estimate.

In the M_{multi} model, the EMM for English was found to be 0.202 ($SE = 0.0211$), while for Spanish (es) it was 0.170 ($SE = 0.0211$). The estimated difference between the two languages was 0.0316 ($SE = 0.0114$, $df = 22$, $t = 2.767$, $p = 0.0113$), indicating a statistically significant effect of language in the multilingual portion of our experimental data. In contrast, in the monolingual scenario, the estimated marginal means were nearly identical for the two languages (English: 0.249, $SE = 0.0142$; Spanish: 0.245, $SE = 0.0142$) and the associated estimated difference was minimal (0.0041, $SE = 0.0201$, $df = 4$, $t = 0.206$, $p = 0.8469$), providing no evidence of a language effect in this setting. These results suggest that the effect of language is significant in the multilingual setting but negligible in the monolingual scenario, shedding light on our initial research question.

Following established reporting practices (Schuff et al., 2023; Card et al., 2020), we further evaluated the magnitude of the significant language effect observed in the multilingual scenario by computing partial eta squared (η_p^2) via a Type III ANOVA of the model coefficients. As it can be seen in Table 2, language exerts a large effect on Δ^h ($\eta_p^2 = 0.60$), indicating that it explains a substantial portion of the variance. However, other higher-order interactions, such as $\ell \times t$ ($\eta_p^2 = 0.57$) and $\ell \times t \times m$ ($\eta_p^2 = 0.26$), also play a significant role. These findings underscore the importance of interpreting the influence of language on system performance within the more nuanced context of the EXIST 2023 task.

Table 2: Partial eta squared (η_p^2) effect sizes with 95% confidence intervals for main and interaction effects in the multilingual scenario.

Effect	η_p^2	95% CI
ℓ	0.60	[0.31, 0.76]
t	0.89	[0.78, 0.93]
m	0.35	[0.06, 0.59]
$\ell \times t$	0.57	[0.24, 0.73]
$\ell \times m$	0.26	[0.02, 0.52]
$t \times m$	0.58	[0.26, 0.74]
$\ell \times t \times m$	0.21	[0.00, 0.46]

To further support this analysis, Appendix C presents two alternate modeling scenarios for interested readers, utilizing the same dataset but with varied fixed-effect configurations. These scenarios, which either exclude interaction terms or modify random effects, demonstrate how minor modeling decisions can greatly influence uncertainty, statistical power, and residuals. By contrasting these examples with our primary model, we highlight the crucial role of model design in drawing reliable conclusions from repeated-measures NLP experiments.

5 Robustness of Linear Modeling

To further validate the applicability of linear mixed models (LMMs), we conducted a supplementary experiment on the data from the multilingual scenario. The aim of this experiment was to provide an example of the robustness and flexibility of LMMs in the context of missing data points, a capacity that makes them superior to more rigid tests such as RM-ANOVA. To this end, we systematically removed increasing proportions of the data—specifically, 5%, 10%, up to 50%—and performed 100 fits of M_{multi} at each level, assessing the p-value of comparing the EMMs for the two levels of `language`: $H_0 : \bar{y}_{\ell=en} - \bar{y}_{\ell=es} = 0$. On each case, we annotated the proportion of cases where the null hypothesis was rejected by the corresponding test. The results (Figure 2) indicate that the LMM successfully identified the language effect in 95% of the simulation runs when 10% of the data were removed, and maintained a detection rate of 70% when 15% of the data were omitted. On the contrary, and despite RM-ANOVA was able to find the significant effect of language in the complete data (F(1,2) = 20.778, p = .045), it failed to produce sensible estimates even when the data ablation level was at the minimum (5%). This example illustrates the practical advantages of linear mixed-effects models (LMMs) in NLP experimentation, where missing data frequently arise due to inconsistent data collection, training interruptions, or unanticipated model failures. LMMs demonstrate resilience to data loss, enabling reliable inferences even under less-than-ideal conditions. In contrast, conventional methods such as repeated-measures ANOVA (RM-ANOVA) remove entire groups whenever a single data point is missing, thereby undermining statistical power or even rendering the analysis infeasible.

