Supplementary materials: The Uneven Benefits of Hiding Social Cues: Demetrication and Well-Being on Instagram

Contents

Appendix A Supplementary Methods

A.1 Randomisation and Baseline Balance

Participants were randomised to the demetricated condition (IG00; $n = 64$) or the control condition (IG01; $n = 76$) during the Smolgram onboarding flow. Because randomisation was conducted algorithmically at the point of account creation, no pre-stratification or blocking was used. Table A1 reports Welch t-tests (continuous variables) and a Pearson χ^2 test (sex) comparing the two conditions on demographic characteristics at enrolment. All demographic covariates were well balanced: age ($t = -0.11$, $p = .909$), daily Instagram use ($t = +0.23$, $p = .821$), and sex ($\chi^2 = 0.09$, $p = .763$). Follower count showed a numerical but non-significant difference ($t = -1.69$, $p = .095$) that was reflected in a standardised mean difference of SMD = 0.29, the only demographic covariate exceeding the conventional 0.10 threshold [1] (Table A2; Figure A1). Follower count was therefore included as a covariate in all ANCOVA models.

Table A1 Baseline Equivalence by Condition (Demographic Covariates)

Variable	Demetricated M (SD)	Control M (SD)	Statistic	p
Age	22.24 (2.85)	22.30 (2.39)	$t = -0.11$	.909
Daily IG use (1-6)	2.65 (1.46)	2.59 (1.28)	$t = +0.23$	.821
Followers	452.54 (297.49)	615.06 (735.90)	$t = -1.69$	.095†
Sex (% female)	70.4%	66.2%	$\chi^2 = 0.09$	.763

Note. $N = 140$ enrolled. Welch t-tests for continuous variables; Pearson χ^2 for sex. † $p < .10$.

Table A2 Standardised Mean Differences at Baseline (Demographic Covariates)

Covariate	SMD	$\lvert\text{SMD}\rvert < 0.10$
Age	0.021	✓
Sex (female)	0.090	✓
Daily IG use (1–6)	0.041	✓
Follower count	0.290	×

Note. SMD threshold follows Austin [1]. Only follower count shows imbalance; this covariate is included in all ANCOVA models.

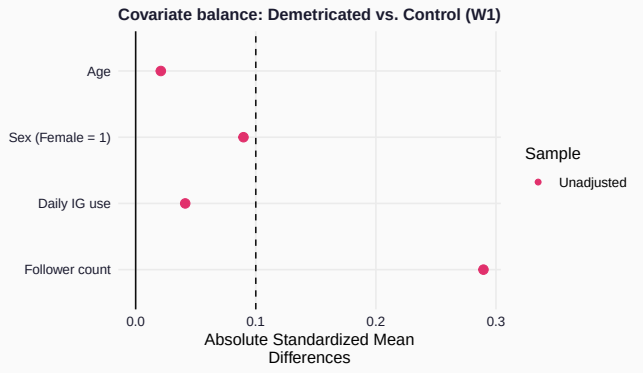

Fig. A1 Love plot of standardised mean differences (SMD) for demographic covariates at baseline. The dashed reference line marks the 0.10 balance threshold. Only follower count exceeded the threshold (SMD = 0.29); this covariate was included in all ANCOVA models.

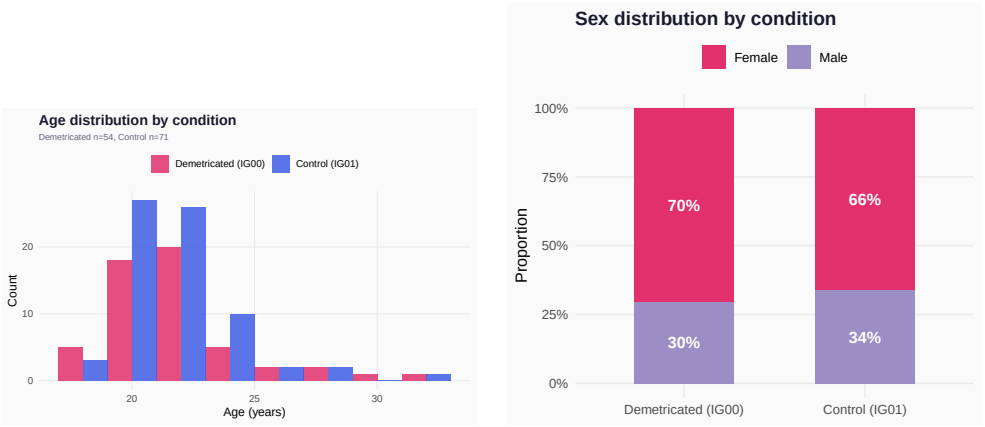

Fig. A2 Age distribution (left) and sex composition (right) by condition at enrolment.

