Supplementary Material

Separating experience reporting from content recall in dream reports across sleep stages

Peiyan Zhao

Supplementary Materials A–F accompany the main text.

A Report arithmetic and measurement assumptions (two tables) — p. 4

B Documentary evidence and native record checks (six tables and one figure) — p. 8

C Secondary model details (six tables) — p. 29

D Joint model and sensitivity specifications (42 tables and six figures) — p. 33

E Verification and reproduction — p. 78

F External source availability and exploratory comparisons (three tables) — p. 79

Reader map

Question or evidence	Where to look
Primary decomposition and contributions	Section A1, Table D8, Table D14, Table D40
Dataset audit and source evidence	Section B1, Table B1, Table B4, Table B6
Content-availability bounds	Section A3, Table A2, Table D22
Uncertain answers and boundary scenarios	Section D9, Section D11, Section D12.6, Table D37, Table D39
Omitting white dreams	Section D10, Section D12.5, Table D26, Table D36
Chronology and computation	Section D3.1, Section D3.2, Table D2, Table D3, Table D28, Table D38

Table and figure index

Item	Title	Page
Table A1	Sufficient assumptions for a latent interpretation	4
Table A2	Bounds on the content-availability gap within protocols	7
Table B1	Documentary eligibility and evidence for all 20 datasets	8
Table B2	Elicitation paths, report reference windows and observed counts	20
Table B3	Recoverable changes in documentary judgments	22
Table B4	Elicitation routes and negative meanings across all 20 datasets	23
Table B5	Reporting contrasts in the six initial EEG datasets	26
Table B6	Native transcript examples in Zhang–Wamsley	27
Figure B1	Sample flow and analysis samples	28
Table C1	Dataset-specific stage odds ratios from the six-dataset primary fit	30
Table C2	Samples for secondary logistic sensitivity analyses	30
Table C3	Stage differences in secondary logistic models	32
Table C4	Interval methods for participant-stratified logistic models	32

Item	Title	Page
Table C5	DATA1 timing and sequence adjustments in secondary logistic models	32
Table C6	DATA1 conditional white-dream proportions in N2 and N3	33
Table D1	Stage sample sizes within participant clusters	35
Table D2	Recoverable analysis chronology	36
Table D3	Sampling diagnostics for fitted models	38
Table D4	Participant and sequence posterior predictive checks	39
Table D5	Hierarchical means, dataset variation and new-dataset predictions	40
Table D6	DATA1 category and awakening-order support by night	41
Table D7	DATA1 contributions before and after timing adjustment	41
Table D8	Joint decompositions across analysis choices	42
Table D9	Complete leave-one-dataset-out refits	44
Table D10	Stage sample sizes for dataset clock terms	44
Table D11	Fixed weights for the six-dataset summary	45
Table D12	Standardised report probabilities by dataset and stage	46
Table D13	Observed pooling and standardisation of EWR	48
Table D14	Signed experience-report shares of the gap	49
Table D15	Standardised category probabilities and stage differences	49
Table D16	Difference between NE and EWR increases after NREM awakening	50
Table D17	Prior and posterior summaries of slope variation	51
Figure D1	Prior and posterior distributions of between-dataset slope variation	51
Table D18	Posterior differences in experience-report gap shares	52
Table D19	Observed and model-based contributions in all six datasets	53
Table D20	Category changes in every modelled dataset	55
Table D21	Fitted shares and common slopes under three anchors	55
Table D22	Selection lower endpoints in the four-dataset fit	56
Figure D2	Fitted shares and three anchoring choices	57
Figure D3	Secondary specifications and weighting targets	58
Table D23	Selected boundary scenarios in the six dataset equal weight summary	59
Table D24	Moved fractions at which the posterior mean of G reaches zero	60
Figure D4	Boundary grid for the six dataset equal weight summary	61
Figure D5	Separate consequences along the two specified paths	62
Figure D6	Dataset contributions and hierarchical stage associations	63
Table D25	Observed stage differences under two report contrasts	64
Table D26	Paired stage differences in the auxiliary joint models	65
Table D27	Weighting targets and marginal odds ratio comparisons	65
Table D28	Sampling diagnostics for the auxiliary models	66
Table D29	Observed stage support for the auxiliary ten-dataset sample	66
Table D30	Same-model refit restricted to participants observed at both stages	67
Table D31	Expanded-denominator scenarios for documented uncertain answers	68

Item	Title	Page
Table D32	Standardisation weight from wholly unobserved participant stages	69
Table D33	Contributions grouped by awakening procedure	69
Table D34	Additional model and standardisation targets	70
Table D35	Conditional white-dream count predictions	70
Table D36	Paired contrasts among sources storing EWR separately	71
Table D37	Uncertain-answer refits and their resolved-record baseline	72
Table D38	Diagnostics for the additional fits	73
Table D39	Uncertain answers by stage and participant cluster	74
Table D40	Contributions and shares under three decompositions	74
Table D41	Negative report shares of the fall in contentful reports	75
Table D42	New protocol predictions across slope priors	76
Table F1	Examples of unresolved requirements for external quantitative synthesis	79
Table F2	Listed or conditionally compatible counts in the exploratory sources	80
Table F3	Descriptive differences between original label increments	80

Supplementary Material A Report arithmetic and measurement assumptions

A1 An illustrative decomposition

Suppose, for illustration, that experience-report probabilities are 0.80 in REM and 0.60 in NREM, and conditional-content probabilities are 0.90 and 0.70. The contentful-report probabilities are 0.72 and 0.42. Their 0.30 difference comprises an experience-report contribution of (0.80 − 0.60) × (0.90 + 0.70)/2 = 0.16 and a content contribution of (0.90 − 0.70) × (0.80 + 0.60)/2 = 0.14.

The share also changes with baseline probabilities even when stage odds ratios are identical. For an odds ratio of two, a baseline probability p becomes $\frac{2p}{1 + p}$. With NREM experience-report and conditional-content probabilities of 0.50 and 0.80, the REM probabilities are 2/3 and 8/9. The experience-report and content contributions are 14.07 and 5.19 pp, giving a 73.08% experience-report share. Reversing the two NREM baselines reverses the contributions and gives a 26.92% share, while retaining an odds ratio of two for both probabilities.

The symmetric product decomposition follows the classical rate decomposition of Kitagawa (1955) and the two-factor presentation of Das Gupta (1993, Section 2.2). It is applied within each dataset and posterior draw, then summed with fixed weights. Supplementary Section D2 preserves the corresponding standardisation definitions.

A2 Experience occurrence and reporting

The probabilities concern reporting after awakening. Interpreting them as experience occurrence or content availability requires assumptions about memory, response criteria and elicitation (Supplementary Table A1). Assessing those assumptions requires evidence beyond category counts.

Supplementary Table A1 Sufficient assumptions for a latent interpretation

Observed probability	Sufficient assumptions	Quantity then identified
Experience report a	Every experience produces an experience report. No experience report occurs without an experience.	Frequency of experience occurrence within the protocol reference window
Conditional content c	Content available at report collection always yields E. Contentful false positives are absent. No experience report occurs without an experience. Experiences without available content always yield an experience report.	Frequency of content available at report collection among experiences

The assumptions are jointly sufficient for the stated interpretations. The following derivation considers weaker restrictions that can preserve the direction of a stage difference.

Within one protocol, let O indicate an experience in the report's reference window and let s denote the deposited last-stage label. Write $g_s = P(O = 1 \mid s)$. Let t_s denote the chance of reporting an experience when one occurred, and let f_s denote the chance when none occurred. The observable probability is

$$a_s = g_s t_s + (1 - g_s) f_s$$

Setting $t_s = 1$ and $f_s = 0$ is sufficient for $a_s = g_s$. With unknown reporting errors, the category counts do not identify g_s. If false reports are absent and $t_s = t > 0$ at both stages, the reporting difference has the same sign as the occurrence difference. Stage-dependent sensitivity or criteria can alter that relationship.

A3 Content availability and conditional reporting

Analysing only experience reports can select different subsets after REM and NREM awakening (Hernán et al., 2004). Our selection analysis asks whether differing tendencies to report an experience without available content could alone explain the conditional-content difference. The bounds use two reporting assumptions: (R1) experience reports are not false, and (R2) available content always produces an E report. This is partial identification under reporting assumptions (Manski, 2003): the identified region contains all compatible latent values. Posterior uncertainty in its endpoints is distinct from the region itself.

