Não replicação de um estudo de leitura autocadenciada: sinal de uma Crise da Replicabilidade à brasileira?

Rafael Luis Beraldo[1]

RESUMO: Apresentamos duas replicações de um experimento anterior da psicolinguística brasileira nas quais os achados relatados originalmente não foram encontrados. O estudo original relatou um claro efeito na compreensão quando um verbo meteorológico no singular inserido em uma relativa cortadora era precedido por um sintagma preposicionado (como em "Nilton ficou nas planícies que venta continuamente.") versus um sintagma nominal no plural (por exemplo, as planícies), essa última condição exibindo as maiores médias de tempo de reação. O efeito foi interpretado como estranhamento da parte dos informantes, como se

ABSTRACT: We present two failed replications of a previous experiment in the Brazilian psycholinguistic literature. The original study reported a clear effect in comprehension when an uninflected meteorological verb located in a non-standard relative clause was preceded by a prepositional phrase (e.g. "Nilton ficou nas planícies que venta continuamente.") compared to a plural nominal phrase (as planícies). In the nominal phrase condition, higher mean reaction times were observed. This effect was interpreted as the participants' expectation that the verb should agree with the plural nominal phrase. The results of our replications, however, do not

[1] Doutorando no Programa de Pós-Graduação em Linguística da Universidade Estadual de Campinas (Unicamp).

concordância entre o sintagma nominal e o verbo. Entretanto, os resultados de nossas replicações não apresentam esse efeito. Considerando que nossas amostras eram uma ordem de magnitude maiores do que a do estudo original, exploraremos a hipótese de que o efeito original decorreu, na realidade, de uma inflação no tamanho do efeito, uma importante questão identificada na literatura que discute a Crise da Replicabilidade. Sublinharemos também a importância da adoção de maiores amostras e da realização de estudos de replicação.

PALAVRAS-CHAVE: replicação; Crise da Replicabilidade; inflação no tamanho de efeitos; Psicolinguística.

support this effect. Since our samples were larger than the study's by an order of magnitude, we explore the hypothesis that the original effect was, in fact, a byproduct of effect size inflation, a relevant issue in the Replication Crisis literature. We also highlight the importance of adopting larger samples and of carrying out replication studies.

KEYWORDS: replication; Replication Crisis; effect size inflation; Psycholinguistics.

Introdução

A Crise da Replicabilidade na psicologia experimental (PASHLER; WAGENMAKERS, 2012), área na qual a (psico)linguística encontra inspiração para seus métodos, pode ser resumida como a suspeita (IOANNIDIS, 2005) e a observação empírica (OPEN SCIENCE COLLABORATION, 2015) de que muitos estudos, se repetidos, não teriam seus achados replicados. Recentemente o tema tem sido abordado por linguistas (SÖNNING; WERNER, 2021; BIN; MOTA, 2022), porém replicações em nossa área ainda são raras (KOBROCK; ROETTGER, 2023) ou mesmo inexistentes (MARSDEN *et al.*, 2018).

1. Replicações do Experimento 2 de Costa (2013)

Apresentamos duas replicações de uma tarefa de leitura autocadenciada (COSTA, 2013) cujos achados não foram repetidos. Tínhamos como objetivo inicial demonstrar a viabilidade da coleta remota de grandes amostras e, para isso, elegemos um experimento com um claro efeito que, não obstante, não foi observado em nossas replicações. Longe de levantar dúvidas quanto aos protocolos remotos, argumentaremos, juntos a Costa (2022), que a não replicação do original ilustra um importante indutor da Crise da Replicabilidade. Trata-se da inflação nas estimativas de tamanho de efeito, consequência das pequenas amostras habitualmente empregadas em áreas como a psicolinguística (GELMAN; CARLIN, 2014).

1.1 Desenho e resultados do experimento original

Verbos meteorológicos, como *chover*, *trovejar* e *nevar*, são tradicionalmente descritos como impessoais por não admitirem sujeito e, consequentemente, não flexionarem em número. A despeito disso, dados anedóticos levantam a possibilidade de que, na gramática dos falantes do português brasileiro (PB), verbos desse tipo exibem flexão de número (COSTA, 2013). Os exemplos listados pelo autor, que incluem enunciados como "uns verões *chovem* mais, outros menos" (p. 20), foram corroborados em uma tarefa de produção eliciada, na qual verbos meteorológicos flexionados no plural foram produzidos em proporção maior do que o esperado caso resultassem de lapsos de fala (COSTA, 2013, p. 64). O Experimento 2 de Costa, alvo de nossas replicações, investiga se o fenômeno se manifestaria na compreensão.