Appendix B Supplementary Results

B.1 Attrition and App Engagement

Of the 140 enrolled participants, 125 (89.3%) completed the Wave 3 survey. Attrition was 15.6% in the demetricated condition (IG00; 10 of 64) and 6.6% in the control condition (IG01; 5 of 76). Although attrition was numerically higher in IG00, the difference was not statistically significant ($\chi^2 = 2.10$, $p = .147$), indicating that differential attrition is unlikely to be a serious threat to internal validity.

In terms of active app use over the 21-day study window, participants in the demetricated condition logged a median of 16 active days ($M = 15.3$, $SD = 6.4$); control participants logged a median of 14 active days ($M = 13.7$, $SD = 7.6$). This

difference was not significant ($t(138) = 1.32$, $p = .190$), suggesting that hiding metrics did not reduce overall app engagement. A Kaplan–Meier log-rank test of cumulative retention over the study period was likewise non-significant ($p = .740$), indicating comparable within-study dropout across conditions (Figure B3).

The analytic sample used in all ANCOVA models ($n = 123$) excludes two additional participants whose accounts had no traceable activity records in the Smolgram database; their Wave 3 survey responses were retained as attrite data for the bounds analyses below.

Table B3 App Engagement and Survey Attrition by Condition

Condition	N enrolled	N W3	Attrition %	N any use	$N \geq 14$ days	Median days
Demetricated (IG00)	64	54	15.6%	64	39	16
Control (IG01)	76	71	6.6%	74	44	14

Note. Active days = unique calendar days with any Smolgram event in the 21-day study window. Differential attrition (W3 completion): $\chi^2 = 2.10$, $p = .147$. Log-rank test: $p = .740$.

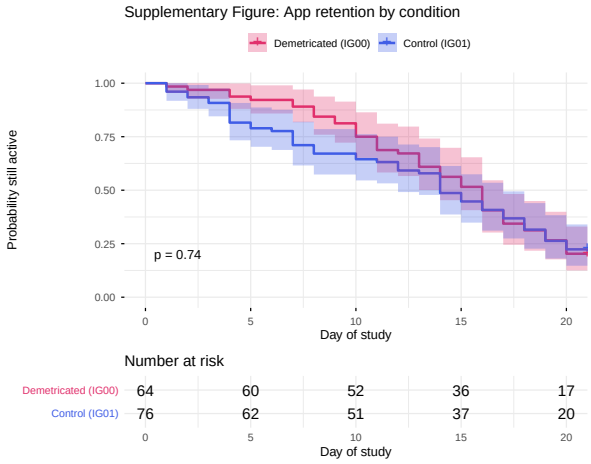

Fig. B3 Kaplan–Meier survival curves showing the probability of remaining active in Smolgram across the 21-day study window. Shaded bands indicate 95% confidence intervals. Log-rank test: $p = .740$.

B.2 Attrition Bounds

To assess robustness to non-random attrition, we computed Manski extreme-value bounds [2] for each primary outcome using the `attrition` package [3]. These bounds assume the worst-case scenario in which missing participants would have scored at the theoretical minimum or maximum of each scale. All bound intervals include zero

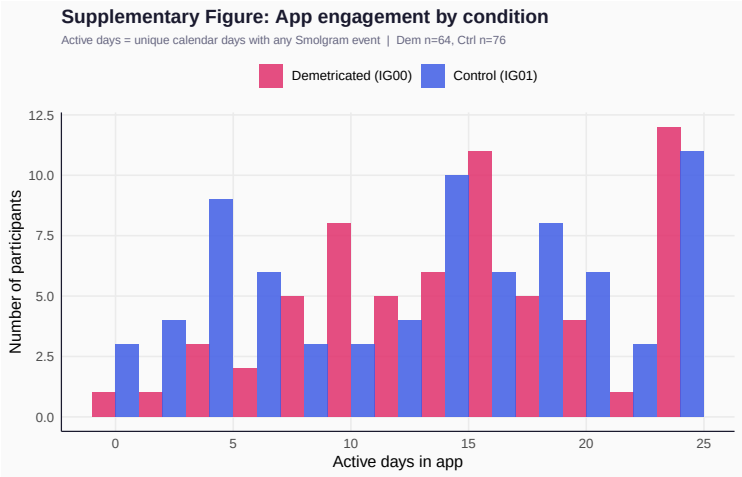

Fig. B4 Distribution of active days in Smolgram by condition. Active days = unique calendar days with any recorded Smolgram event.