Within each protocol, let K indicate content available at report collection, conditional on an experience having occurred, and let $k_s = P(K = 1 \mid O = 1, s)$. Availability includes the effects of encoding and retrieval. To derive the bounds, first assume R2 and allow false experience reports only in EWR. Let m_s denote the probability of reporting an experience when O = 1 and K = 0. Then

$$p_s(E) = g_s k_s$$

$$p_s(EWR) = g_s(1 - k_s) m_s + (1 - g_s) f_s$$

$$1 - c_s = \frac{g_s(1 - k_s) m_s + (1 - g_s) f_s}{g_s k_s + g_s(1 - k_s) m_s + (1 - g_s) f_s}$$

Under these assumptions, $t_s = k_s + (1 - k_s) m_s$. Setting $f_s = 0$ removes false experience reports from the conditional ratio. With the same positive m_s at both stages, conditional content can preserve ordering even when its magnitude differs from k_s. The bound calculation additionally enforces the observed joint category probabilities.

With $f_s = 0$, the restricted model gives $p_s(E) = g_s k_s$, $p_s(EWR) = g_s(1 - k_s) m_s$ and $p_s(NE) = 1 - g_s[k_s + (1 - k_s) m_s]$. Consequently,

$$c_s = \frac{k_s}{k_s + (1 - k_s) m_s}$$

$$\operatorname{logit} k_R - \operatorname{logit} k_N = \operatorname{logit} c_R - \operatorname{logit} c_N + \log(m_R / m_N)$$

The ratio at which the latent contrast is zero is the NREM conditional-content odds divided by the REM odds:

$$r^* = \frac{c_N(1 - c_R)}{c_R(1 - c_N)}$$

This ratio must also be feasible for the joint distribution. Since $g_s = p_s(E) + p_s(EWR)/m_s$ cannot exceed one,

$$\frac{p_s(EWR)}{1 - p_s(E)} \leq m_s \leq 1$$

The compatible latent content probability therefore lies in $p_s(E) \leq k_s \leq c_s$. A zero REM–NREM contrast is feasible only when the two stage intervals overlap, and the contrast obeys

$$p_R(E) - c_N \leq k_R - k_N \leq c_R - p_N(E)$$

The bounds are calculated separately within each protocol. For a fixed-weight average $H = \Sigma_d w_d (k_{Rd} - k_{Nd})$, the compatible interval has endpoints $\Sigma_d w_d L_d$ and $\Sigma_d w_d U_d$, where $L_d = p_{Rd}(E) - c_{Nd}$ and $U_d = c_{Rd} - p_{Nd}(E)$. These averages concern within-protocol conditional probabilities and allow reporting tendencies to differ between protocols. Supplementary Table A2 separates the joint-mean configurations from posterior uncertainty in their endpoints.

Within each stage, the minimum compatible content availability treats every NE report as an experience whose content was unavailable. The maximum treats every NE report as no experience and requires every experience without available content to yield EWR. The lower bound on the REM-minus-NREM gap combines the REM minimum with the NREM maximum; the upper bound reverses these choices.

The bounds require R1 and R2. Response criteria, fatigue, reluctance or attribution to an earlier period can violate them.

Supplementary Table A2 Bounds on the content-availability gap within protocols

Dataset or weights	Bounds or mean endpoints (pp)	Lower endpoint 95% CrI	Upper endpoint 95% CrI
LODE	−11.0 to 38.5	−22.0 to −0.3	28.6 to 47.7
TWC USA	−9.3 to 28.9	−24.4 to 8.3	7.8 to 51.4
SCAN	11.9 to 42.1	−2.8 to 30.0	28.1 to 56.8
Sleep Talking	−35.2 to 36.5	−67.0 to −7.2	−3.8 to 63.7
DATA1	23.1 to 60.9	8.8 to 36.7	48.4 to 70.2
Aamodt evening	−15.0 to −2.9	−58.7 to 13.3	−49.1 to 25.1
Six datasets: record	2.9 to 38.6	−7.5 to 12.0	28.3 to 46.8
Six datasets: equal	−5.9 to 34.0	−16.9 to 4.1	22.2 to 43.7
Four datasets: record	9.7 to 49.8	1.4 to 17.5	42.4 to 55.9
Four datasets: equal	5.7 to 44.2	−1.8 to 13.8	36.4 to 51.9

Bounds assume R1 and R2 (Section A3). Individual-protocol rows use joint posterior mean probabilities from the six-dataset fit. Six- and four-dataset average rows use their respective fits and average within-protocol endpoints, allowing reporting tendencies to differ. Endpoint CrIs describe posterior uncertainty in the bounds.

Applying the formula directly to pooled report probabilities instead gives bounds of 1.5 to 38.3 pp with record weights and −6.1 to 34.0 with equal weights. Lower-endpoint CrIs are −8.8 to 10.5 and −17.0 to 3.9; upper-endpoint CrIs are 27.3 to 46.9 and 22.1 to 43.9. These pooled-mixture calculations use a different averaging target from Supplementary Table A2.

Under these assumptions, DATA1 requires more content available at collection after REM than NREM awakening across all compatible reporting tendencies. Its mean lower bound is 23.1 pp, with a 95% CrI of 8.8 to 36.7. SCAN's mean lower bound is 11.9 pp, with an interval of −2.8 to 30.0. LODE's bounds span −11.0 to 38.5 pp, allowing differential reporting as an explanation there (Supplementary Table A2).

Supplementary Material B Documentary evidence and native record checks

B1 Sources and audit decisions

Supplementary Table B1 gives the eligibility overview. ED01–ED20 identify the full deposited description files for the highest amendments in DREAM version 9 (Wong et al., 2025b). Line numbers run from the first line of each complete file, including metadata, headings and blank lines; they do not restart within the Experimental description field. ED identifiers follow Set IDs; the Set ID in each file header identifies it when archive folder numbers differ. Each entry adds the source study, population, design and NE meaning. Original papers establish study context, while contributor coding descriptions document the DREAM mapping. NF identifiers locate additional native files. Tables B4–B6 give the classification of elicitation routes and NE meanings, correspondence with the initial EEG datasets and native transcript examples.

The entries distinguish missing documentation from a lost category boundary and record the content threshold and assignment role. Participant-explicit means that the response conveys the participant's memory or impression; a researcher may assign its final code. Elicitation route and NE meaning are separate dimensions; neither determines category-contrast eligibility on its own. Judgments for Sets 11 and 14 have medium confidence. Unless an entry names a coder, the source materials do not identify the individual who assigned the final category code.

Supplementary Table B1 Documentary eligibility and evidence for all 20 datasets

Set	Dataset	Conditional content	Experience reporting	Evidence and response basis
1	Zhang–Wamsley	Ineligible	Ineligible	ED01 am 0 L79, 81–83 NF01 Basis: participant-explicit
2	Oudiette N1	Eligible	Eligible	ED02 am 0 L56–59 Basis: not documented
3	LODE	Eligible	Eligible	ED03 am 0 L58, 65 Basis: participant-explicit
4	TWC USA	Eligible	Eligible	ED04 am 1 L137 Basis: participant-explicit
5	Donders	Indeterminate	Indeterminate	ED05 am 0 L80 NF05 Basis: not documented
6	REM Turku	Eligible	Eligible	ED06 am 0 L137, 145, 147, 149 Basis: participant-explicit