Uma tarefa de leitura autocadenciada com janela móvel aferiu a sensibilidade dos informantes (n = 32) à flexão de número em verbos meteorológicos dentro de relativas cortadoras (COSTA, 2013, p. 65). Por hipótese, essa sensibilidade se demonstraria em tempos de reação (TR) diferentes nas condições singular (SG) e plural (PL) do verbo. Também foram manipulados os antecedentes do verbo, localizados na sentença matriz, sendo ou sintagmas nominais (NP) ou sintagmas preposicionais (PP). Constituíram-se, assim, quatro condições experimentais (veja o Quadro 1). A variável-resposta de TR foi coletada nas posições P6 (o verbo meteorológico) e na zona de espraiamento P7 (um advérbio).

Quadro 1: Estímulos-alvo do Experimento 2 de Costa (2013) nas quatro condições, com antecedentes do tipo sintagma nominal (NP) ou sintagma preposicionado (PP) e verbos meteorológicos com flexão (PL) e sem flexão (SG). Os verbos e os advérbios estão nas posições finais P6 e P7 dos estímulos.

Condição	Estímulo-alvo de exemplo
NP.PL	Paulo visitou as nações que nevam excessivamente.
NP.SG	Letícia viu as serras que troveja diariamente.
PP.PL	André correu nas praias que chovem repetidamente.
PP.SG	Nilton ficou nas planícies que venta continuamente.

Fonte: Elaboração própria.

Os informantes, alunos da Universidade do Estado do Rio de Janeiro (UERJ), foram expostos a quatro sentenças em cada uma das quatro condições experimentais, perfazendo 16 estímulos-alvo. O triplo de sentenças distratoras (48) foi apresentado, havendo, ao todo, 64 estímulos. Após cada estímulo, perguntas de compreensão com respostas binárias mediam a atenção dos informantes. Anteriormente à fase experimental houve uma breve fase de familiarização, constituída de 10 sentenças.

Um modelo linear de efeitos mistos (MLM) foi ajustado aos dados de TR normalizados na posição P7, tendo como efeito fixo a interação entre tipo de antecedente e número do verbo meteorológico e tendo como efeitos aleatórios informantes e itens (COSTA, 2013, p. 69). Foi encontrado um

efeito (t = −3,869, p = 0,001) na interação entre a condição verbo no singular e a variável tipo de antecedente.

Em resumo, as diferenças no TR não foram estatisticamente significativas nas condições em que o verbo estava no plural (NP.PL e PP.PL). Entretanto, foi encontrada uma interação estatisticamente significativa entre NP.SG × PP.SG (veja o primeiro painel da Figura 1 a seguir). O TR mais alto foi encontrado na condição NP.SG (por exemplo, "as serras que troveja") e o TR mais baixo, na condição PP.SG ("nas praias que chove"). O resultado indica um "estranhamento" (COSTA, 2013, p. 71) quando o verbo está no singular e é precedido por um sintagma nominal plural ("as serras"), como se esse sintagma ocupasse a posição de sujeito do verbo, desencadeando concordância. O mesmo não ocorreu no caso do antecedente preposicionado ("nas serras"), para o qual não se verificaram diferenças quando o verbo estava no singular ou no plural.

1.2 Objetivos e expectativas das replicações

Nosso principal objetivo foi investigar a viabilidade da coleta remota de grandes amostras via web e a qualidade dos dados resultantes, considerando as medidas de distanciamento social impostas pela pandemia de COVID-19. Para atingi-lo, buscamos replicar os achados do Experimento 2 de Costa (2013), coletando dados remotamente, com uma amostra de informantes uma ordem de magnitude maior.

Esperávamos obter achados análogos ao original, isso é, obter um efeito na interação NP.SG × PP.SG. Quando o antecedente era um PP e o verbo estava na condição SG, por exemplo, estimativa do logRT foi igual a −0,23000 no original, ou seja, houve uma redução no tempo de reação associada a essas duas condições.