6 Recommendations

To complete our contribution, we propose the following general methodological recommendations:

1. Embrace variability Our study underscores the importance of explicitly modeling inherent variability rather than relying on aggregation-based approaches that may obscure meaningful differences. LMMs provide a rigorous statistical framework for addressing the hierarchical and crossed structures that frequently arise in NLP experiments, such as repeated runs, multiple datasets, and annotator variability. Compared to conventional methods, LMMs more effectively partition variance, yielding robust population-level inferences even in the presence of small sample sizes or missing data.

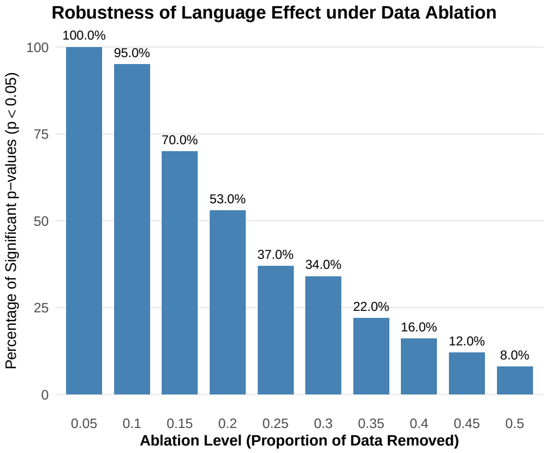

Figure 2: Percentage of significant p-values by number of missing data points. Randomly removing 20% of the data from the model fits reduced the chances of detecting the language effect by approximately 50%.

2. Keep Model Complexity Low Although the inclusion of additional predictors may seem advantageous in achieving a more comprehensive representation of underlying effects, excessive model complexity can hinder interpretability and compromise the reliability of parameter estimates. As demonstrated in this study, the presence of substantial interaction effects necessitates a nuanced interpretation of regression coefficients. Consequently, increasing the number of modeled effects exacerbates the challenge of deriving meaningful insights within a given research framework. A pragmatic approach documented in the literature (Bates et al., 2018), which we followed in this work, is to begin with a parsimonious model that focuses on key fixed effects and a minimal set of essential random effects.

3. Report Model Diagnostics. To ensure the validity of statistical inferences derived from fitting regression models, it is critical to assess whether model assumptions—such as normality and homoscedasticity in the case of LMMs—are adequately met. Nevertheless, in practice, the reporting of residual diagnostics—ranging from graphical assessments to simulation-based techniques—is frequently omitted, despite their potential to reveal assumption violations. Tools such as `DHARMa` enable the detection of overlooked anomalies and facilitate more robust model evaluation. Transparent reporting of these diagnostic checks strengthens the credibility of statistical conclusions and mitigates the risk of drawing inferences based on model misspecification.

7 Conclusion

In this study, we have presented a rigorous methodology for estimating population-level effects from repeated-measures experimental data, demonstrated through a case study rooted in the cross-lingual evaluation of language models. By explicitly accounting for the repeated-measures structure, linear mixed-effects models enable more reliable statistical inferences and help mitigate common pitfalls frequently encountered in the literature, such as inflated Type I error rates and underpowered comparisons. While our case study focuses on cross-lingual evaluation, the methodological insights and recommendations outlined in this work are directly applicable to virtually any NLP domain involving repeated measurements, including model robustness assessments, human annotation studies, and dataset-level comparisons. We hope these findings encourage the broader adoption of hierarchical and multilevel modeling within the NLP community, ultimately fostering more reproducible and interpretable NLP experiments.