Set	Dataset	Conditional content	Experience reporting	Evidence and response basis
7	Children dreaming	Ineligible	Eligible	ED07 am 0 L78 Basis: participant-explicit
8	SCAN	Eligible	Eligible	ED08 am 0 L66, 77–80 Basis: participant-explicit
9	Multiple awakenings	Ineligible	Eligible	ED09 am 0 L62, 67 Basis: participant-explicit
10	Sleep Talking	Eligible	Eligible	ED10 am 0 L67 Basis: participant-explicit
11	Young adults	Ineligible	Eligible	ED11 am 0 L68, 75 Basis: participant-explicit
12	Older adults	Indeterminate	Indeterminate	ED12 am 0 L74 NF12 Basis: not documented
13	Siclari serial	Eligible	Eligible	ED13 am 2 L76 NF13 Basis: participant-explicit
14	UFRN	Ineligible	Ineligible	ED14 am 0 L64, 73 Basis: participant-explicit
15	Motor tDCS	Ineligible	Ineligible	ED15 am 0 L66, 93 Basis: not documented
16	DATA1	Eligible	Eligible	ED16 am 0 L73, 79–82 NF16 Basis: participant-explicit
17	Aamodt evening	Eligible	Eligible	ED17 am 0 L111, 133 Basis: participant-explicit
18	Aamodt morning	Indeterminate	Eligible	ED18 am 0 L67 NF18 Basis: not documented

Set	Dataset	Conditional content	Experience reporting	Evidence and response basis
19	Kumral audiobook	Ineligible	Ineligible	ED19 am 0 L64, 69–71 Basis: participant-explicit
20	MEG Kyushu	Eligible	Eligible	ED20 am 0 L64, 69, 71–77 NF20 Basis: participant-explicit

Category-boundary decisions and evidence locators. ED identifies a deposited description, am an amendment, L a line and NF an additional source. NE wording and study context are given in the individual entries. Native-source checks covered Sets 1, 3, 5, 12, 13, 16, 18 and 20.

Set 1 Zhang–Wamsley (Zhang & Wamsley, 2019)

Source and design. Zhang and Wamsley (2019). Healthy young adults; overnight laboratory study following a spatial-learning or control task. Repeated sleep-onset probes, later N2/REM awakenings and a final morning report; experimenters called the participant's name. The negative code records absence of described mental content, without separately asking about a dream impression.

Source ED01, amendment 0, Experimental description, line 79: "Reports were coded for the presence of dream content".

Additional passage, lines 81–83: "Participants were not asked to assess whether".

Reports were evaluated for recalled mental content. The deposit says participants were not asked whether they had dreamed without recall. Codes 2 and 0 distinguish recalled content from its absence; no separate EWR code. This reflects the original study's intentional No Content grouping, which combined "I can't remember" and "There was nothing"; mapping it to NE does not restore a separate experience boundary (Zhang & Wamsley, 2019, Methods).

Assignment: Two judges, described as blind to stage. Content threshold: Any described mental content.

Both category contrasts are ineligible because the coding combines absence of recalled content with absence of reported experience. Native transcripts provide examples of experience without content recall among NE records (Table B6).

NF01: ExperimentalDescription.set01.txt, RecordsSet01.csv and Reports(1).csv; case identities 422_SO7 and 512_SO8; coding description DREAM categorization procedure.

Native record check. All 308 record identities and deposited fields matched the export. The 308 report texts were linked and the 70 NE texts examined. Cases 422_SO7 and 512_SO8 explicitly described experience without recalled content. Individual reports were retained as deposited.

Set 2 Oudiette N1 (Oudiette_N1Data)

Source and design. Lacaux, Andrillon, Arnulf and Oudiette, contributor description ED02 in Wong et al. (2025b). Adults aged 18–35; a 30-minute daytime nap between creativity tasks, with four alarm probes at six-minute intervals. NE denotes nothing having passed through the mind before the probe.

The deposit lists no original publication; a paper from the same research programme is not assumed to document these exact probes.

Source ED02, amendment 0, Experimental description, lines 56–59: “We classified participants' reports as:”.

Participants were asked to describe aloud what went through their mind before the alarm (ED02, line 52). The categorisation distinguishes something, nothing, and something without remembered mental content at the probe. Any clarifying follow-up and the basis of EWR assignment are unspecified. E, NE and WR map to E, NE and EWR.

Content threshold: A thought or image at the probe; WR denotes no recalled mental content.

The assignment definitions preserve both required boundaries. The participant’s role in establishing experience without content is unspecified.

Set 3 LODE

Source and design. Elce et al. (2023), LODE contributor deposit. Healthy Italian-speaking adults; at-home recordings with a portable electroencephalography (EEG) headband and morning voice reports over at least 14 days. Awakenings were spontaneous. The requested period was immediately before final awakening. NE denotes neither a memory nor an impression of an experience. The deposit is the direct source for these 190 records; overlap with the larger Elce et al. (2025) morning-report study is not assumed.

Source ED03, amendment 0, Experimental description, line 65: “All verbal reports are evaluated and categorized as:”.

Additional passage, line 58: “All participants are instructed to produce a verbal report”.

Participants were instructed to report each morning what had been going through their mind just before final awakening, including when they recalled no experience or could not remember its details. The coding was 2 = content; 1 = impression of experience without recalled content; 0 = no memory or impression of experience. E/EWR reports referring only to experiences earlier in the night were excluded. Most reports were collected within 10 min; contentful and white-dream reports were retained only within 1 h, whereas NE reports were explicitly exempt from this limit (ED03, lines 58, 86–88).

Assignment: Reports evaluated by researchers. Content threshold: At least one content of any kind, including a perception, thought or emotion.

The impression-without-content assignment rule preserves both boundaries. Participants were preinstructed to report absent experiences and white dreams; researchers assigned codes from those reports. No structured awakening question separating the two is documented. This report-based route is distinct from NE’s meaning, which excludes both memory and an impression of experience (Section B4).

Staging and channels. All recordings were reinspected after automated DREEM staging. The source lists seven final-epoch corrections from REM to N2, two from N2 to REM, one from N2 to N1 and one from N1 to N2 (ED03, lines 90–95), but no final-epoch agreement rate or confusion matrix. All

190 native records flag separate EOG and EMG channels as absent. Records.csv has no report-delay variable, so the effect of the asymmetric one-hour retention rule cannot be estimated record by record.

The local identity check matched all 190 supplied native Case IDs to the locked v9 records and found no differences in Filename, Subject ID, Experience, Last sleep stage, Duration, EEG sample rate or Number of EEG channels after the recorded normalisations. The saved six-dataset model used the same record identities. The supplied files were checked against the DREAM export. Their provenance before supply and participant overlap with Elce et al. (2025) were not established.

Set 4 TWC USA (TWC_USA)

Source and design. Konkoly et al. (2021), USA procedure; contributor description ED04. Young adults recalling at least one dream per week; morning laboratory naps with targeted lucidity cues, eye signals and attempted two-way communication. The deposit includes experimenter-induced and spontaneous awakenings. NE combines no remembered material with the participant's belief that sleep did not occur. The deposit contains 19 participant identities; the published USA recruitment total was 22.

Source ED04, amendment 1, Experimental description, line 137: "Karen Konkoly rated reports according to the DREAM categorization."

The stored coding is based on whether the participant reports mental content, a remembered dream without its content, or remembering nothing. 2 = any mental content; 1 = remembered dream without content; 0 = remembering nothing or not falling asleep.

Assignment: Karen Konkoly. Content threshold: Any sort of mental content from sleep.

The separate impression-without-content rule preserves both coded contrasts. NE also includes perceived failure to sleep. The five retained NE records lack original answers separating this from absent recall, motivating the documentary-restriction refit in Section D12.3.

Set 5 Donders (Dream Database from Donders)

Source and design. Konkoly et al. (2021), Netherlands procedure, for the identified dialogue subset; ED05 for additional pilots. A mixed collection of lucid-dream communication, dreamed hand-movement and unspecified pilots, mainly in morning naps. Participants arrived with high REM pressure; the deposit does not provide one uniform recruitment or awakening rule. The linked article does not validate the uncrosswalked pilot records.

Source ED05, amendment 0, Experimental description, line 80: the categorization-procedure field reads "N/A".

Assignment: Study-specific response and category-assignment routes for negative or uncertain reports are undocumented. Content threshold: Not documented.

Both category contrasts are indeterminate.

NF05: ExperimentalDescription.Set5.txt, Records.Set5.csv and README.Set5.txt; categorization-procedure field and Case 3 stage field.

Native record check. Seven record identities represented five participant clusters and all were E. One native stage was blank where the export labelled Wake; split recording parts did not independently establish separate awakenings. These issues were retained without recoding.