(Table B4), consistent with the conclusion that any treatment effect is small and that the current findings are not an artefact of selective dropout. Because differential attrition was non-significant, Lee trimming bounds [4], which require unequal attrition rates, are not applicable here.

Table B4 Manski Extreme-Value Bounds for Treatment Effect Under Worst-Case Attrition

Outcome	Lower bound	Upper bound	95% CI lower	95% CI upper
SPANE-P (positive affect)	−0.197	+0.506	−0.421	+0.717
SPANE-B (affect balance)	−0.222	+0.999	−0.605	+1.346
Flourishing	−0.148	+0.435	−0.345	+0.619
Perceived social support	−0.110	+0.556	−0.342	+0.735

Note. Attrited participants assigned their theoretical minimum or maximum possible score [**?**]. All bounds intervals include zero, consistent with non-significant differential attrition. Lee trimming bounds not applicable (non-differential attrition).

B.3 Pre-registered Treatment Effects: All Outcomes (H1–H5)

We pre-registered five hypotheses concerning the effect of demetrication on perceived social support (H1), self-disclosure (H2), self-discrepancy (H3), emotional well-being (H4; SPANE-P and SPANE-B), and subjective well-being (H5; Flourishing). All hypotheses were tested using ANCOVA with the Wave 3 outcome regressed on treatment, the Wave 1 baseline of the same outcome, and demographic covariates (age, sex, IG use, follower count; $n = 123$; Table B5). Treatment was coded 1 for the control condition (metrics visible, IG01) and 0 for the demetricated condition (IG00), so a negative coefficient indicates that the demetricated group scored higher.

H1: Perceived social support. Demetrication was associated with significantly higher perceived social support at Wave 3 ($b = -0.140$, $SE = 0.069$, $t(116) = -2.01$, $p = .046$, 95% CI $[-0.277, -0.002]$). The effect was small in magnitude but directionally consistent with the hypothesis that removing visibility metrics reduces social comparison pressure and thereby preserves or enhances users' sense of social connection.

H2: Self-disclosure. No significant treatment effect on self-disclosure was observed ($b = +0.051$, $SE = 0.110$, $p = .646$). The result is directionally opposite to H2 (demetrication was expected to *increase* disclosure), though the point estimate is negligible and the confidence interval spans both directions.

H3: Self-discrepancy. The proxy self-discrepancy score (Euclidean distance between Wave 1 general BFI-15 subscales and Wave 3 IG-context BFI-15 subscales) showed no significant treatment effect ($b = -0.121$, $SE = 0.134$, $p = .367$). The direction is consistent with H3 (demetrication reducing the gap between real and online self), but the effect is imprecisely estimated.

H4a–H4b: Emotional well-being. Neither SPANE-P (positive affect; $b = -0.079$, $p = .391$) nor SPANE-B (affect balance; $b = -0.137$, $p = .337$) showed a significant main effect of treatment. As reported in the main text, however, these outcomes were significantly moderated by IG-context neuroticism: participants higher in IG-context neuroticism benefited more from demetrication on both SPANE-P ($b_{\text{int}} = +0.388$, $p = .007$) and SPANE-B ($b_{\text{int}} = +0.497$, $p = .025$).

H5: Flourishing. No significant main effect was found for Flourishing ($b = -0.103$, $SE = 0.070$, $p = .146$), though the direction was consistent with H5. As with the SPANE outcomes, the Flourishing effect was significantly moderated by IG-context neuroticism ($b_{\text{int}} = +0.236$, $p = .035$).

Table B5 Pre-registered Treatment Effects on All Outcomes (ANCOVA, $n = 123$)

Outcome	n	b	SE	t	df	p	95% CI
Social connection							
Support (H1)	123	−0.140	0.069	−2.01	116	.046*	[−0.277, −0.002]
Self-presentation							
Self-disclosure (H2)	123	+0.051	0.110	+0.46	116	.646	[−0.168, +0.269]
Self-discrepancy[a] (H3)	123	−0.121	0.134	−0.91	116	.367	[−0.387, +0.144]
Well-being							
SPANE-P (H4a)	123	−0.079	0.092	−0.86	116	.391	[−0.261, +0.103]
SPANE-B (H4b)	123	−0.137	0.142	−0.96	116	.337	[−0.419, +0.145]
Flourishing (H5)	123	−0.103	0.070	−1.46	116	.146	[−0.242, +0.036]

[a] Self-discrepancy = Euclidean distance between Wave 1 general BFI-15 subscales and Wave 3 IG-context BFI-15 subscales. General Big Five traits not re-administered at Wave 3.

