A comment on Varma et al. (2024) "A systematic review and meta-analysis of experimental methods for modulating intrusive memories following lab-analogue trauma exposure in non-clinical populations"

Ian Hussey[1], Martin Götz[2], & Saein Lee[3]

Varma et al. (2024) report findings from a systematic review and meta-analysis that assesses the causal claim that intrusive memories, induced from lab-analogue trauma exposure, can be increased or decreased through experimental techniques in non-clinical populations. This literature is particularly noted for the specific claim that playing the video game Tetris can reduce intrusive memories following lab-analogue trauma exposure. The article reports an overall very small effect size for experimental techniques (vs. controls) across various outcome measures, intervention types, and delay periods, with Hedges' g = 0.16, 95% CI [0.09, 0.23], p < .001. Additionally, the original article reports that Tetris was found to cause large decreases in intrusive memories compared to control both at post-treatment (g = 0.80, 95% CI [0.66, 0.93], p < .001) and at follow-up (g = 0.77, 95% CI [0.07, 1.48], p < .001). We assessed the computational reproducibility of a subset of the results we deemed more important using the authors' data and code. The distributed code did not implement many of the analyses reported, including the claim in the abstract, and included implementations of many other analyses not reported in the article. Where code was available, results were reproduced; however, when we had to reimplement analyses, most results could not be reproduced. We note that the claim in the abstract merges two different estimands: intervention studies aiming to decrease intrusions and manipulation studies aiming to increase them; these partially offset each other, rendering the estimate uninterpretable. We assessed robustness reproducibility by reanalysing main moderator analyses using only the subset of intervention studies that sought to decrease intrusions. Conclusions remained to be statistically significant, but were substantially larger in the robustness analyses than in the reproduced analyses. The original study's conclusions are partially computationally reproducible and broadly robust, indicating that interventions can reduce intrusions, albeit with considerable variations in effect sizes compared to the original study.

[1] University of Bern, https://orcid.org/0000-0001-8906-7559, ian.hussey@unibe.ch
[2] University of Zurich, https://orcid.org/0000-0003-1415-1240, martin.goetz@uzh.ch
[3] University of Zurich, https://orcid.org/0000-0003-2835-0937, lee@psychologie.uzh.ch

1. Introduction

In this report, prepared as part of a collaboration between the Institute for Replication and Nature Human Behaviour (Brodeur et al., 2024), we investigate whether Varma et al.'s (2024) analytical results are computationally reproducible and further examine their robustness by correcting exclusions to align the analysis with the estimand described in the abstract (i.e., interventions to decrease intrusion frequency, excluding manipulations that attempted to increase intrusions).

We assessed computational reproducibility in three ways: (1) the completeness of the code for calculating relevant results reported in the article, (2) the presence or absence of discrepancies between the numerical results reported in the article and those reproduced from the data and code, and (3) the accuracy of the numeric results recorded in the data compared to our re-extractions of these effect sizes from the original articles.

Given the large number of results reported by Varma et al. (2024), we selected a subset to attempt computational reproduction. This selection was based on 1) the claim mentioned in the abstract regarding the overall meta-analysis result, and 2) what we deemed to be a set of main results reported in the article itself. Specifically, we attempted to reproduce:

1. The meta-analysis results in Figure 4a for "all direction" predictions, which include the result reported in the abstract.
2. The first row of each subset in the (moderator) meta-analyses reported in Figure 5 at the "immediate post" (i.e., post intervention) and "delayed post" (i.e., follow up) timepoints.

We used the replication package provided by Varma et al. (2024): osf.io/phu7w.

To better understand the code and enhance its reproducibility, we converted the .R files in the replication package to .Rmd (i.e., RMarkdown) files. We made minor edits to all directory paths, changing them from absolute to relative paths, so that the replication package should now function on any machine with the required dependencies without further alteration. This follows common practices for distributing code on GitHub.

2. Computational Reproducibility

2.1 Summary of Reproducibility

Table 1 provides a summary of the availability and reproducibility of code, data, and results.

Table 1. Summary of availability and reproducibility

	Fully	Partial	No
Raw data provided	x		
Cleaning code provided		x	
Analysis data provided	x		
Analysis code provided		x	
Reproducible from raw data			x
Reproducible from analysis data		x	

Raw data was fully available, including the statistical results reported in original articles, which were converted to effect sizes for meta-analysis and included in the replication package.