Set 6 REM Turku (REM_Turku)

Source and design. Sikka et al. (2019); contributor description ED06. Healthy Finnish-speaking adults with good sleep quality; two laboratory nights. Tone-induced reports followed five minutes of continuous phasic REM sleep and included a morning report. A negative initial answer was followed by a question distinguishing no dream experience from an impression without specific content. NE corresponds to the former response.

Source ED06, amendment 0, Experimental description, line 145: “Dream experiences = participants remembered having a dream”.

Additional passage, line 137: “In case the participants reported”.

Additional passage, line 147: “Without recall (”.

Additional passage, line 149: “No recall = participants”.

After an initial no-dream response, researchers asked whether the person had no dream or felt they had dreamed without specific content recall. 2 = dream with some content; 1 = felt dream without specific content; 0 = no dream experience.

Assignment: Participant responses obtained by researchers. Content threshold: At least some specific content.

The explicit follow-up to a negative response preserves both boundaries. This REM-only dataset contributes no main REM–NREM stage difference.

Set 7 Children dreaming (ChildrenDreaming)

Source and design. Reda et al. (2021), covering some contributed recordings; ED07 for the DREAM mapping. Children and adolescents aged 9–14 with developmental dyslexia or control status; overnight laboratory recording with a morning sleep and dream diary after ten hours of sleep or spontaneous awakening. The negative category includes no recall or uncertainty without a feeling of having dreamed. The positive code combines contentful and white-dream reports.

Source ED07, amendment 0, Experimental description, line 78: “2 = Experience or Experience without recall”.

The diary records a feeling of having dreamed; lack of recall or an I-do-not-know response without that feeling forms the negative category. Code 2 merges experience with and without content recall; code 0 retains no clear feeling of dreaming.

Assignment: Participant diary response with deposited mapping. Content threshold: No separate content threshold can be recovered after the positive categories are merged.

Experience reporting is eligible because the positive code retains reports with and without content. Merging those reports prevents the content contrast.

Set 8 SCAN (SCANDataset)

Source and design. Eichenlaub et al. (2018); contributor description ED08. Healthy students selected for frequent dream recall; a laboratory night after a ten-day waking-life diary. A buzzer triggered reports after ten minutes of N3 or REM sleep. NE is defined by not remembering anything, whereas EWR requires a reported impression of dreaming. DREAM retains 18 participant clusters from a study that recruited 20. The paper's schedule allowed a third SWS awakening if the first two failed to yield recall; later probes targeted REM. SWS awakenings occurred earlier than REM (reported means 110.0 and 347.9 min after sleep onset), coupling stage and time.

Source ED08, amendment 0, Experimental description, lines 77–80: "The dream reports were rated as following:".

Additional passage, line 66: "What was going through your mind immediately before you were woken up?"

An initial prompt asks what was going through the mind immediately before awakening, followed by a request for any further remembered material. 2 = any mental content; 1 = participant-reported impression of a dream without content; 0 = not remembering anything.

Assignment: Researcher ratings. Content threshold: Any sort of mental content from sleep.

The participant-impression rule preserves both category contrasts. The EWR/NE mapping is documented in contributor ED08. The paper reports 48 REM awakenings with 46 dream reports; two awakenings with reports were excluded from its EEG analyses because of poor-quality EEG. The remaining EEG sample contained 46 REM awakenings (44 with reports) and 39 SWS awakenings (19 with reports), matching the deposit (Eichenlaub et al., 2018, Results).

Set 9 Multiple awakenings

Source and design. Scarpelli, Bartolacci, et al. (2020); contributor description ED09. Young adults in two laboratory nights with repeated REM and NREM awakenings after five minutes of stable stage; participants were called by name over an interphone. NE represents a negative experience response; uncertain answers coded jointly as NE/EWR are separately excluded.

Source ED09, amendment 0, Experimental description, line 67: "2 = dream experience, 0 = No dream experience".

Additional passage, line 62: "We did not collect dream contents".

Participants reported whether they had or had not a dream experience immediately before awakening. 2 = experience; 0 = no experience; -2 = combined no experience/experience without recall, excluded.

Assignment: Participant categorical response. Content threshold: Content was not collected, so an E-versus-EWR threshold is unavailable.

Experience reporting is eligible among resolved classifications; conditional content is ineligible. A positive deposited label in this dataset does not establish provision of content.

Set 10 Sleep Talking

Source and design. De Gennaro and Scarpelli, contributor description ED10 in Wong et al. (2025b); related protocol in Mangiaruga et al. (2022). Frequent sleep talkers aged 19–30; adaptation and

experimental laboratory nights with uninterrupted sleep followed by a morning diary. The final awakening followed about seven hours of undisturbed sleep; its induction method and the diary's exact reference period are not specified. NE includes no recall or uncertainty without a clear dream impression. ED10 lists no previous publication. Mangiaruga et al. describes matching recruitment and recording procedures but 13 participants, whereas this deposit has 12; an exact participant/record crosswalk is unavailable.

Source ED10, amendment 0, Experimental description, line 67: "2 = Experience with content recall".

Diary reports distinguish recalled content, a feeling of dreaming without content, and no clear feeling of dreaming. 2 = content report; 1 = feeling of dream without content recall; 0 = no recall or uncertainty without a clear feeling of dreaming.

Assignment: Participant diary basis. Content threshold: At least one dream report with content recall; a more detailed threshold is not specified.

The diary preserves both coded contrasts but its reference-period instruction is unavailable. A report after about seven hours of uninterrupted sleep may refer to an earlier period than the final scored stage. Section D12.3 excludes Sleep Talking together with TWC USA. The four-dataset subset excludes it with Aamodt evening.

Set 11 Young adults (Dream_YoungAdults)

Source and design. Marzano et al. (2011); contributor description ED11. Healthy university students selected from several recording protocols; adaptation and undisturbed laboratory nights followed by an experimenter-induced morning awakening after five stable minutes of N2 or REM. The diary asks about the feeling of having dreamed. The negative code is absence of dream recall under that instruction; EWR is not retained separately. The ad hoc recruitment and recall-group balancing limit population interpretation.

Source ED11, amendment 0, Experimental description, line 75: "Dream Recall = 2; Absence of dream recall = 0".

Additional passage, line 68: "namely the feeling of having or having not dreamt".

The morning diary is described as measuring the feeling of having or not having dreamed. Dream Recall = 2; Absence of dream recall = 0. No separate EWR classification is retained.

Assignment: Participant diary. Content threshold: No separate content classification.

Experience reporting is eligible on the documented feeling-of-dreaming basis; conditional content is ineligible. The experience-boundary judgment has medium confidence.

Set 12 Older adults

Source and design. Scarpelli et al. (2019) and Scarpelli, D'Atri, et al. (2020); contributor description ED12. Older adults screened for cognitive, neurological, psychiatric and sleep disorders; an overnight laboratory study. Experimenters called the participant's name in the morning after five stable minutes of N2 or REM and requested all remembered mental activity.

Source ED12, amendment 0, Experimental description, line 74: the categorization-procedure field reads “N/A”.

The deposit contains E and NE and states that no white dreams were reported.

Assignment and content threshold: The sources do not establish how the original recall/non-recall (REC/NREC) codes map to DREAM categories, an EWR assignment rule or the recalled-content threshold needed for the two contrasts.

Both category contrasts are indeterminate.

NF12: ExperimentalDescription.set12.txt, Records.set12.csv and Records-README.set12.txt; description lines 65–74 and the generic code dictionary.

Native record check. Forty native records matched the export: 18 E and 22 NE, with 20 N2 and 20 REM records.

Set 13 Siclari serial (Tononi Serial Awakenings)

Source and design. Siclari et al. (2017); contributor description ED13. Adult volunteers in a serial-awakening laboratory paradigm with immediate experience interviews. The deposited description reuses the experiment 1 protocol; its retained records are N2 only. NE is the explicit no-experience category of the three-way interview. The expanded deposit is not mapped case by case to the published experiments.

Source ED13, amendment 2, Experimental description, line 76: “We have converted dream categories in Records.csv as DE = Experience”.

The linked study explicitly asks about experience with remembered content, experience without remembered content, and no experience. DE = E; DEWR = EWR; NE = NE.

Assignment: Structured participant report. Content threshold: Remembered content of the most recent experience; the three report states are distinguished.