Note. Treatment: 1 = control (metrics visible, IG01), 0 = demetricated (IG00); negative b = demetricated scored higher. ANCOVA controls for Wave 1 baseline, age, sex, IG use, and follower count. † $p < .10$; $* p < .05$; $** p < .01$; $*** p < .001$.

B.4 Sensitivity: Activity-Threshold Analyses

A concern for ecological validity in field experiments using custom apps is whether results are driven by superficial users who engaged minimally with the platform. To address this, we re-estimated each model in subsamples defined by minimum active-days thresholds $k \in \{0, 1, 5, 10, 14\}$, where $k = 0$ corresponds to the full analytic sample ($n = 123$) and $k = 14$ retains only participants who were active for at least two weeks of the 21-day window ($n = 72$).

Main effects (Table B6). The main effect of demetrication on perceived social support (H1) was significant at the full-sample threshold ($b = -0.140$, $p = .046$) and at $k = 1$ ($p = .046$), but attenuated and became non-significant at higher thresholds ($k \geq 5$; all $p > .10$). This pattern suggests the support effect may be weaker among heavier users, though sample size reduction at higher k also increases uncertainty. Main effects on SPANE-P, SPANE-B, and Flourishing were consistently null across all thresholds.

Treatment × IG-context neuroticism interactions (Table B7). By contrast, the interaction effect on SPANE-P was significant across all thresholds ($k = 0$: $p = .007$; $k = 5$: $p = .045$; $k = 10$: $p = .037$; $k = 14$: $p = .018$), demonstrating that the neuroticism moderation is robust to users with minimal engagement being included or excluded. The SPANE-B interaction was significant at $k \in \{0, 1, 5\}$ and marginal at $k \in \{10, 14\}$ ($p = .059, .055$). The Flourishing interaction was significant at $k \in \{0, 1\}$ and non-significant at higher thresholds, suggesting it is somewhat sensitive to sample composition. The support interaction was consistently null across all thresholds ($p > .25$).

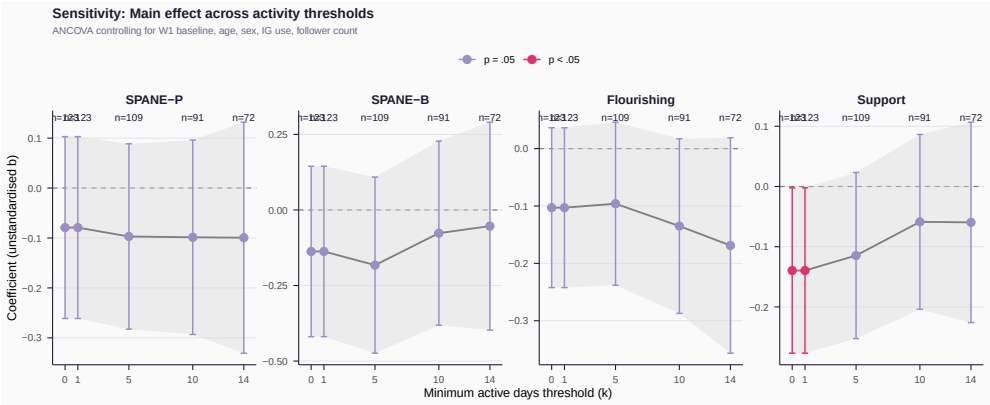

Fig. B5 Sensitivity of main ANCOVA treatment effects across minimum active-days thresholds (k). Each point is the treatment coefficient with 95% CI from the subsample with $\geq k$ active days. Annotations show per-panel n. Pink = $p < .05$.