1.3 Métodos e materiais das replicações e diferenças em relação ao original

Duas replicações do Experimento 2 de Costa (2013) foram conduzidas. Na Replicação 1 (R1), o desenho original foi seguido à risca e os mesmos materiais foram utilizados. Entretanto, o protocolo foi aplicado remotamente, isso é, via web. A diferença é relevante: se no ambiente controlado do laboratório, buscamos maximizar a atenção dos informantes e padronizar o equipamento, a realização de experimentos remotos diminui o grau de controle, adicionando possíveis fontes de distração e de variação de

equipamento. A despeito desses obstáculos, a coleta remota de dados de cronometria mental completou mais de uma década (SPROUSE, 2011) e, aliada à filtragem criteriosa dos eventuais informantes desatentos e ao descarte de observações problemáticas, a qualidade dos dados coletados remotamente pode ser comparável àquela da coleta presencial (RODD, 2024).

A Replicação 2 (R2) foi motivada por uma questão teórica: os estímulos-alvo de Costa (2013) tinham como zona de espraiamento (e, portanto, posição crítica) a última palavra da sentença, conforme ilustrado na Tabela 1. Entretanto, os dados de TR poderiam estar sendo afetados pelo efeito de integralização (*wrap-up effect*) (JUST; CARPENTER, 1980), ou seja, a observação empírica de que a posição final de uma sentença exibe tempos de reação naturalmente maiores como consequência da resolução do processamento. Seria plausível, assim, supor ou que o efeito observado originalmente fosse um artefato da integralização, ou que o tamanho real do efeito pudesse estar sendo mascarado. Desse modo, um objetivo subsidiário foi verificar, por meio de uma replicação suplementar, se o controle da integralização, realizado a partir da adição de duas novas posições finais aos estímulos-alvo originais (Quadro 2), impactaria ou não nos achados.

Quadro 2: Exemplo de estímulo-alvo modificado para separar a zona de espraiamento (P7) do final da sentença. Duas novas posições, ocupadas sempre por preposição e nome, foram adicionadas aos 16 estímulos-alvo originais.

P1	P2	P3	P4	P5	P6	P7	P8	P9
Paulo	visitou	as	nações	que	nevam	excessivamente	no	inverno

Fonte: Elaboração própria.

Ambas as replicações foram implementadas como uma tarefa de leitura autocadenciada com janela móvel no PCIBex (ZEHR; SCHWARZ, 2018) usando as ferramentas padrão desse pacote de *software*, com 16 estímulos-alvo e 48 sentenças distratoras, além de 10 estímulos de familiarização. Ao fim de cada estímulo, uma pergunta de compreensão binária era apresentada, o informante devendo escolher entre "sim" e "não". A taxa de acerto foi um dos critérios para determinar a atenção dos informantes, como veremos a seguir.

1.4 Informantes, critérios de exclusão e descarte

Para a R1, recrutamos informantes nas listas de e-mail da Universidade Federal do Rio Grande do Norte (UFRN). Contamos com 281 informantes, a sua maioria (65%) mulheres, em média com 29 anos (s = 10) e residentes no estado do Rio Grande do Norte (90%). Para a R2, recrutamos informantes em redes sociais e pelas listas de e-mail da Universidade Estadual de Campinas (Unicamp). Ao todo foram 261 informantes, em sua maioria (60%) também se identificando como mulheres, em média com 38 anos (s = 12), desta vez residentes no estado de São Paulo (74%). Excluímos aqueles informantes que responderam ter aprendido outra língua em vez do português ou juntamente com o português durante a infância. Na R1, 2% dos informantes foram excluídos, restando 275. Na R2, essa taxa de exclusão foi maior, atingindo 7% dos informantes, restando 242.

Uma vez que nossas replicações foram realizadas remotamente, é fundamental verificar a qualidade das observações obtidas (SAUTER; DRASCHKOW; MACK, 2020; CRUMP; MCDONNELL; GURECKIS, 2013; SPROUSE, 2011), excluindo informantes com comportamento desviante e descartando dados que deem indícios de desatenção. Como não há consenso sobre os parâmetros para a limpeza dos dados, consideramos os fatores desatenção dos informantes e *distribuição dos TRs* para decidirmos quanto à rejeição dos informantes e ao descarte de observações[2].