Limitations

We acknowledge some limitation in our approach. The bilingual dataset used in our experiments, despite being designed for cross-lingual analysis, may still exhibit unaccounted differences in difficulty, annotation consistency, or domain coverage between its English and Spanish subsets. We addressed this by introducing normalized baselines, but inherent disparities across languages (lexical richness, writing style, label imbalance) can affect system performance independently of any language-specific model shortcomings. For that reason, the findings made in this paper may not be directly translatable to other tasks and domains. The sexism-detection domain chosen here reflects real-world social media data, but may not map directly to other domains such as biomedical text, legal documents, or conversational agents. Domain shifts can alter the distribution of language features and labeling conventions, potentially diminishing the generality of our findings. However, the methodology presented in this paper is still useful for evaluating robustness in hierarchical modeling approaches and can be extended to other multilingual and multi-domain settings. Future work should explore additional language pairs, larger datasets, and alternative annotation frameworks to further validate our proposed methodology. Furthermore, while our case study focuses on fixed factors such as language, task, and mode, other important sources of variation—such as random initializations, pretraining corpus differences, or training data splits—can also be incorporated into LMMs when sample sizes permit (see Sanchez Carmona et al. (2025) for an example). Finally, we acknowledge the use of commercial AI assistant chatbots to assist in the writing of the manuscript.

Acknowledgments

This work has been funded by the European Union - NextGenerationEU through the ‘Recovery, Transformation and Resilience Plan’, by the Ministerio para la transformación digital y de la función pública and by UNED via cooperation agreement C039-21OT, and by the research project "Learning Technologies and Collaborative Systems" (LTCS) from UNED (2009I/PPRO/00).

Alejandro Benito-Santos’ work was supported by grant JDC2022-048408-I funded by MICIU/AEI/10.13039/501100011033 and by “European Union NextGenerationEU/PRTR”

The views and opinions expressed are solely those of the author(s) and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

References

Rodrigo Agerri and Eneko Agirre. 2023. Lessons learned from the evaluation of Spanish Language Models. *Procesamiento del Lenguaje Natural*, 70(0):157–170. Tex.copyright: Copyright (c) 2023 Procesamiento del Lenguaje Natural.

Enrique Amigo and Agustín Delgado. 2022. Evaluating Extreme Hierarchical Multi-label Classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5819, Dublin, Ireland. Association for Computational Linguistics.

Enrique Amigó, Fernando Giner, Julio Gonzalo, and Felisa Verdejo. 2020. On the foundations of similarity in information access. *Information Retrieval Journal*, 23(3):216–254.

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):10.1016/j.jml.2012.11.001.

Kamil Bartoń. 2024. *MuMIn: Multi-Model Inference*. R package version 1.48.4.

Douglas Bates, Reinhold Kliegl, Shravan Vasishth, and Harald Baayen. 2018. Parsimonious Mixed Models. *arXiv preprint*. ArXiv:1506.04967 [stat].

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023. Non-Repeatable Experiments and Non-Reproducible Results: The Reproducibility Crisis in Human Evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3676–3687, Toronto, Canada. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An Empirical Investigation of Statistical Significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With Little Power Comes Great Responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Herbert H. Clark. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4):335–359.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability Analysis for Natural Language Processing: Testing Significance with Multiple Datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics. 47 citations (Crossref) [2023-04-13].

William Gantt, Benjamin Kane, and Aaron Steven White. 2020. Natural Language Inference with Mixed Effects. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 81–87, Barcelona, Spain (Online). Association for Computational Linguistics.

Florian Hartig. 2024. *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. R package version 0.4.7.

David M. Howcroft and Verena Rieser. 2021. What happens if you treat ordinal ratings as interval data? Human evaluations in NLP are even more underpowered than you think. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8932–8939, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore. Association for Computational Linguistics.

Rricha Jalota, Koel Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2023. Translating away Translationese without Parallel Data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7086–7100, Singapore. Association for Computational Linguistics.

Russell V. Lenth. 2024. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.10.5.

Shinichi Nakagawa. 2004. A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology*, 15(6):1044–1045.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694. 26 citations (Crossref) [2023-10-04] Place: Cambridge, MA Publisher: MIT Press.

Laura Plaza, Jorge Carrillo-de-Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2023. Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 316–342, Cham. Springer Nature Switzerland.

Federico Ranaldi, Elena Sofia Ruzzetti, Dario Onorati, Leonardo Ranaldi, Cristina Giannone, Andrea Favalli, Raniero Romagnoli, and Fabio Massimo Zanzotto. 2024. Investigating the impact of data contamination of large language models in text-to-SQL translation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13909–13920, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual Offensive Language Identification with Cross-lingual Embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2017. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. *arXiv preprint*. ArXiv:1707.09861 [cs, stat] version: 1.