The original three-way interview and deposited crosswalk support both category contrasts (Siclari et al., 2017, first Results paragraph). The native records are all N2, so they do not contribute a REM–NREM stage difference.

NF13: ExperimentalDescription13.txt, Records13.csv and AdditionalData.csv; coding procedure, line 76; Siclari et al. (2017), first Results paragraph.

All 287 amendment-2 records matched: 124 E, 56 EWR and 107 NE, all N2. The deposited procedure reproduces experiment 1 in Siclari et al. (2017); amendment 2 added subjects 37–39. The paper states that experiment 2 shared no participants with experiments 1 or 3 (Methods, “Study participants in experiment 2”). Overlap would therefore require the expanded Set 13 deposit to include experiment 2 participants, whose case assignments are not specified. Set 13 contributes no records to the six-dataset REM–NREM model. Additional content data cover E reports and do not recover EWR subtypes.

Set 14 UFRN (Brain Institute - Federal University of Rio Grande do Norte)

Source and design. Araujo, Palhano-Fontes, Andrade and Mota-Rolim, ED14 in Wong et al. (2025b); associated trial publication Palhano-Fontes et al. (2019). Overnight hospital recordings; the deposit

describes 41 participants from a control group of an associated trial. Experimenters awakened them near 06:00 and asked if a dream was remembered before requesting a report. The negative category is no remembered dream. ED14 states that these EEG data were unpublished; the trial article is credited as associated context, not as verification of this 41-person EEG sample or its treatment status.

Source ED14, amendment 0, Experimental description, line 64: “After awakening, subjects were then asked if they remembered a dream.”

Additional passage, line 73: “All dream reports were categorized”.

The documented initial question asks whether a dream is remembered; a verbal report is requested after a positive answer. Reports are described using recalled dreams and non-recall; no separate route for experience without recalled content is documented.

Assignment: Participants provide free reports, subsequently categorised. Content threshold: Content is requested after a remembered-dream response; a minimum content threshold is not documented.

Both category contrasts are ineligible under the audit’s medium-confidence interpretation that this recall-gated procedure retains a recall/non-recall boundary. Its medium confidence reflects the missing explicit EWR assignment rule.

Set 15 Motor tDCS (Noreika_Motor_tDCS)

Source and design. Noreika et al. (2020); contributor description ED15. Ten participants in two non-consecutive laboratory nights with REM sensorimotor tDCS or sham stimulation. A tone awakened them one minute after stimulation for a free report and bodily-experience questionnaire. The deposited collection is selected for dream-containing reports, so it supplies no usable negative-response denominator.

Source ED15, amendment 0, Experimental description, line 93: “White dream reports”.

Additional passage, line 66: “All reports contain dreams”.

Free reports were followed by content questions. The original analysis explicitly excluded white-dream and non-perceptual mentation reports. The deposited category account describes dream-containing reports only.

Assignment: Researcher collection and selection; exact DREAM assignment role not specified. Content threshold: Selected reports containing dream experience and content.

The documented selection of dream-containing reports prevents either category contrast.

Set 16 DATA1

Source and design. Wong et al. (2020); Noreika et al. (2009); Noreika (2023). Healthy Finnish-speaking adults retained after an adaptation-night screen for sleep and clear reporting; ten participated in four experimental nights. Experimenter-induced serial awakenings sampled the first three to four hours of sleep. The classification remains a retrospective participant report. The original target was sustained early-night N2/N3, not REM. Wong et al. (2020), Supplementary Document 1, reports sleep-onset REM intrusions under this serial-awakening protocol.

Source ED16, amendment 0, Experimental description, line 73: “All collected dream reports were divided by two blind raters”.

Additional passage, lines 79–82: “- Dreamless sleep -> No experience”.

Following no recalled dream, the white-dream item (item N2 in the source questionnaire) asks about an impression of a dream experience without specific content recall. Dreamful = E; white dream = EWR; dreamless = NE; uncertain = combined NE/EWR, excluded.

Assignment: Two study raters classified participant reports; disagreements were discussed, sometimes with a third rater. Content threshold: Any reportable experience, including perceptions, sensations, thoughts or feelings; EWR has no specific recalled content.

The interview and assignment rules support both category contrasts.

NF16: ExperimentalDescription18.txt and Records18.csv identify DATA1 (Set 16). Reports.csv is archived in Noreika (2023), Version 1. Wong et al. (2020), Supplementary Document 2, p. 2, gives the white-dream item; Supplementary Document 1, p. 2, gives the content-analysis procedure. Files named Set18eighteen instead identify the separate Aamodt morning cohort.

The native file and report checks matched all 324 cases: 107 E, 63 EWR, 118 NE, 34 uncertain answers and two Unknown. Two blank native stages corresponded to deposited Wake labels; both were excluded from the REM–NREM analyses. Reports.csv supplied night, awakening order and classification fields rather than all verbatim interviews. The elapsed-lights-off values had an unspecified absolute unit and did not identify sleep onset. The original study’s rater agreement concerns that classification procedure. The original paper’s selected analysis sample (nine participants) is not the entire ten-participant deposit. No current case-level SOREM marker confirms the origin of each of the 42 classified REM records. The post hoc N2/N3 comparison and its denominators are reported in Table C6.

Set 17 Aamodt evening (Aamodt_evening_sleep)

Source and design. Aamodt et al. (2023); contributor description ED17. Adults underwent evening laboratory sleep after rising three hours earlier than usual. Remote electronic alarms prompted repeated immediate interviews, mainly in N2. The questionnaire explicitly distinguishes no experience from experience without recall. The study recruited 27 participants; 25 participant clusters contribute to the retained sleep sample, with only two REM records.

Source ED17, amendment 0, Experimental description, line 133: “Based on the answer to Question 1, awakening reports were classified”.

Additional passage, line 111: “Describe the most recent experience”.

Question 1 asks about the last thing going through the mind; if needed, participants disambiguate NE and DEWR before the interview continues. Any explicitly recalled content = DE/E; reported dream without specific remembered content = DEWR/EWR; no experience = NE.

Assignment: Interview reports classified under the stated rule. Content threshold: Explicit recall of any dream content at all.

Both report boundaries are documented in Aamodt et al. (2023, Section 4.4) and the evening-cohort deposit. Only two retained records are REM. The cohort remains in the six-dataset primary fit but is excluded from the four-dataset subset.

Set 18 Aamodt morning (Aamodt_morning_sleep)

Source and design. Aamodt et al. (2021); contributor description ED18. Healthy adults underwent morning laboratory sleep after a full night of sleep deprivation. Electronic alarms prompted repeated immediate interviews, mainly in N2. NE is labelled no experience.

Source ED18, amendment 0, Experimental description, line 67: “What was the last thing going through your mind?”

The description lists NE/DEWR/DE, which are deposited as DREAM categories.

Assignment and content threshold: The morning-cohort DEWR assignment rule and the E/EWR boundary are undocumented. The evening-cohort questionnaire concerns a separate experiment.

Experience reporting is eligible, while conditional content is indeterminate.

NF18: ExperimentalDescription.Set18eighteen.txt, Records.Set18eighteen.csv and README.Set18eighteen_.txt; description lines 67–69 and 94–96, for the morning cohort.

Native record check. All 97 morning-cohort records matched the export: 60 E, 21 EWR, 14 NE and two Unknown. Published N2 category counts also differed from the deposit and could not be reconciled record by record.

Set 19 Kumral audiobook (Kumral audiobook study)

Source and design. Kumral et al. (2025); contributor description ED19, named Kumral et al. 2023 after the preprint. Healthy young adults listened to assigned audiobooks while falling asleep. Following an adaptation night, experimenters awakened them approximately every 90 minutes during the experimental night and repeatedly prompted mental-content recall. No separate question assessed a dream impression without recalled content; the negative code therefore cannot establish absence of experience.

Source ED19, amendment 0, Experimental description, lines 69–71: “We did not assess whether participants might have been dreaming”.

Additional passage, line 64: “They were asked up to three times”.

Repeated prompts request remembered mental content, followed by a detailed dream report when recall is present. 2 = recalled experience; 0 = no experience. Dreaming without content recall was explicitly not assessed.