A *taxa de desatenção média* foi calculada dividindo as respostas incorretas pelo total de perguntas de compreensão, considerando distratoras e estímulos-alvo. Essa taxa foi igual a 3,2% na R1 e a 2,4% na R2. Excluímos informantes cuja taxa de desatenção superasse os 10%, medida que consideramos conservadora. Como resultado, 5,4% dos informantes foram excluídos da R1 e 2,4% foram excluídos da R2, restando, respectivamente, 260 e 236 informantes. Em seguida, abandonos temporários do experimento foram identificados. Por exemplo, um informante na R1 deixou o experimento por 24 horas, completando a tarefa no dia seguinte. Consideramos como desatento o informante que houvesse demorado 30 segundos ou mais para responder a qualquer estímulo. Esse filtro levou à exclusão de 16 informantes na R1 e de 8 informantes na R2. A medida reduziu consideravelmente o desvio padrão do TR global, sem descaracterizar a sua média.

[2] Os materiais, o código do experimento, os dados e as análises completas podem ser acessados em um repositório na Open Science Foundation: <https://osf.io/pes2j/>.

Por fim, descartamos os dados ainda observando a distribuição dos tempos de reação. Em protocolos psicolinguísticos de cronometria mental, o objetivo quase sempre é encontrar diferenças entre os TRs nas condições manipuladas. Portanto, o descarte de observações com TRs altos é indesejada, pois o efeito estudado pode ter sido sua causa. Por outro lado, TRs muito longos podem indicar desatenção. Observando que havia relativamente poucas observações após os 5000 ms, decidimos por manter observações apenas entre 100 ms e 5000 ms.

Com a exclusão e descarte acima, retivemos 244 informantes (taxa de exclusão de 9%) e 8.436 observações (descarte a 24%) na R1; e retivemos 228 informantes (taxa de exclusão de 5,4%) e 6.656 observações (descarte a 17%) na R2. Esses números estão alinhados com outros experimentos remotos de cronometria mental (SPROUSE, 2011).