Martin Riddell, Ansong Ni, and Arman Cohan. 2024. Quantifying contamination in evaluating code generation capabilities of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14116–14137, Bangkok, Thailand. Association for Computational Linguistics.

Vicente Ivan Sanchez Carmona, Shanshan Jiang, and Bin Dong. 2025. Towards Robust Comparisons of NLP Models: A Case Study. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4973–4979, Abu Dhabi, UAE. Association for Computational Linguistics.

Hendrik Schuff, Lindsey Vanderlyn, Heike Adel, and Ngoc Thang Vu. 2023. How to do human evaluation: A brief introduction to user studies in NLP. *Natural Language Engineering*, pages 1–24.

Anders Søgaard. 2013. Estimating effect size across datasets. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 607–611, Atlanta, Georgia. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

Yan Xue, Xuefei Cao, Xingli Yang, Yu Wang, Ruibo Wang, and Jihong Li. 2023. We Need to Talk About Reproducibility in NLP Model Comparison. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9424–9434, Singapore. Association for Computational Linguistics.

A EXIST 2023: Task Descriptions

1. **Sexism Identification (Task 1):** A binary classification task in which systems determine whether a tweet expresses ideas related to sexism, including explicitly sexist messages, descriptions of sexist situations, or critiques of sexist behavior.

2. **Source Intention Classification (Task 2):** A ternary classification task categorizing the author's intention in sexist tweets into three classes: (a) *Direct*—tweets that are explicitly sexist or incite sexism; (b) *Reported*—tweets narrating a sexist situation experienced by women in the first or third person; and (c) *Judgmental*—tweets condemning sexist situations or behaviors.

3. **Sexism Categorization (Task 3):** A multi-label, hierarchical classification task assigning one or more categories to sexist tweets:

   - *Ideological and Inequality:* Discrediting feminism, rejecting gender inequality, or portraying men as victims of oppression.
   - *Role Stereotyping and Dominance:* Expressing false notions about women's suitability for certain tasks or asserting male superiority.
   - *Objectification:* Presenting women as objects, focusing on physical attributes, or reinforcing unrealistic beauty standards.
   - *Sexual Violence:* Containing sexual suggestions, requests, or harassment of a sexual nature.
   - *Misogyny and Non-Sexual Violence:* Expressing hatred and non-sexual violence towards women.

B Model Diagnostics

To verify that our linear mixed model adequately captures the relationship between log(score norm baseline norm) and the fixed effect of language (while accounting for the random intercepts of class run and task:mode), we employed the DHARMa R package for diagnostic checks. DHARMa uses a simulation-based approach to generate scaled residuals, which allows for straightforward graphical and statistical assessments of normality, homoscedasticity, and outliers. In the figure below, the Q–Q plot on the left compares the observed residual distribution to an ideal uniform distribution, while the box plot on the right summarizes residuals by category. The non-significant Kolmogorov–Smirnov, dispersion, and outlier tests all indicate that the model's assumptions are sufficiently met, suggesting our multilingual model is well-specified for the data at hand.