Assignment: Participant verbal reports with deposited assignment account. Content threshold: Any recalled mental experience.

The study did not assess whether participants had dreamed without recalling content, preventing recovery of both required boundaries. Both category contrasts are ineligible. The deposited name

Kumral et al. 2023 refers to the earlier preprint; the corresponding peer-reviewed article is Kumral et al. (2025).

Set 20 MEG Kyushu

Source and design. Motomura, Takeichi, Ishii, Tsutsumi, Kawabata, Takahashi, Hosogai, Shigeto and Hiramatsu, contributor description ED20 in Wong et al. (2025b). An evening hypnagogic MEG/EEG pilot with repeated N1/N2 probes; one participant is represented in the deposit. Brief beeps prompted interviews, while spontaneous signals could also trigger reporting. The native no-dream branch asks about dreaming without remembered content. NE is the negative experience code after that route; the deposit lists no original publication.

Source ED20, amendment 0, Experimental description, line 69: “We defined ‘dreams’ as any mental experience during sleep”.

Additional passage, line 64: “The interview questions are described in the file”.

Additional passage, lines 71–77: “* 2 = Experience”.

The questionnaire first asks whether the participant was dreaming; its no-dream branch asks whether they were dreaming but unable to remember content. 2 = E; 1 = EWR; 0 = NE; combined and unknown codes are separately listed and are excluded when unresolved.

Assignment: Participant interview basis. Content threshold: Specific dream content elicited through open description and sensory follow-ups.

The negative-response follow-up supports both category contrasts (List of questions.docx, paragraphs 18–19).

NF20: ExperimentalDescription.txt, Records.csv and List of questions.docx; questionnaire paragraphs 18–19, after the initial no-dream answer.

Native record check. The 31 records contained 24 E, four EWR and three Unknown. Twenty-five classified sleep records remained, all NREM and all E-or-EWR, providing no REM–NREM stage difference. The full-deposit counts differ from the analysis counts for this reason.

B2 Elicitation paths and report reference windows

Supplementary Table B2 Elicitation paths, report reference windows and observed counts

Set	EWR elicitation and assignment	Report reference window	E/EWR/NE by stage	Records by stage	EWR among experience reports (%)
2	Open description before probe; three-way coding rule; individual coder unspecified. Basis: undocumented	Before the nap probe alarm (ED02 L52, L57–59).	NREM: 23/8/6	NREM: 37	NREM: 25.8

Set	EWR elicitation and assignment	Report reference window	E/EWR/NE by stage	Records by stage	EWR among experience reports (%)
3	Free report and impression of dreaming without content; researcher coding. Basis: participant response	Immediately before final awakening; earlier-only E/EWR excluded. Most ≤10 min; E/EWR ≤1 h; NE explicitly exempt from 1-h rule.	NREM: 72/11/44 REM: 46/3/14	NREM: 127 REM: 63	NREM: 13.3 REM: 6.1
4	Initial question wording unspecified; reports classified; Karen Konkoly (researcher coding). Basis: participant response	Linked to preceding continuous sleep. Ambiguous sleep-period links excluded; exact prompt unspecified (ED04 L130–132).	NREM: 8/0/4 REM: 18/1/1	NREM: 12 REM: 20	NREM: 0.0 REM: 5.3
6	Follow-up after a negative initial response; coding role not specified. Basis: participant response	Last image before awakening, then the whole most recent dream (ED06 L137–140).	REM: 123/8/3	REM: 134	REM: 6.1
8	Open report and further recall assessment; study raters. Basis: participant impression	Immediately before awakening; further recall prompted (ED08 L66).	NREM: 19/7/13 REM: 44/0/2	NREM: 39 REM: 46	NREM: 26.9 REM: 0.0
10	Diary including a feeling of dreaming; individual coder unspecified. Basis: participant response	Morning diary after final awakening. Exact reference period unspecified; whole-night recall unconfirmed (ED10 L67, L77).	NREM: 9/0/8 REM: 2/1/2	NREM: 17 REM: 5	NREM: 0.0 REM: 33.3
13	Explicit three-way question; source categories mapped to DREAM; individual coder unspecified. Basis: participant response	Most recent experience before alarm, linked experiment 1 protocol (ED13 L67–69; Siclari 2017 Methods/Results).	NREM: 124/56/107	NREM: 287	NREM: 31.1
16	No recalled dream followed by impression question; two study raters. Basis: participant response	Most recent experience/dream before awakening (NF16; Wong et al., 2020 Supplementary Document 2, p. 2).	NREM: 70/62/112 REM: 37/1/4	NREM: 244 REM: 42	NREM: 47.0 REM: 2.6
17	Open question with NE/DEWR clarification if needed; coder not specified. Basis: participant response	Most recent experience before alarm, after falling asleep; immediate interview (ED17 L107–111).	NREM: 108/26/15 REM: 1/1/0	NREM: 149 REM: 2	NREM: 19.4 REM: 50.0

Set	EWR elicitation and assignment	Report reference window	E/EWR/NE by stage	Records by stage	EWR among experience reports (%)
20	Initial no-dream response followed by recall question; coder not specified. Basis: participant response	Experience before awakening during sleep-onset trials (ED20 L64; NF20 questionnaire).	NREM: 22/3/0	NREM: 25	NREM: 12.0

Counts are E/EWR/NE within the indicated stage after analysis exclusions; the percentage is EWR/(E + EWR). All 10 content-eligible datasets are shown, including single-stage datasets and small samples. NREM pools available substages. The reference-window column distinguishes the period requested from collection timing; an unspecified period is not inferred to be the whole night. ED identifiers and line numbers identify source descriptions in Section B1. Elicitation and category assignment are recorded separately.

B3 Audit decision history

Supplementary Table B3 Recoverable changes in documentary judgments

Set	Change	Date	Triggering evidence
1	Experience reporting eligibility: indeterminate to ineligible	9 Oct 2026	Native reports 422_SO7 and 512_SO8 describe experience without content recall but are deposited as NE.
12	Experience reporting eligibility: eligible to indeterminate	9 Oct 2026	The supplied REC/NREC coding lacks a study-specific mapping to experience without recalled content.
16	Response basis: rater-inferred to participant-explicit; eligibility unchanged	1 Oct 2026	Wong et al. (2020), Supplementary Document 2, p. 2, the white-dream item, explicitly asks about an impression of dreaming; Document 1 specifies researcher assignment.

The table records dated, source-supported revisions to documentary judgments.

B4 Elicitation routes and negative report meanings

Main Table 3 and Supplementary Table B4 separate the route used to elicit or assign reports from the meaning of NE. NE can additionally include uncertainty or perceived wakefulness regardless of route. Content-based procedures with a lost boundary and dream-selected samples are identified separately. The initial-question column records whether the prompt refers to dreams or to mental activity, including the broader definitions supplied to participants.

Supplementary Table B4 Elicitation routes and negative meanings across all 20 datasets

Dataset	Elicitation route	Meaning of NE	Initial question target	Procedure and evidence
1 Zhang–Wamsley	Content-based coding	Original No Content merges absent recall and nothing experienced; two EWR-like native reports map to DREAM NE.	Any mental activity; recall prompt	Content description requested; dream-without-recall not assessed. ED01 L73, 79, 81–83; NF01; Table B6.
2 Oudiette N1	Route unresolved	Three coded categories distinguish nothing through the mind from experience without content.	What went through the mind	Three-way coding documented; participant-level impression probe not established. ED02 L52, 56–59.
3 LODE	Report-based coding	NE: no memory or impression; EWR: an impression without content.	Everything through the mind before awakening	Explicit preinstructions request absent-experience and white-dream reports. Researchers classify participant memory/impression; no structured distinction within an awakening interview documented. ED03 L58, 65; Section B1.
4 TWC USA	Report-based coding	NE combines nothing remembered with perceived failure to fall asleep.	Initial wording not reported	Explicit impression without content retained as EWR; no separate probe route established. ED04 am 1 L137.
5 Donders	Route unresolved	Current negative-code meaning cannot be verified.	Not documented	Classification procedure not supplied; native codes do not resolve the mapping. ED05 L80; NF05.
6 REM Turku	Structured question	NE: denied experience; EWR retained after distinguishing it from initial “no dreams”.	Last image, then whole dream	Fixed follow-up after a negative answer asks whether there was no experience or a dream without recall. ED06 L137, 145, 147, 149.
7 Children dreaming	Report-based coding	NE combines no recall with “I don’t know” without a clear feeling of dreaming.	Feeling of dreaming; initial wording unreported	A feeling without content joins positive reports; E/EWR are not separable. ED07 L78.
8 SCAN	Report-based coding	NE: not remembering anything; a dream impression without content is EWR.	What was going through the mind	“Can you remember anything else?” is documented, not an impression-specific negative-branch probe. ED08 L66, 77–80.