Cleaning code was only partially available. There was no code provided to convert the statistical results of the individual studies into effect sizes and to invert some of their directions to standardize the interpretation (i.e., to create the *yi_pos* and *vi* variables used in the analyses). The cleaning code was available for converting these effect sizes and variances into various subsets for analyses in the (moderator) meta-analyses.

Analysis data was completely available, with all rows for the *yi_pos* and *vi* variables seemingly provided, allowing us to reimplement the overall meta-analysis and reproduce the primary result reported in the abstract.

Analysis code, however, was only partially available. Only some of the meta-analysis models appeared to be implemented in the code. Notably absent were the overall meta-analysis result reported in the abstract and in Figure 4a, as well as the main moderator and specific intervention meta-analyses reported in Figure 5. Additionally, the available code included a large number of additional moderator meta-analyses on data subsets (i.e., split by outcome type) that were not reported in the article.

Results were not reproducible from raw data, as no cleaning code for effect size and variance calculation was provided, and no attempt was made to reimplement this. Consequently, the results were not reproducible from the raw data.

Results were partially reproducible from analysis data. Once reimplemented, we could precisely reproduce the effect size estimate reported in the abstract, and most of the main results we attempted to reproduce from Figure 4a. Most of the results for the moderator and specific intervention analyses from Figure 5 could not be reproduced.

2.2 Computational reproduction of Figure 4a's main results

Table 2 presents the reproduction of the key findings from Varma et al.'s (2024) Figure 4a. It is important to note that the result reported in the abstract refers to "overall intrusion frequency," which encompasses "diary-based intrusion frequency," "lab-based intrusion frequency," and "questionnaire-based intrusion frequency" outcome variables. This comprehensive meta-analysis was not included in the publicly available code. However, upon reimplementing this code, we successfully reproduced the reported results.

2.3 Computational reproduction of Figure 5's main results

Varma et al. (2024) reported the results of moderator meta-analyses and meta-analyses of individual intervention types in their Figure 5. We attempted to reproduce the results of what we deemed to be the most important results, as previously described. Results of the reproduction can be found in Table 3. As discussed previously, these moderator meta-analyses were not implemented in the publicly available code. When we reimplemented them, the results generally did not reproduce those reported in Varma et al.'s (2024) Figure 5.

Table 2. Reproduction of results from Varma et al.'s (2024) Figure 4a.

Source	Outcome	*N*	*n*	*k*	Hedges' *g*	Lower	Upper	*p*
Article	Overall intrusion frequency*	10,765	139	370	0.16	0.09	0.23	< .001
	Diary-based intrusion frequency	8,985	121	248	0.14	0.07	0.21	< .001
	Lab-based intrusion frequency	2,859	39	79	0.30	0.17	0.43	< .001
	Questionnaire-based intrusion frequency	1,715	17	43	0.13	-0.04	0.31	.137
	Intrusion-related distress	5,690	66	136	0.04	-0.03	0.11	.290
	Intrusion symptoms	4,730	55	100	0.07	-0.01	0.16	.089
Reproduced	Overall intrusion frequency*	10,765	139	370	0.16	0.09	0.23	< .001
	Diary-based intrusion frequency	8,985	121	248	0.14	0.07	0.21	< .001
	Lab-based intrusion frequency	2,859	39	79	0.30	0.17	0.43	< .001
	Questionnaire-based intrusion frequency	1,715	17	43	0.13	-0.04	0.31	.137
	Intrusion-related distress	5,690	66	136	0.04	-0.03	0.11	.290
	Intrusion symptoms	4,730	55	100	0.07	-0.01	0.16	.089
Comparison	Overall intrusion frequency*	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
	Diary-based intrusion frequency	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
	Lab-based intrusion frequency	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
	Questionnaire-based intrusion frequency	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
	Intrusion-related distress	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
	Intrusion symptoms	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Notes: N = number of participants, *n* = number of experiments, *k* = number of effect sizes, Lower and Upper = 95% Confidence Intervals. *Not implemented in original code; reimplemented during computational reproduction.

Table 3. Reproduction of a subset of results from Varma et al.'s (2024) Figure 5.