Table B6 Sensitivity: Main Effect of Demetrication Across Minimum Active-Days Thresholds

Outcome	Min. days (k)	n	b	95% CI	p
SPANE-P	≥ 0	123	−0.079	[−0.261, +0.103]	.391
	≥ 1	123	−0.079	[−0.261, +0.103]	.391
	≥ 5	109	−0.097	[−0.283, +0.089]	.302
	≥ 10	91	−0.099	[−0.293, +0.096]	.317
	≥ 14	72	−0.099	[−0.331, +0.132]	.394
SPANE-B	≥ 0	123	−0.137	[−0.419, +0.145]	.337
	≥ 1	123	−0.137	[−0.419, +0.145]	.337
	≥ 5	109	−0.182	[−0.474, +0.109]	.217
	≥ 10	91	−0.077	[−0.381, +0.228]	.618
	≥ 14	72	−0.053	[−0.397, +0.291]	.758
Flourishing	≥ 0	123	−0.103	[−0.242, +0.036]	.146
	≥ 1	123	−0.103	[−0.242, +0.036]	.146
	≥ 5	109	−0.096	[−0.238, +0.046]	.183
	≥ 10	91	−0.135	[−0.287, +0.017]	.081†
	≥ 14	72	−0.169	[−0.357, +0.019]	.077†
Support	≥ 0	123	−0.140	[−0.277, −0.002]	.046*
	≥ 1	123	−0.140	[−0.277, −0.002]	.046*
	≥ 5	109	−0.115	[−0.253, +0.023]	.102
	≥ 10	91	−0.059	[−0.204, +0.086]	.423
	≥ 14	72	−0.060	[−0.226, +0.107]	.477

Note. Subsample retains participants with $\geq$ k active days. ANCOVA controls Wave 1 baseline, age, sex, IG use, follower count. † $p < .10$; $* p < .05$; $** p < .01$; $*** p < .001$.

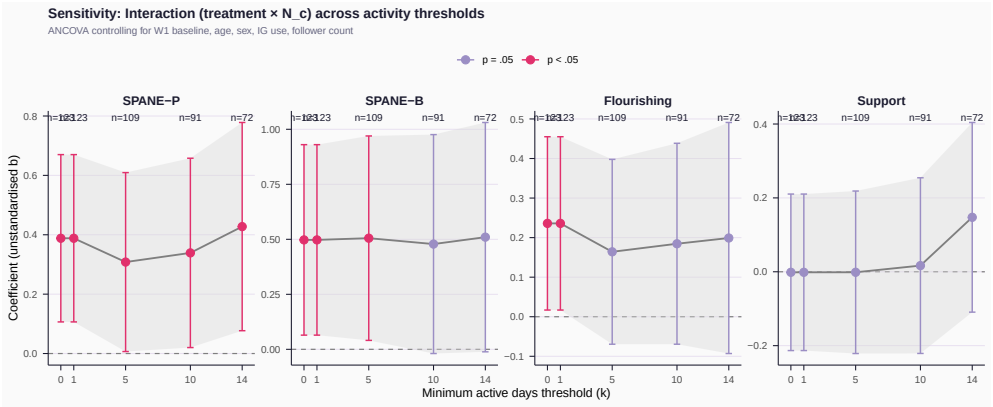

Fig. B6 Sensitivity of the treatment × IG-context neuroticism interaction across minimum active-days thresholds (k). Pink = $p < .05$, muted purple = $p \geq .05$.

Table B7 Sensitivity: Treatment × IG-Context Neuroticism Across Minimum Active-Days Thresholds

Outcome	Min. days (k)	n	b	95% CI	p
SPANE-P	≥ 0	123	+0.388	[+0.106, +0.670]	.007**
	≥ 1	123	+0.388	[+0.106, +0.670]	.007**
	≥ 5	109	+0.308	[+0.007, +0.609]	.045*
	≥ 10	91	+0.339	[+0.020, +0.658]	.037*
	≥ 14	72	+0.428	[+0.077, +0.778]	.018*
SPANE-B	≥ 0	123	+0.497	[+0.064, +0.930]	.025*
	≥ 1	123	+0.497	[+0.064, +0.930]	.025*
	≥ 5	109	+0.505	[+0.041, +0.970]	.033*
	≥ 10	91	+0.478	[−0.019, +0.976]	.059†
	≥ 14	72	+0.510	[−0.011, +1.031]	.055†
Flourishing	≥ 0	123	+0.236	[+0.017, +0.455]	.035*
	≥ 1	123	+0.236	[+0.017, +0.455]	.035*
	≥ 5	109	+0.164	[−0.069, +0.398]	.166
	≥ 10	91	+0.185	[−0.069, +0.438]	.152
	≥ 14	72	+0.199	[−0.093, +0.491]	.178
Support	≥ 0	123	−0.001	[−0.213, +0.210]	.989
	≥ 1	123	−0.001	[−0.213, +0.210]	.989
	≥ 5	109	−0.001	[−0.221, +0.219]	.990
	≥ 10	91	+0.017	[−0.221, +0.254]	.889
	≥ 14	72	+0.147	[−0.109, +0.404]	.255

Note. Subsample retains participants with $\geq k$ active days. ANCOVA controls Wave 1 baseline, age, sex, IG use, follower count. † $p < .10$; $* p < .05$; $** p < .01$; $*** p < .001$.