Dataset	Elicitation route	Meaning of NE	Initial question target	Procedure and evidence
9 Multiple awakenings	Route unresolved	NE: negative experience judgment; combined NE/EWR answers are excluded.	Dream experience; defined as any distinct mental activity	Experience yes/no collected without content; an impression-specific negative follow-up is not documented. ED09 L62, 67.
10 Sleep Talking	Report-based coding	NE combines no recall with uncertainty without a clear dream impression.	Feeling of dreaming; initial wording unreported	Separate feeling-without-content rule; exact follow-up route and report reference period unresolved. ED10 L67, 77.
11 Young adults	Report-based coding	NE: absence of dream recall under a feeling-of-dreaming instruction.	Feeling of having dreamed	Morning diary elicits a feeling of having or not having dreamt; positive feeling can lack content. Medium-confidence experience boundary. ED11 L68, 75.
12 Older adults	Route unresolved	REC/NREC codes lack a verified E/EWR/NE interpretation.	All remembered mental activity	Description and native metadata do not establish the response or assignment route. ED12 L74; NF12.
13 Siclari serial	Structured question	Explicit no-experience alternative distinguished from experience without recall.	Most recent experience	Three alternatives: experience, experience without content, no experience. This is not evidence of a sequential negative follow-up. ED13 am 2 L76; NF13; Siclari et al. (2017), first Results paragraph.
14 UFRN	Content-based coding	Negative dream-recall answer; an experience-without-content boundary is not secured.	Whether a dream is remembered	Initial “remember a dream?” gate, with free report only after yes. ED14 L64, 73.
15 Motor tDCS	Dream-selected sample	Retained reports are selected dreams; general negative responses are not represented.	Everything through the mind, then bodily experience	Two white dreams and three non-perceptual reports excluded by the source procedure. ED15 L66, 93.
16 DATA1	Structured question	Dreamless/no-experience response distinguished from white dream; uncertain answers excluded.	Dream experience; defined as every image or thought	Fixed impression follow-up after a negative answer: the white-dream item in the linked interview. ED16 L73, 79–82; NF16; Wong et al. (2020), Supplementary Document 2, p. 2.

Dataset	Elicitation route	Meaning of NE	Initial question target	Procedure and evidence
17 Aamodt evening	Structured question	NE and experience without recall are distinguished; NE is an explicit no-experience answer.	Last thing through the mind; any mental activity	NE versus EWR (the source's DEWR) clarified "if necessary", not a documented fixed follow-up for every negative answer. ED17 L107–111, 133.
18 Aamodt morning	Route unresolved	Experience boundary retained; exact morning NE/EWR route remains unresolved.	Last thing through the mind	Three categories mentioned, but the evening clarification route is not imputed to the morning protocol. ED18 L67; NF18.
19 Kumral audiobook	Content-based coding	Negative mental-content recall; dreaming without recall not assessed.	What was going through the mind; recall prompts	Content questions could be repeated up to three times; no separate EWR category. ED19 L64, 69–71.
20 MEG Kyushu	Structured question	Negative experience code after a branch asking about dreaming without recall.	Dreaming; defined as any mental experience	Fixed negative-branch follow-up in native List of questions.docx, paras 18–19. ED20 L64, 69, 71–77; NF20.

Among the 14 experience-report-eligible datasets, structured routes occur in Sets 6, 13, 16, 17 and 20; report-based coding in Sets 3, 4, 7, 8, 10 and 11; and unresolved routes in Sets 2, 9 and 18. Sets 4, 7 and 10 broaden the negative category. Eleven eligible datasets have NE definitions without explicit uncertainty or perceived wakefulness, and eight also support conditional content. Fixed negative-answer follow-ups occur in Sets 6, 16 and 20, conditional clarification in Set 17, and three alternatives in Set 13. Sources are in Section B1.

SCAN codes participant-reported dream impressions as EWR and documents "Can you remember anything else?" after an initial report. DATA1 asks about an impression after a negative answer and excludes uncertain answers. Both preserve EWR, but their documented questions differ.

B5 Correspondence with the initial DREAM EEG analyses

Wong et al. (2025a), p. 3, lists six EEG datasets as Lacaux, Zhang, Scarpelli 1, Scarpelli 2, Siclari and Sikka. Table B5 matches them to version 9 datasets by name, authors and procedure. The published EEG analyses used version 1, whereas the audit uses version 9. A record-level version crosswalk is needed to assess how coding differences affect those analyses.

Supplementary Table B5 Reporting contrasts in the six initial EEG datasets

Dataset and article label	v9 set	Audit support	Meaning of deposited E versus NE	v9 sleep counts and EWR status	Evidence
Oudiette_N1Data Lacaux	2	Experience and conditional content	Contentful versus negative report; EWR retained separately but omitted from this contrast	23/8/6 8 EWR omitted	ED02 L56–59
Zhang & Wamsley 2019 Zhang	1	Neither intended boundary	Described content versus its absence; native NE records include explicit experience without content	156 content / 56 no-content EWR split lost	ED01 L79, L81–83; NF01
De Gennaro Multiple Awakenings Scarpelli 1	9	Experience reporting only	Experience judgment versus negative judgment; content was not collected	313 positive / 121 NE EWR split unmeasured 21 uncertain answers excluded	ED09 L62, L67
De Gennaro Young Adults Scarpelli 2	11	Experience reporting only	Feeling-of-dreaming diary contrast; EWR not retained separately	40 positive / 25 NE EWR split unmeasured	ED11 L68, L75
Tononi Serial Awakenings Siclari	13	Experience and conditional content	Contentful versus negative report; EWR retained separately but omitted from this contrast	124/56/107 56 EWR omitted	ED13 L76; NF13
REM Turku Sikka	6	Experience and conditional content	Contentful versus negative report; EWR retained separately but omitted from this contrast	123/8/3 8 EWR omitted	ED06 L137, L145, L147, L149

In datasets retaining all three categories, E versus NE excludes EWR and differs from E-or-EWR versus NE and from E versus EWR (main Figure 1). Added counts describe highest-amendment v9 REM/N1/N2/N3 records. Triplets are E/EWR/NE only where all three labels survive. Elsewhere positive or no-content labels are retained explicitly; unmeasured EWR is not zero. The 21 uncertain-answer Set 9 sleep records are outside its 434 classified records.

White dreams have already supported these separate contrasts within one laboratory. In experiment 1 of Siclari et al. (2017), a 20-person N2 subset with all three report categories showed lower parieto-occipital 1–4 Hz power before EWR than before NE; the E-versus-EWR low-frequency comparison was not significant. Contentful reports showed higher 20–50 Hz power than EWR over medial and lateral frontoparietal regions. Siclari et al. (2018) linked EWR versus NE to local slow-wave and fast-spindle differences. For E versus EWR, the differences were larger and steeper high-amplitude frontal slow waves, a higher type-I/type-II slow-wave ratio, and stronger central 18–30 Hz increases after type-I slow waves, predominantly on the right. Seven of its 14 participants overlapped with the 2017 experiment 2 sample. These are correlates of report categories, not of experience independent of report.

Figure 4 of Wong et al. (2025a), p. 5, pools 1,462 pre-report epochs and displays all three report categories. For human-scored NREM, its square mean markers show automated wake probability visually near 0.20 for E, 0.13–0.14 for EWR and 0.13 for NE. These are approximate readings of the displayed mean markers. The article reports a Bayesian ANOVA with category effects and a post hoc wake-like NREM association (p. 3); it does not report the proposed EWR-versus-NE and E-versus-

EWR contrasts within the subset of datasets selected by the present audit. This global hypnodensity measure describes different physiological features from local spectral power and microevents.