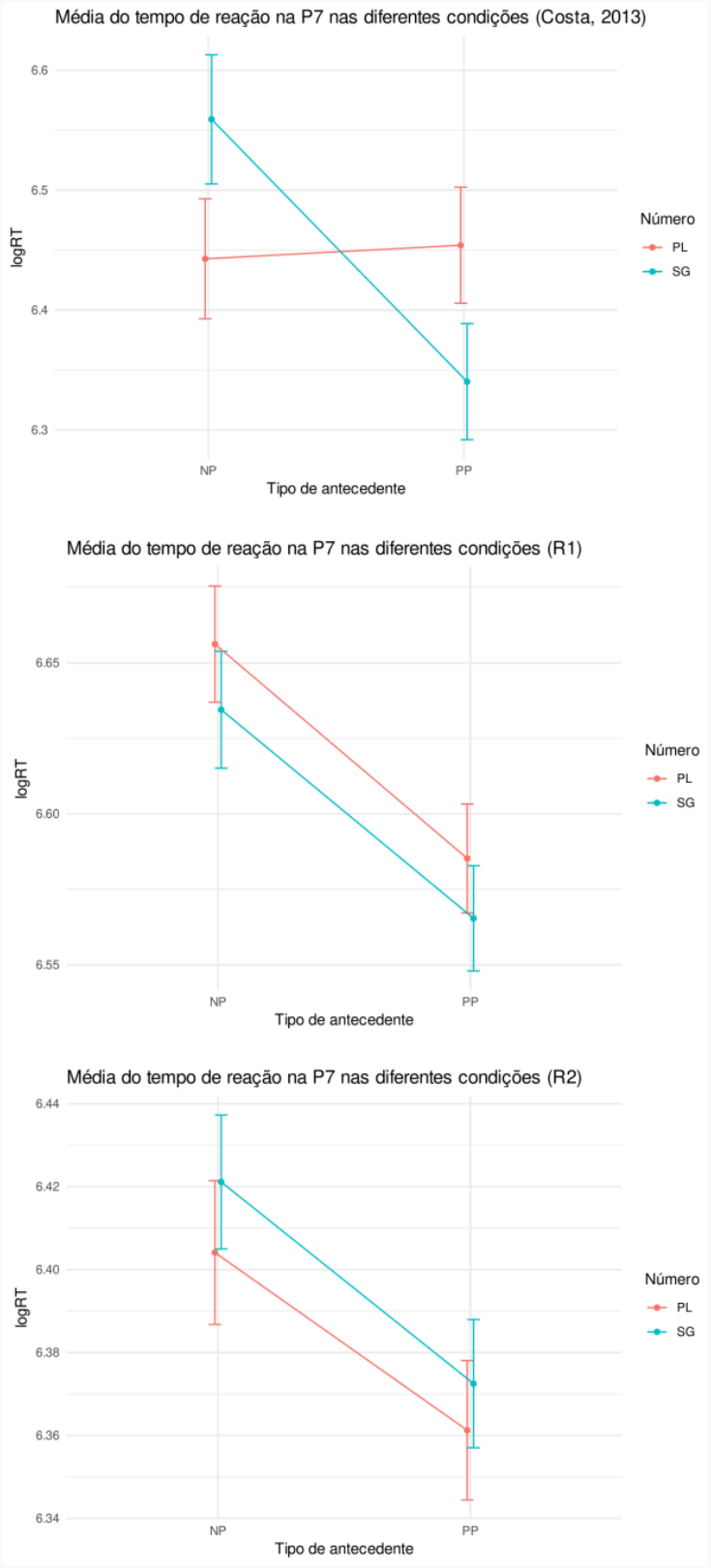
1.5 Resultados das replicações

Investiguemos, descritivamente e por meio de estatística inferencial, os dados de TR na posição crítica (P7). Como veremos, não replicamos o efeito de interação NP.SG × PP.SG, encontrado no Experimento 2 de Costa (2013).

Na Figura 1, o gráfico mais à esquerda apresenta os resultados originais (COSTA, 2013, p. 70), enquanto os outros dois apresentam resultados das replicações. O efeito observado originalmente – a clara interação NP.SG × PP.SG – não foi observado nas replicações, que apresentaram resultados muito similares entre si. Apesar de, na R1, os TRs terem sido menores para os verbos no singular (condição em azul) em relação aos verbos no plural (em vermelho), as barras do erro padrão se confundem, não sendo possível afirmar que haja, de fato, diferenças no TR a depender da condição número do verbo, o que condiz com os achados originais. Parece haver um agrupamento de acordo com o tipo de antecedente: em ambas as replicações, o TR foi menor quando os verbos (no singular ou no plural) eram antecedidos por sintagmas preposicionais (PP).

Implementamos MLMs com estrutura idêntica à do experimento original. Como não observamos, na Figura 1, a interação originalmente relatada, não esperávamos que ela iria se provar estatisticamente significativa. Por outro lado, encontramos indícios de que pode haver uma diferença no TR a depender do tipo de antecedente do verbo, isto é, se o verbo era precedido por um sintagma nominal (NP) ou por um sintagma preposicional (PP). Verificaremos, portanto, se essa diferença é capturada pelo modelo.

Figura 1: Médias do tempo de reação



Fonte: elaborado pelo autor

Ajustamos um modelo linear de efeitos mistos aos dados da R1 seguindo a estrutura do modelo original (COSTA, 2013, p. 69), tendo logRT como variável resposta, a interação entre o tipo de antecedente e o número do verbo como efeito fixo, e informantes e sentenças como efeitos aleatórios. O modelo, cujos valores estão reproduzidos na Tabela 1, não identificou estimativas estatisticamente significativas em nenhum dos tratamentos experimentais. A estimativa para a condição NP.PL ficou em 6,66 logRT (aproximadamente 780 ms), mais alta do que a estimativa original de 6,44 logRT (aproximadamente 625 ms).

Tabela 1: MLM ajustado para as observações da posição P7 da Replicação R1 de Costa (2013).

Efeitos Fixos	Estimativa (logRT)	Erro Padrão	Valor *t*	Valor *p*
Intercepto (NP.PL)	6,662	0,048	138,056	0,000
Antecedente PP	-0,069	0,059	-1,170	0,264
Número SG	-0,023	0,058	-0,392	0,702
PP.SG	0,000	0,083	0,005	0,996

Fonte: Elaboração própria.

Um MLM com a mesma estrutura foi ajustado aos dados da R2, com os valores reproduzidos na Tabela 2. Novamente, não foram observados, dentre as estimativas, valores que pareçam indicar diferenças entre os tratamentos. Nesse caso, a estimativa para a condição NP.PL ficou em 6,40 logRT (aproximadamente 600 ms), abaixo da original.

Tabela 2: MLM ajustado para as observações da posição P7 da Replicação R2 de Costa (2013).

Efeitos Fixos	Estimativa (logRT)	Erro Padrão	Valor *t*	Valor *p*
Intercepto (NP.PL)	6,407	0,032	201,855	0,000
Antecedente PP	-0,044	0,032	-1,358	0,199
Número SG	-0,016	0,032	-0,508	0,602
PP.SG	-0,005	0,045	0,101	0,921

Fonte: Elaboração própria.