Source	Analysis	Outcome	Timepoint	*N*	*n*	*k*	Hedges' *g*	Lower	Upper	*p*
Article	Procedures	Behavioural	Immediate post	6,042	77	226	0.23	0.15	0.31	< .001
	Mechanisms	Imagery	Immediate post	2,933	37	96	0.31	0.18	0.43	< .001
	Individual Techniques	Trauma Reminder + Tetris vs Trauma Reminder	Immediate post	607	11	21	0.80	0.66	0.93	< .001
	Procedures	Behavioural	Delayed post	1,018	9	31	0.26	-0.02	0.54	.069
	Mechanisms	Imagery	Delayed post	837	8	27	0.27	-0.05	0.59	.104
	Individual Techniques	Trauma Reminder + Tetris	Delayed post	273	3	6	0.77	0.07	1.48	.032
Reproduced *	Procedures	Behavioural	Immediate post	6,655	84	368	0.17	0.09	0.25	< .001
	Mechanisms	Imagery	Immediate post	3,069	39	141	0.26	0.13	0.39	< .001
	Individual Techniques	Trauma Reminder + Tetris vs Trauma Reminder	Immediate post	607	11	27	0.72	0.56	0.87	< .001
	Procedures	Behavioural	Delayed post	1,204	12	51	0.10	-0.12	0.33	.400
	Mechanisms	Imagery	Delayed post	954	10	40	0.10	-0.19	0.39	.500
	Individual Techniques	Trauma Reminder + Tetris	Delayed post	273	3	9	0.62	0.02	1.21	.040
Comparison	Procedures	Behavioural	Immediate post	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
	Mechanisms	Imagery	Immediate post	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
	Individual Techniques	Trauma Reminder + Tetris vs Trauma Reminder	Immediate post	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE
	Procedures	Behavioural	Delayed post	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Mechanisms	Imagery	Delayed post	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

	Individual Techniques	Trauma Reminder + Tetris	Delayed post	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE

Notes: N = number of participants, *n* = number of experiments, *k* = number of effect sizes, Lower and Upper = 95% Confidence Intervals. *None of these analyses were implemented in the original code; they were reimplemented during computational reproduction.

2.4 Computational reproduction of data extractions for extreme effect sizes

While Varma et al. (2024) did report that their results were generally robust to excluding outlier effect sizes, it should be noted that these outliers were assessed for in a data-driven manner, as is common. If systemic errors in effect size extraction or conversion are prevalent in a dataset, such issues may not be detected by such data-driven exclusion strategies.

Errors in data extraction are highly prevalent in meta-analyses (e.g., Gøtzsche et al., 2007; Hussey, 2025; Maassen et al., 2020). In particular, standard errors are often confused for standard deviations when calculating Standardized Mean Difference effect sizes (e.g., Cohen's *d*, Hedges' *g*), which can significantly inflate them and result in implausibly large effect size (Harrer et al., 2025). As such, Cochrane's recently adopted INSPECT-SR Trustworthiness Assessment tool highlights the importance of assessing, among other things, the mathematical feasibility and plausibility of *SD*s and effect sizes (Wilkinson et al., 2025). Although a comprehensive validation of effect size extractions and conversions was beyond the scope of our reanalysis, we did a small number of spot checks to attempt to computationally reproduce some of the most extreme (absolute) Hedges' *g* values in the dataset.

The Hedges' $g = 2.67$ reported for Cheung and Bryant (2017) was the largest in Varma et al.'s (2024) dataset. We found strong indications that Cheung and Bryant (2017) mislabelled Standard Errors as Standard Deviations, and these values were taken at face value by Varma et al. (2024). Details of these validations of Cheung and Bryant (2017) are available in the supplementary materials. The recalculated effect is much lower: $g = 0.69$.

Similarly, two effect sizes were included in Varma et al.'s (2024) data from James et al. (2015): *g*s = 1.16 and 1.36. Varma et al.'s (2024) dataset notes that these effect sizes were extracted from James et al.'s (2015) published plots. Extracting data from plots can be useful, but it is necessarily more imprecise than calculating results from the original data. Varma et al. (2024) may have overlooked that James et al. (2015) provided open data. We recalculated effect sizes from the data itself and found a similar effect size (in one case) and a much smaller effect (in the other case): *g*s = 1.12 and 0.97.

These reproductions were intended to illustrate the general point that effect size extractions are frequently erroneous, but did not comprehensively evaluate the prevalence of extraction errors in Varma et al.'s (2024) data or any impact on their results, which remains unknown.

3. Robustness Reproduction and Replication with New Data

After reviewing the code and data, we developed sensitivity analyses to better understand how the code executed the analyses. This process, along with the reproductions we generated, clarified the (mis)alignments between the reported results and the code' implementation or lack thereof. Consequently, these sensitivity analyses were not preregistered.