B.5 IPTW-Weighted Robustness Check

The love-plot analysis (Section 1.1; Figure A1) showed that follower count was the only covariate with a meaningful baseline imbalance (SMD = 0.29). To supplement the covariate-adjusted ANCOVA, we additionally estimated treatment effects using inverse probability of treatment weighting [IPTW; 5]. Propensity scores were estimated via logistic regression of treatment assignment on the four demographic covariates (age, sex, IG use, follower count). Average treatment effect (ATE) weights ($1/\hat{e}$ for treated, $1/(1-\hat{e})$ for control) were trimmed at the 99th percentile to reduce influence of extreme weights (range after trimming: 1.01–3.01).

IPTW main effects (Table B8). After weighting, the pattern of null main effects on well-being outcomes is unchanged. The support effect attenuates from $b = -0.140$ ($p = .046$) to $b = -0.123$ ($p = .072$), suggesting it is partially attributable to the follower-count imbalance and should be interpreted cautiously. H2 and H3 remain null under weighting.

IPTW interactions (Table B9). The treatment × IG-context neuroticism interaction was robust to demographic weighting for SPANE-P ($b = +0.346$, $p = .018$) and SPANE-B ($b = +0.469$, $p = .038$). The Flourishing interaction attenuated slightly to marginal ($b = +0.210$, $p = .065$), while the support interaction remained firmly null ($b = -0.007$, $p = .948$). These results support the conclusion that the neuroticism

moderation of well-being outcomes is not explained by the demographic imbalances observed at baseline.

Table B8 IPTW-Weighted Treatment Effects: All Pre-registered Outcomes ($n = 123$)

Outcome	b	SE	95% CI	p
Social connection				
Support (H1)	−0.123	0.068	[−0.257, +0.011]	.072†
Self-presentation				
Self-disclosure (H2)	+0.057	0.107	[−0.154, +0.268]	.596
Self-discrepancy[a] (H3)	−0.138	0.133	[−0.401, +0.124]	.299
Well-being				
SPANE-P (H4a)	−0.096	0.090	[−0.273, +0.081]	.286
SPANE-B (H4b)	−0.138	0.140	[−0.416, +0.140]	.327
Flourishing (H5)	−0.098	0.069	[−0.234, +0.038]	.157

Note. ATE weights derived from logistic propensity score model on demographic covariates (age, sex, IG use, follower count); weights trimmed at 99th percentile. † $p < .10$; ∗ $p < .05$.

Table B9 IPTW-Weighted Treatment × IG-Context Neuroticism Interactions ($n = 123$)

Outcome	b	SE	95% CI	p
Social connection				
Support (H1)	−0.007	0.108	[−0.220, +0.206]	.948
Well-being				
SPANE-P (H4a)	+0.346	0.145	[+0.060, +0.633]	.018*
SPANE-B (H4b)	+0.469	0.224	[+0.026, +0.912]	.038*
Flourishing (H5)	+0.210	0.112	[−0.013, +0.432]	.065†

Note. Positive b = higher IG-context neuroticism amplifies the demetrication benefit. ATE weights from demographic propensity scores (trimmed at 99th percentile). † $p < .10$; ∗ $p < .05$; ∗∗ $p < .01$.