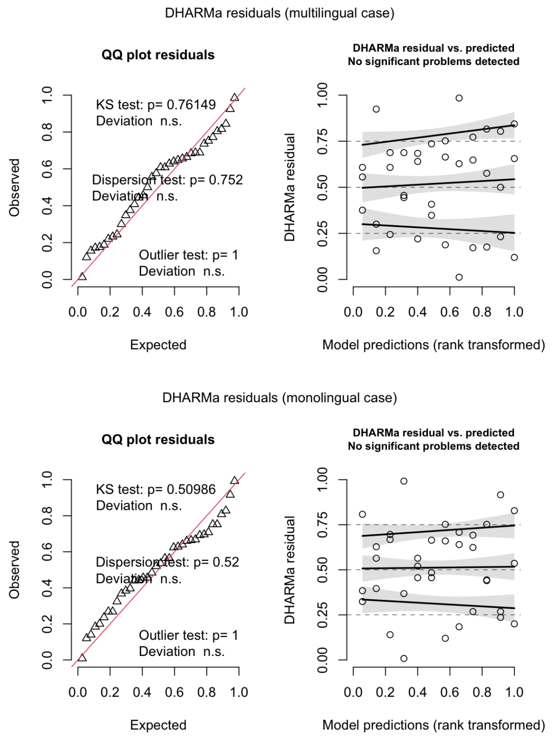

Figure 3: DHARMa simulation-based diagnostic plots for the models fitted to the data in the multilingual (top) and monolingual (bottom) scenarios. The Q–Q plot (left) shows that observed scaled residuals closely follow the expected distribution, and the box plot (right) indicates no significant language within-group deviations. Non-significant Kolmogorov–Smirnov, dispersion, and outlier tests confirm that the models' residuals meet standard assumptions.

C Additional LMM Specification Examples

To complement our main analysis, we present two additional modeling examples using the same dataset. These illustrate common choices researchers face when applying linear mixed-effects models (LMMs) in repeated-measures NLP experiments. By varying how fixed and random effects are specified, we highlight the consequences for interpretability and statistical inference.

C.1 Example 1: Treating Architecture as Fixed, Task as Random

To illustrate the implications of simplifying model structure, we refit a linear mixed-effects model in which `architecture` is treated as a fixed effect and `task` as a random intercept. This scenario reflects a common analytical choice in which researchers seek to compare known model families (e.g., BERT vs. RoBERTa) across a representative sample of tasks, while omitting higher-order interactions involving `task` or `mode`. In this specification, the outcome variable Δ^h is modeled as a function of `language`, `architecture`, and their interaction, with a random intercept for `task` to account for task-level variability. The `mode` factor is excluded entirely from the model, under the assumption that its contribution is either negligible or redundant in the presence of the other terms. The model specification is:

```
delta ~ language * architecture + (1 | task)
```

This simplified model produced a clean fit with no singularities or convergence issues. The estimated marginal means (EMMs) for `language` revealed a mean fraction-of-headroom of 0.202 for English ($SE = 0.056$) and 0.170 for Spanish ($SE = 0.056$), resulting in a mean difference of 0.0316 ($SE = 0.0222$, $t = 1.43$, $p = 0.164$). Although the effect size is identical to that obtained in the main full-factorial model (Section 3.5), the estimated standard error is nearly double, leading to a loss of statistical significance. The model's explanatory power also diminished substantially: the conditional R^2, which accounts for both fixed and random effects, was 0.690, suggesting that despite the random intercept for `task` still captures some meaningful structure, it is insufficient to the purpose of the study.

These discrepancies are not accidental but illustrative of a broader principle in hierarchical modeling: although point estimates for main effects may remain stable across specifications, their associated uncertainty is highly sensitive to model structure. The full-factorial model reduces residual variance by explicitly modeling how `language` interacts with `task` and `mode`. In contrast, the simplified model effectively reallocates all interaction-related variance to the residual term, thereby inflating standard errors and reducing power. The example illustrates how simplifying fixed-effect structure—especially by omitting theoretically plausible interactions—can severely weaken statistical inference, even when point estimates of the variable of interest remain unchanged. For researchers working with repeated-measures designs, particularly in small-N scenarios, such simplifications should be approached with caution. When the research objective involves making population-level claims, retaining a rich interaction structure is often essential for reliable inference.

C.2 Example 2: Omitting All Interactions (Additive-Only Model)

To further explore the impact of model simplification, we fit an additive-only linear mixed-effects model in which all fixed effects are specified without interactions. In this configuration, the outcome variable Δ^h is modeled as an additive function of `language`, `task`, and `mode`, with a random intercept for `architecture` to account for system-level heterogeneity. No interaction terms are included, meaning the model assumes that the effects of each predictor are strictly independent. This approach is frequently motivated by a desire for interpretability and reduced complexity, especially when the sample size is limited. The model specification is:

```
delta ~ language + task + mode
        + (1 | architecture)
```