3.1 Robustness of the overall meta-analysis to excluding "increase" studies

While the meta-analysis estimate reported in the abstract could be reproduced when we reimplemented code for this analysis, we observe, however, that estimate is uninterpretable because it is fundamentally misaligned with its interpretation. Varma et al. (2024) state, "results showed that techniques (behavioural, pharmacological, neuromodulation) significantly reduced intrusion frequency (g = 0.16, 95% confidence interval [0.09, 0.23])" (p. 1,968). However, this overall analysis combines results from both (a) intrusion intervention studies (referred to by Varma et al. (2024) as "decrease" studies), and (b) intrusion induction studies (referred to by Varma et al. (2024) as "increase" studies), without adjusting for the direction of interpretation for either effect. Consequently, the overall estimate is a blend of (mostly positive) effect sizes from intervention studies and (mostly negative) effect sizes from induction studies.

To clarify this point, we conducted an overall meta-analysis divided by "increase" vs. "decrease" studies. The overall g = 0.16, 95% CI [0.09, 0.23] result reported in the abstract is, in effect, a combination of the overall effect for increase/induction studies (g = -0.13, 95% CI [-0.25, -0.02], p = .025) and the overall effect of decrease/intervention studies (g = 0.31, 95% CI [0.23, 0.39], p < .001), which have opposite causal goals and therefore conficting estimands. In our view, because it represents a mixture of two fundamentally different estimands, the overall meta-effect size reported in the abstract lacks meaningful or useful interpretation.

Nonetheless, Varma et al.'s (2024) verbal assertion that "results showed that techniques (behavioural, pharmacological, neuromodulation) significantly reduced intrusion frequency" (p. 1,968) is supported by the results of the robustness test in the "decrease" studies, as both are statistically significant, although the effect in the subset is more than twice as large as the reported one.

3.2 Robustness of the moderator meta-analyses to excluding "increase" studies

The same concern applies to all the analyses reported in Figure 5, which integrate results from both "increase" and "decrease" studies. As a robustness test, we conducted the subset of moderator meta-analyses mentioned earlier in the reproduction section,

focusing solely on the “decrease” studies to achieve a clearer estimand. The Tetris-related analyses were not included here, as the originally described meta-analyses were already applied to exclusively “decrease” studies.

Table 4. Key moderator analyses using the subset of “increase” studies

Analysis	Outcome	Timepoint	*N*	*n*	*k*	Hedges' *g*	Lower	Upper	*p*	Δ*g*
Procedures	Behavioural	Immediate post	5,069	63	192	0.31	0.23	0.39	< .001	0.14
Mechanisms	Imagery	Immediate post	2,281	30	82	0.40	0.25	0.55	< .001	0.14
Procedures	Behavioural	Delayed post	1,018	9	34	0.26	-0.04	0.56	.091	0.16
Mechanisms	Imagery	Delayed post	837	8	29	0.27	-0.08	0.62	.128	0.17

Notes: N = number of participants, *n* = number of experiments, *k* = number of effect sizes, Lower and Upper = 95% Confidence Intervals, Δ*g* = change in Hedges’ *g* compared to the reproduced results for the moderator analyses using both “increase” and “decrease” studies.

The robustness test results reveal the same conclusions when focusing solely on statistical significance (see Tables 3 and 4). Regarding effect sizes, these were approximately one and a half times larger in the robustness tests compared to the reproduced analyses at the "immediate post” timepoint and roughly twice as large at the “delayed post” timepoint.

4. Conclusion

Varma et al.’s (2024) findings are only partially computationally reproducible, based on our efforts to reproduce a subset of the most significant results. Several key findings, including the one highlighted in the abstract, are not implemented in the provided code. When we re-executed the analyses as described, except for the overall meta-analysis’s result reported in the abstract, none of the corresponding results could be reproduced. However, the main result in the abstract, despite being reproducible, is arguably difficult to interpret as it combines two very different estimands: studies that induce intrusions and those that that reduce them. Although this was acknowledged as potentially presenting “a misleading picture of results” (Varma et al., 2024, p. 1970), it remains the headline claim in the abstract. We contend that results that are known to be misleading or difficult to interpret should not be presented without qualification as the main results supporting the core argument.

Varma et al.’s (2024) general assertions remain to be supported by the reproductions and robustness tests: (1) interventions aimed at reducing intrusions in non-clinical populations

have a non-zero efficacy, and (2) inductions designed to provoke intrusions in non-clinical populations also exhibit non-zero efficacy. However, substantial discrepancies or uncertainties persist regarding the magnitude of these effects.