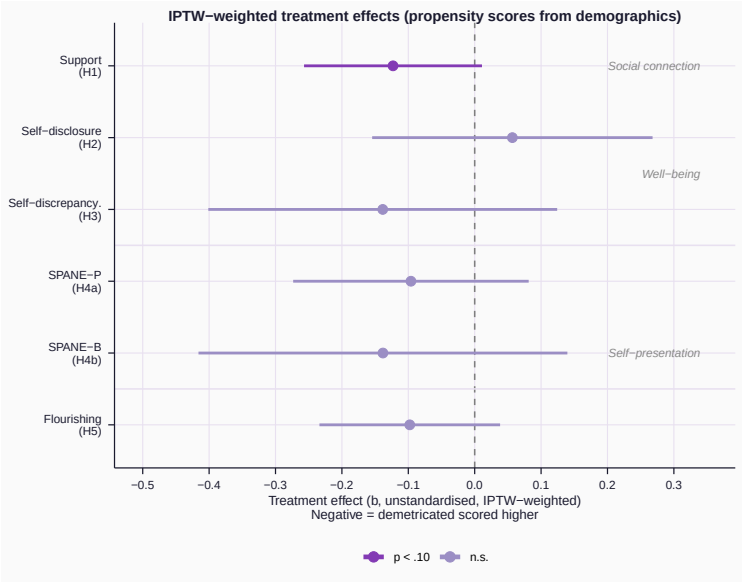

Fig. B7 IPTW-weighted treatment effects on all pre-registered outcomes. Propensity scores estimated from demographic covariates only (age, sex, IG use, follower count). Negative b = demetricated scored higher.

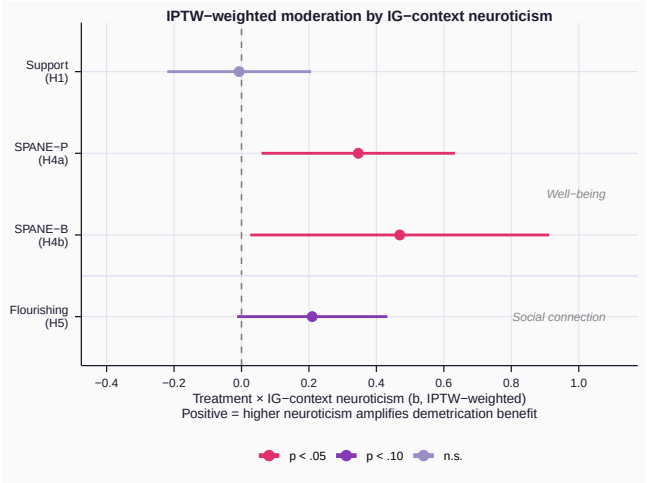

Fig. B8 IPTW-weighted treatment × IG-context neuroticism interactions. Positive b = higher IG-context neuroticism amplifies the demetrication benefit. Pink = $p < .05$; purple = $p < .10$.

B.6 Full Behavioral Outcome Comparisons

Table B12 and Figure B9 report behavioral event counts across all seven logged activity types. Every effect was in the same direction (demetricated users engaged less), though

only photo views reached conventional significance ($p = .032$) and comments viewed was marginal ($p = .073$). Non-parametric Wilcoxon rank-sum tests were consistent with the Welch t-tests; no additional effects reached $p < .05$ under the non-parametric approach, indicating that the pattern is not attributable to outliers driving parametric results.

The consistent negative direction across all behavioral categories—profile views ($d = -0.20$), photo views ($d = -0.35$), reels ($d = -0.18$), comments ($d = -0.29$), likes ($d = -0.20$), stories ($d = -0.23$), and story items ($d = -0.14$)—suggests a broad suppression of passive content consumption rather than metric-specific effects. A one-sample sign test on the seven effect sizes confirms that all seven are negative ($p = .016$, binomial).