1.5 Discussão

Os resultados para a posição crítica (P7) das Replicações R1 e R2 foram bastante similares, não replicando a interação NP.SG × PP.SG originalmente encontrada, como pode ser visto na Figura 1 e nas estimativas dos MLMs. Uma diferença importante entre as replicações é a queda nos erros padrão e as estimativas menores do TR em todas as condições na R2. Atribuímos essas diferenças entre os dados das replicações à separação da zona de espraiamento (P7) da zona do efeito de integralização (P8 à P9), realizada na R2 por meio da inserção de um PP no final dos estímulos-alvo.

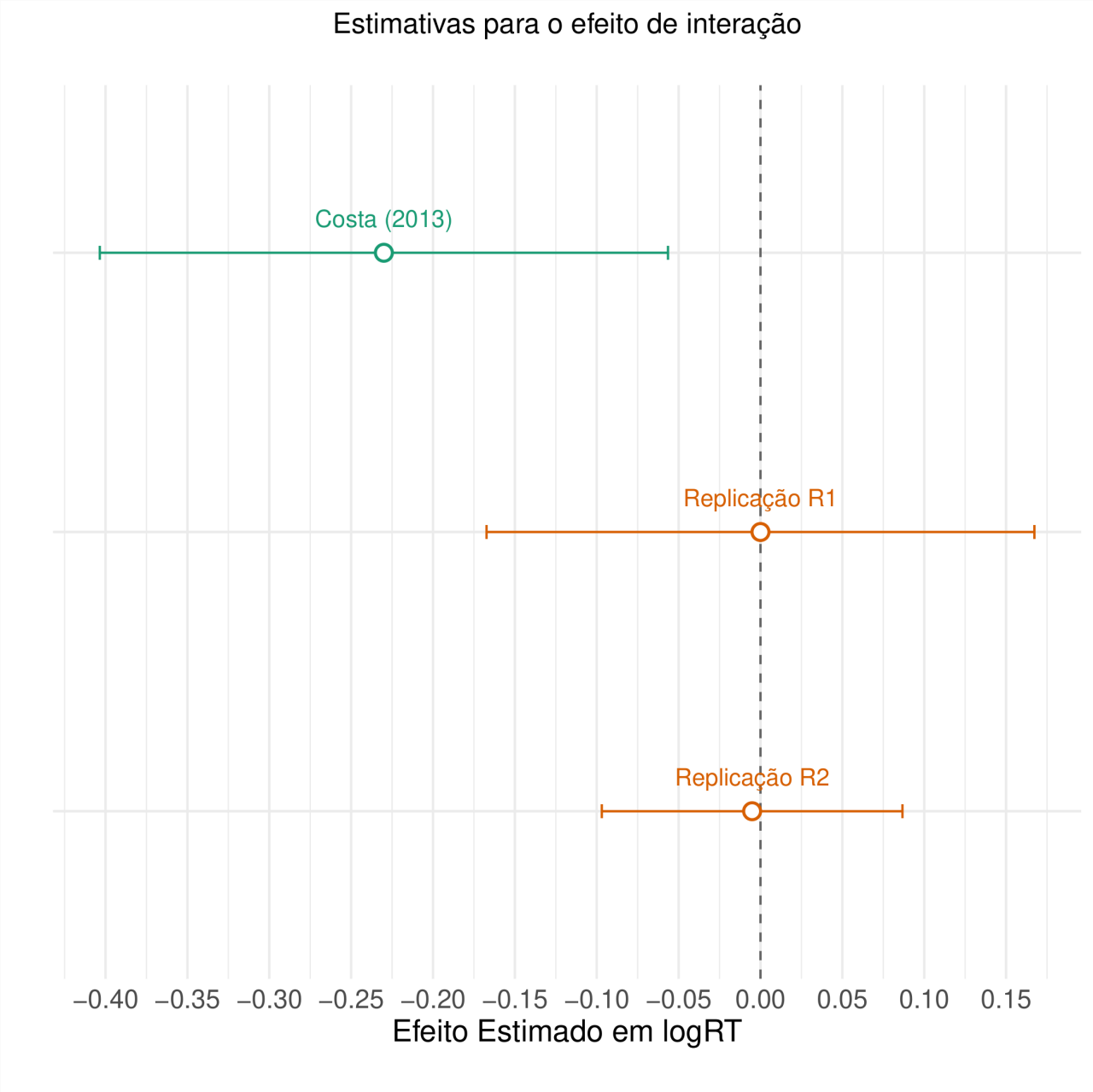
A obtenção de dados com menor variância, levando a melhores estimativas dos parâmetros populacionais, deveria ser um objetivo compartilhado por todos os psicolinguistas. Gelman e Carlin (2014) trazem um importante alerta sobre os erros de magnitude e de sinal positivo ou negativo nos efeitos relatados na literatura em psicologia (e psicolinguística, por extensão) quando amostras pequenas são empregadas[3]. Segundo esses e outros autores (e.g. IOANNIDIS, 2005), dados de experimentos com poucos informantes tendem a ser marcados por maior variabilidade. Como testes estatísticos capturam a razão *tamanho do efeito* (por exemplo, a diferença entre as médias dos grupos) dividido por *variabilidade*, quando essa última medida é alta, os efeitos tendem a ficar diluídos. Apenas tamanhos de efeito muitos grandes – e, potencialmente, inflados – resultarão em testes estatísticos que passam pelo crivo da significância estatística. A inflação nos tamanhos de efeito é um problema que afeta sistematicamente a literatura (BAKKER; VAN DIJK; WICHERTS, 2012) e que merece a atenção dos psicolinguistas.