This model converged without singularity issues and yielded a clean fit according to standard diagnostics. The estimated marginal means (EMMs) for `language` were 0.202 for English ($SE = 0.023$) and 0.170 for Spanish ($SE = 0.023$), resulting in a marginal difference of 0.0316 ($SE = 0.0218$, $t = 1.45$, $p = 0.158$). As expected, the point estimate of the language effect mirrors that of the full-factorial model in the main text (0.0316) once again, but the exclusion of interaction terms leads to inflated uncertainty and a corresponding loss of statistical significance. The marginal R^2 was 0.556 and the conditional R^2 was 0.632, indicating moderate explanatory power largely driven by the fixed-effects structure.

However, in this case residual diagnostics using the `DHARMa` package revealed a notable deviation from model assumptions. While the Q-Q plot and standard dispersion tests did not indicate significant violations, the right-hand quantile plot in Figure 4 showed significant deviations between observed and expected residuals. The red quantile bands, generated from simulated residuals, reveal systematic miscalibration in the model's predictive distribution. This suggests that while the model explains the central tendency of the data reasonably well, it may fail to capture tail behavior or nonlinear dependencies, especially across combinations of `task` and `mode` that are conditionally structured in the full model but ignored here.

These results reinforce the conclusions drawn in Example 1: although point estimates of key effects may remain stable across different model specifications, the omission of experimentally-motivated interactions can lead to underestimation of variability, distort residual structure, and diminish the model's ability to generalize. The additive-only model is particularly vulnerable to this issue because it assumes strictly independent contributions of all factors—a strong assumption rarely justified in complex, factorial designs. The application of simpler, additive-only models to repeated-measures data should therefore be approached cautiously, as those may overlook critical sources of variance and yield misleadingly narrow conclusions.

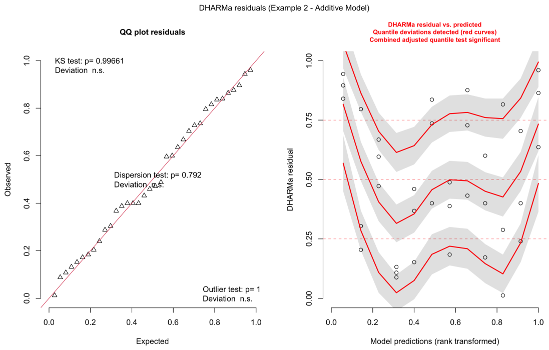

Figure 4

D Additional Tables

Model Name	**Trained In**	**Parameters**
distilbert-base-uncased	English	66M
CenIA/distillbert-base-spanish-uncased	Spanish	66M
distilbert-base-multilingual-cased	Multilingual	134M
bert-base-cased	English	110M
dccuchile/bert-base-spanish-wwm-cased	Spanish	110M
bert-base-multilingual-cased	Multilingual	177M
roberta-base	English	125M
PlanTL-GOB-ES/roberta-base-bne	Spanish	125M
xlm-roberta-base	Multilingual	270M

Table 3: Summary of model architectures, training language, and parameter count.

Task	**Mode**	$\mathcal{G}^{en}_{\mathrm{ICM}}$	$\mathcal{G}^{es}_{\mathrm{ICM}}$	$\mathcal{G}^{en}_{\mathrm{ICM}} - \mathcal{G}^{es}_{\mathrm{ICM}}$	μ_{en}	μ_{es}	$\mu_{en} - \mu_{es}$
t1	H	±0.980	±0.999	0.020	0.571	0.617	-0.046
	S	±3.114	±3.118	0.004	0.436	0.468	-0.032
t2	H	±1.445	±1.601	0.156	0.347	0.411	-0.065
	S	±6.118	±6.243	0.125	0.219	0.251	-0.031
t3	H	±2.040	±2.239	0.199	0.304	0.326	-0.022
	S	±9.126	±9.607	0.482	0.205	0.218	-0.014
Mean	-	-	-	-	0.347	0.382	-0.036

Table 4: Gold standard test set ICM scores (used for normalizing system raw ICM scores) and average baseline scores for English and Spanish datasets.