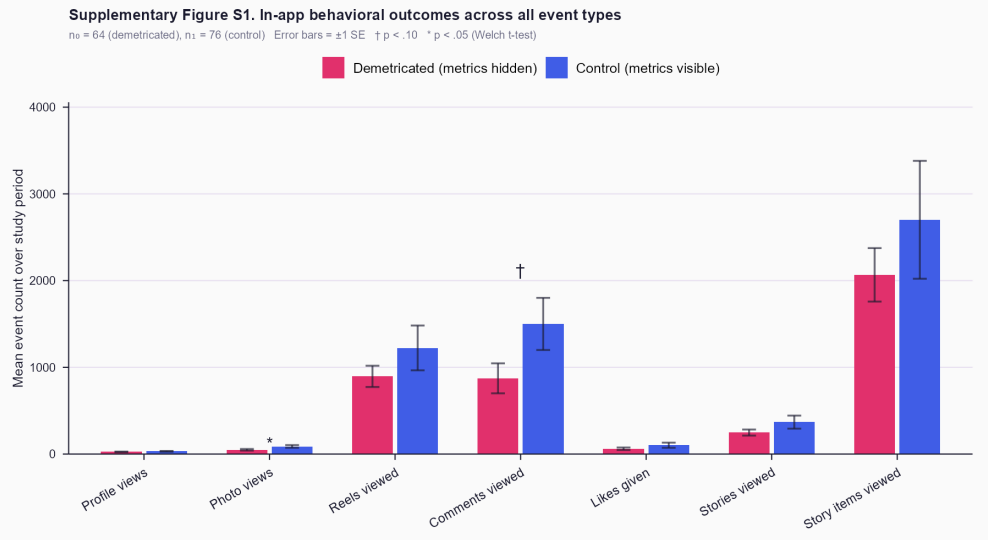

Fig. B9 Supplementary Figure S1. In-app behavioral outcomes across all seven event types. Mean event counts over the three-week study period by condition ($n_0 = 64$ demetricated, $n_1 = 76$ control), aggregated from Smolgram app logs. Error bars represent ± 1 standard error. $\dagger\, p < .10$; $*\, p < .05$ (Welch t-test). Cohen's d values range from -0.14 (story items) to -0.35 (photo views); all are negative.

References

[1] Austin, P.C.: Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. Communications in statistics-simulation and computation **38**(6), 1228–1234 (2009)

[2] Manski, C.F.: Nonparametric bounds on treatment effects. American Economic Review **80**(2), 319–323 (1990)

Table B10 Main Effects of Demetrication on Well-Being Outcomes (ANCOVA, $n = 123$)

Outcome	b	SE	t	df	p	95% CI
SPANE-P (positive affect)	−0.08	0.09	−0.86	116	.391	[−0.26, 0.10]
SPANE-B (affect balance)	−0.14	0.14	−0.96	116	.337	[−0.42, 0.15]
Flourishing	−0.10	0.07	−1.46	116	.146	[−0.24, 0.04]
Perceived social support	−0.14	0.07	−2.01	116	.046	[−0.28, −0.002]

Note. b = unstandardized regression coefficient for treatment (IG01 = control, metrics visible; IG00 = demetricated, metrics hidden). All models control for Wave 1 baseline score, age, sex, Instagram use frequency, and follower count. Negative b indicates lower scores in the control condition relative to the demetricated condition. CI = confidence interval.

Table B11 Treatment × IG-Context Neuroticism Interactions on Well-Being ($n = 123$)

	Interaction term					Simple slopes (b_{treat})		
Outcome	b_{int}	SE	t	df	p	−1 SD	Mean	+1 SD
SPANE-P	0.39	0.14	2.73	114	.007	−0.31	−0.07	+0.18
SPANE-B	0.50	0.22	2.28	114	.025	−0.43	−0.12	+0.19
Flourishing	0.24	0.11	2.13	114	.035	−0.24	−0.10	+0.05
Support	−0.001	0.11	−0.01	114	.989	−0.13	−0.13	−0.13

Note. b_{int} = unstandardized coefficient for the Treatment × IG-Context Neuroticism interaction. Simple slopes represent the estimated treatment effect (b_{treat}) at −1 SD, the mean, and +1 SD of IG-context neuroticism. All models control for Wave 1 baseline, age, sex, Instagram use frequency, and follower count. IG-context neuroticism was mean-centred prior to computing the interaction term.

Table B12 In-App Behavioral Outcomes by Condition ($n_0 = 64$, $n_1 = 76$)

	Demetricated		Control					
Outcome	M	SD	M	SD	t	df	p	d
Profile views	25.1	24.1	31.6	37.6	−1.23	129	.222	−0.20
Photo views	49.2	73.1	88.0	134.3	−2.17	119	.032	−0.35
Reels viewed	895.8	980.3	1223.7	2252.6	−1.15	106	.254	−0.18
Comments viewed	873.1	1382.1	1500.5	2620.8	−1.81	117	.073	−0.29
Likes given	61.2	114.6	102.0	258.2	−1.24	107	.218	−0.20
Stories viewed	247.9	278.7	368.3	656.2	−1.45	105	.150	−0.23
Story items viewed	2066.5	2465.7	2701.0	5925.8	−0.85	104	.397	−0.14

Note. Counts represent total app events logged over the three-week study period, aggregated from Smolgram app data. d = Cohen's d (pooled SD). Welch t-tests used due to unequal variances. All effects are in the same direction (demetricated < control). Wilcoxon rank-sum tests (not shown) were consistent with the t-tests; no effects reached $p < .05$ under Wilcoxon.

[3] Coppock, A.: attrition: An R Package for Analyzing Experiments with Attrition (2019). https://github.com/acoppock/attrition

[4] Lee, D.S.: Training, wages, and sample selection: Estimating sharp bounds on treatment effects. Review of Economic Studies **76**(3), 1071–1102 (2009)

[5] Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)