A Figura 2 apresenta uma análise dos tamanhos do efeito na interação tipo de antecedente × número do verbo no experimento original e nas replicações (COSTA, 2022, manuscrito não publicado). O intervalo de confiança (IC) a 95%, representado pelas barras horizontais, foi calculado pela multiplicação do erro padrão por dois. O IC a 95% indica que, em futuras repetições da amostragem, o parâmetro populacional (no caso, o tamanho do efeito) estará dentro do intervalo estimado 95% das vezes. Há uma relação entre o IC e a significância estatística: se esse intervalo contiver zero, então o efeito estudado – nesse caso, a interação entre tipo de antecedente × número do verbo – estará ausente. Na figura, as estimativas do efeito da interação

[3] Na verdade, o problema vem do baixo poder estatístico, problema típico dos estudos na psicologia e na linguística experimental. Atentamos, aqui, para uma das suas principais causas: as pequenas amostras.

incluem zero nas Replicações R1 e R2, ao contrário da estimativa do tamanho do efeito do encontrado originalmente, indicando que o resultado original representa uma estimativa inflada de um efeito potencialmente inexistente.

Figura 2: Estimativa do efeito da interação



Fonte: elaborado pelo autor

2. Conclusão à luz da Crise da Replicabilidade

A condição da falseabilidade, pedra fundamental na construção do conhecimento científico (POPPER, 2002, p. 66), depende da realização de replicações. Achados que não tenham sido repetidos devem ser encarados com uma salutar dose de suspeita (RASTLE *et al.*, 2023). Ao apresentarmos uma replicação falha – até onde sabemos, a primeira na psicolinguística brasileira a se identificar como tal e discutir suas causas – acendemos um alerta importante em relação à superestimativa nos tamanhos de efeito. Como vimos, é possível que o efeito encontrado pelo Experimento 2 de Costa (2013) tenha sido contingencial, resultado de sua pequena amostra. O emprego de amostras de tamanho adequado, apesar de assunto com suas complexidades (VASISHTH, 2023), tem a vantagem da regressão à média, ou seja, de que as variações entre informantes, idiossincráticas por natureza, tendem a ter menor impacto na estimativa final do tamanho do efeito.

Por fim, é importante frisar alguns aspectos do presente estudo. Primeiramente, a não replicação do Experimento 2 de Costa (2013) não implica "refutarmos" a observação de que verbos meteorológicos podem exibir, em PB, flexão de número. Costa apresenta dados inequívocos de produção eliciada, além de ser trivial encontrar, usando a Internet como corpus, novos exemplos desse fenômeno. Isso posto, a não replicação do original lança dúvidas quanto a um efeito de expectativa de que verbos meteorológicos estejam flexionados em certos contextos que se manifeste na compreensão, ou, pelo menos, quanto à sensibilidade do protocolo da leitura autocadenciada na captura desse efeito. Em segundo lugar, salientamos que os critérios de exclusão de informantes e descarte de observações não foram responsáveis por apagar um possível efeito. Análises sem qualquer filtragem revelaram resultados equivalentes e podem ser conferidas em detalhes em <https://osf.io/pes2j/>. Por fim, para responder à pergunta no título deste artigo, concordamos com a crescente leitura de que não se deve falar em uma "crise" nas ciências (NATIONAL ACADEMIES OF SCIENCE, 2019), mas em uma natural revisão dos métodos até então considerados de excelência para a acomodação das novas possibilidades que surgem com as inovações tecnológicas e que vão desde o compartilhamento de dados em repositórios *online*, passando pela verificação de resultados por meio de replicações sistemáticas, até a adoção de novas ferramentas estatísticas.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 88887.479688/2020-00.