Conflict of interest

All authors declare they have no conflict of interest.

Code and data availability

All data and code can be found here: github.com/ianhussey/verification-of-varma-et-al-2024

References

Brodeur, A., Dreber, A., Hoces De La Guardia, F., & Miguel, E. (2024). Reproduction and replication at scale. *Nature Human Behaviour*, *8*(1), 2–3. https://doi.org/10.1038/s41562-023-01807-2

Bub, K., & Lommen, M. J. J. (2017). The role of guilt in posttraumatic stress disorder. *European Journal of Psychotraumatology*, *8*(1), 1407202. https://doi.org/10.1080/20008198.2017.1407202

Cheung, J., & Bryant, R. A. (2017). The impact of appraisals on intrusive memories. *Journal of Behavior Therapy and Experimental Psychiatry*, *54*, 108–111. https://doi.org/10.1016/j.jbtep.2016.07.005

Frasquilho, F. M. (2004). *The role of peri-traumatic visuo-spatial and verbal interference on the development of intrusions.* [Doctoral thesis]. University College London.

Gøtzsche, P. C., Hróbjartsson, A., Marić, K., & Tendal, B. (2007). Data extraction errors in meta-analyses that use standardized mean differences. *JAMA*, *298*(4). https://doi.org/10.1001/jama.298.4.430

Harrer, M., Miguel, C., Hussey, I., Cristea, I. A., Van Ballegooijen, W., Basic, D., Wang, Y., Pfund, R. A., Quero, S., Van Spreckelsen, P., Schnurr, P. P., Van Straten, A., Furukawa, T. A., Papola, D., & Cuijpers, P. (2025). *Implausible effects of psychological interventions meta-epidemiological study and development of a simple flagging tool*. Psychiatry and Clinical Psychology. https://doi.org/10.1101/2025.11.12.25340062

Hussey, I. (2025). Verification report: A critical reanalysis of Vahey et al. (2015) “A meta-analysis of criterion effects for the Implicit Relational Assessment Procedure (IRAP) in the clinical domain.” *Journal of Behavior Therapy and Experimental Psychiatry*, *87*, 102015. https://doi.org/10.1016/j.jbtep.2024.102015

James, E. L., Bonsall, M. B., Hoppitt, L., Tunbridge, E. M., Geddes, J. R., Milton, A. L., & Holmes, E. A. (2015). Computer game play reduces intrusive memories of experimental trauma via reconsolidation-update mechanisms. *Psychological Science*, *26*(8), 1201–1215. https://doi.org/10.1177/0956797615583071

Lau-Zhu, A., Henson, R. N., & Holmes, E. A. (2019). Intrusive memories and voluntary memory of a trauma film: Differential effects of a cognitive interference task after encoding. *Journal of Experimental Psychology: General*, *148*(12), 2154–2180. https://doi.org/10.1037/xge0000598

Maassen, E., van Assen, M. A. L. M., Nuijten, M. B., Olsson-Collentine, A., & Wicherts, J. M. (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. *PLOS ONE*, *15*(5), e0233107. https://doi.org/10.1371/journal.pone.0233107

Page, S., & Coxon, M. (2017). Preventing post-traumatic intrusions using virtual reality. In B. K. Wiederhold, G. Riva, C. Fullwood, A. Attrill-Smith, & G. Kirwan (Eds.), *Annual review of cybertherapy and telemedicine 2017* (Vol. 15, pp. 129–134). Interactive Media Institute.

Varma, M. M., Zeng, S., Singh, L., Holmes, E. A., Huang, J., Chiu, M. H., & Hu, X. (2024). A systematic review and meta-analysis of experimental methods for modulating intrusive memories following lab-analogue trauma exposure in non-

clinical populations. *Nature Human Behaviour*, *8*(10), 1968–1987. https://doi.org/10.1038/s41562-024-01956-y

Wilkinson, J., Heal, C., Flemyng, E., Antoniou, G. A., Aburrow, T., Alfirevic, Z., Avenell, A., Barbour, V., Berghella, V., Bishop, D. V. M., Bordewijk, E. M., Brown, N. J. L., Christopher, J., Clarke, M., Dahly, D., Dennis, J., Dicker, P., Dumville, J., Frankish, H., … Kirkham, J. J. (2025). *INSPECT-SR: A tool for assessing trustworthiness of randomised controlled trials*. Epidemiology. https://doi.org/10.1101/2025.09.03.25334905