Referências

BAKKER, M.; VAN DIJK, A.; WICHERTS, J. M. The Rules of the Game Called Psychological Science. **Perspectives on Psychological Science**, v. 7, n. 6, p. 543–554, nov. 2012.

BIN, P. R.; MOTA, M. B. Pré-registro de estudos na linguística experimental. **Cadernos de Linguística**, v. 3, n. 1, p. e616, 23 mar. 2022.

COSTA, I. DE O. **Verbos meteorológicos no plural em orações relativas do Português Brasileiro: sintaxe e processamento**. Dissertação de mestrado — Rio de Janeiro: PUC-Rio, 2013.

COSTA, I. DE O. **Poder estatístico e tamanho da amostra nos estudos em psicolinguística expesrimental: uma abordagem compreensiva usando simulações**. Rio de Janeiro, 15 set. 2022.

CRUMP, M. J. C.; MCDONNELL, J. V.; GURECKIS, T. M. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. **PLoS ONE**, v. 8, n. 3, p. e57410, 13 mar. 2013.

GELMAN, A.; CARLIN, J. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. **Perspectives on Psychological Science**, v. 9, n. 6, p. 641–651, nov. 2014.

GODOY, M. C.; NUNES, M. A. Uma comparação entre ANOVA e modelos lineares mistos para análise de dados de tempo de resposta. **Revista da ABRALIN**, v. 19, n. 1, p. 1–23, 17 jul. 2020.

IOANNIDIS, J. P. A. Why Most Published Research Findings Are False. **PLoS Medicine**, v. 2, n. 8, p. e124, 30 ago. 2005.

IOANNIDIS, J. P. A. Why Most Discovered True Associations Are Inflated. **Epidemiology**, v. 19, n. 5, p. 640–648, set. 2008.

JUST, M. A.; CARPENTER, P. A. A theory of reading: from eye fixations to comprehension. **Psychological Review**, v. 87, n. 4, p. 28, 1980.

KOBROCK, K.; ROETTGER, T. B. Assessing the replication landscape in experimental linguistics. **Glossa Psycholinguistics**, v. 2, n. 1, 30 mar. 2023.

MARSDEN, E. *et al.* Replication in Second Language Research: Narrative and Systematic Reviews and Recommendations for the Field. **Language Learning**, v. 68, n. 2, p. 321–391, jun. 2018.

NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, AND MEDICINE. **Reproducibility and Replicability in Science**. Washington, DC: The National Academies Press, 2019.

OPEN SCIENCE COLLABORATION. Estimating the reproducibility of psychological science. **Science**, v. 349, n. 6251, p. aac4716, 28 ago. 2015.

PASHLER, H.; WAGENMAKERS, E. Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? **Perspectives on Psychological Science**, v. 7, n. 6, p. 528–530, 1 nov. 2012.

POPPER, K. **The Logic of Scientific Discovery**. London: Routledge, 2002.

RASTLE, K. et al. Beware influential findings that have not been replicated. **Journal of Memory and Language**, v. 129, p. 104390, fev. 2023.

RODD, J. M. Moving experimental psychology online: How to obtain high quality data when we can't see our participants. **Journal of Memory and Language**, v. 134, p. 104472, fev. 2024.

SAUTER, M.; DRASCHKOW, D.; MACK, W. Building, hosting and recruiting: a brief introduction to running behavioral experiments online. **Brain Sciences**, v. 10, n. 4, p. 251, 24 abr. 2020.

SÖNNING, L.; WERNER, V. The replication crisis, scientific revolutions, and linguistics. **Linguistics**, v. 59, n. 5, p. 1179–1206, 27 set. 2021.

SPROUSE, J. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. **Behavior Research Methods**, v. 43, p. 155–167, 2011.

VASISHTH, S. Some Right Ways to Analyze (Psycho)Linguistic Data. **Annual Review of Linguistics**, v. 9, n. 1, p. 273–291, 17 jan. 2023.

ZEHR, J.; SCHWARZ, F. **PennController for Internet Based Experiments (IBEX)**. 2018.