The reported NREM totals imply 70 EWR epochs: 778 total − 456 E − 252 NE. This subtraction uses the three-category pool and its reported totals. The corresponding REM subtraction gives nine EWR epochs (342 − 282 − 51). The E-versus-NE classifiers omit EWR. The figure motivates prospective or preregistered tests of the report distinctions, with arousal, retrieval and response criteria assessed before interpreting a neural correlate as experience itself.

Classifier coding. Oudiette N1 (Set 2), REM Turku (Set 6) and Siclari serial (Set 13) support E + EWR versus NE and E versus EWR. Multiple awakenings (Set 9) and Young adults (Set 11) support only the merged-positive-versus-NE contrast. Zhang–Wamsley's lost boundary (Set 1) does not support either intended contrast. For experience reporting, include EWR in the positive class; for conditional content, compare E with EWR and omit NE. The machine-readable report-contrast crosswalk in the archive's data directory extends these rules to all 20 datasets, retains indeterminate judgments and observed labels, and leaves unmeasured EWR counts unspecified. Applying it to the version 1 EEG recordings requires record-level case matching, channel checks and preprocessing.

B6 Native examples of experience without recalled content coded NE

Reports(1).csv and RecordsSet01.csv were joined by Case ID and checked against the highest-amendment DREAM export for Set 1. Both examples have native Experience = 0 and Last sleep stage = 2 (N2), with the same deposited NE classification. Their texts describe prior experience without recalled content. The original No Content category intentionally combined such absent-content answers; interpreting it as NE in DREAM loses the separate experience boundary. The original codes were retained and Set 1 remained excluded.

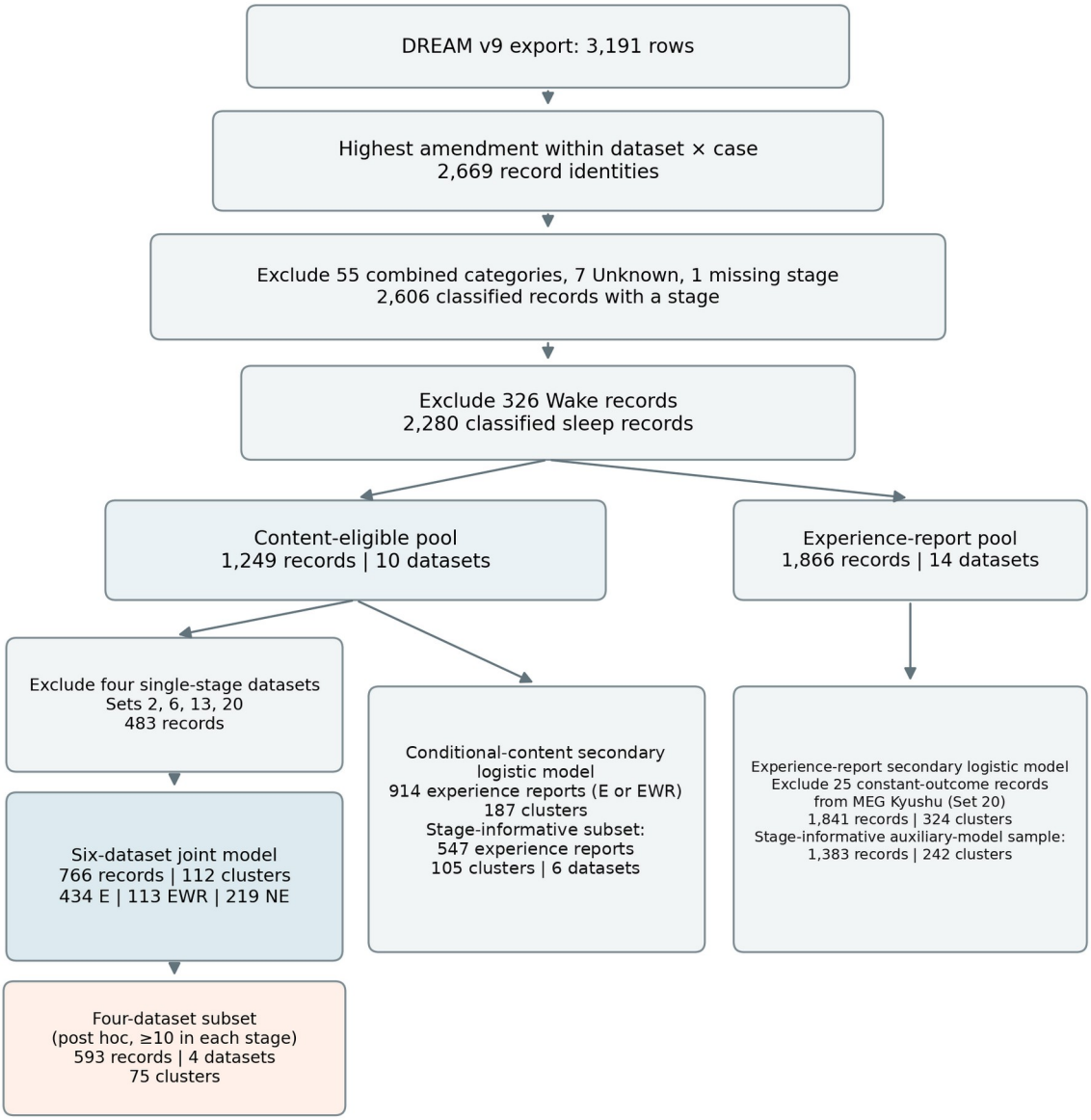
Supplementary Table B6 Native transcript examples in Zhang–Wamsley

Case ID	Code	Native report text
422_SO7	0 (NE)	Yeah I was definitely imagining something, but I can't remember, yeah I don't know, umm... I'm done.
512_SO8	0 (NE)	I was definitely dreaming but I can't remember it, that's all, sorry.

Source: NF01, Reports(1).csv, Text of Report, and RecordsSet01.csv, Experience and Last sleep stage; deposited coding procedure in ExperimentalDescription.set01.txt, L79 and L81–83. Original record IDs and text are retained for traceability. The deposited-record checks concern version 9; Section B5 explains the version 1 correspondence.

B7 Record selection and analysis samples

Supplementary Figure B1 Sample flow and analysis samples



Eligibility pools overlap. The joint sample retains all three categories; the conditional-content secondary logistic model retains E and EWR. Clusters are dataset–participant combinations.

Supplementary Material C Secondary model details

Supplementary Table C1 gives dataset-specific odds ratios from the joint model in the same REM-versus-NREM direction as Supplementary Figure D6c. These conditional model odds ratios have Bayesian CrIs. The secondary logistic contrasts involving REM use the same direction, with frequentist confidence intervals; Supplementary Table C3 also identifies a separate within-NREM stage difference.

Secondary logistic models with dataset intercepts and a common stage slope used the broader samples eligible for each category contrast. Participant-stratified conditional logistic regression examined the subset of clusters varying in both stage and outcome. These approaches address different questions: stratification accounts for stable participant differences, whereas cluster-robust standard errors address uncertainty under within-participant dependence (McNeish, 2026). Participant-cluster CR1 intervals used a normal reference. Conditional-score sandwich, model-Hessian and profile-likelihood intervals are specified in Tables C2–C5.

We also evaluated uncertainty in the secondary logistic models with one stage slope, treating datasets as clusters: ten informative datasets for experience reporting and six for conditional content. Dataset-cluster CR1 and delete-one-dataset CV3 intervals used a t reference with $G - 1$ degrees of freedom. A null-imposed, studentised linearised wild cluster bootstrap enumerated 1,024 Rademacher patterns for experience reporting and 46,656 Webb six-point patterns for content (MacKinnon et al., 2025; Webb, 2023). The statistic is the stage log odds ratio divided by its dataset-cluster CR1 standard error; bootstrap P values use the enumerated studentised distributions. These approximate procedures assume independent dataset clusters and test a single stage slope. They do not assess hierarchical credible-interval coverage.

Sensitivity analyses of the secondary logistic models examined participant-explicit eligibility, NREM substages and REM versus N2. Clock-time analyses compared unadjusted, linear and quadratic specifications on identical complete-case samples, preserving overnight continuity and ordinary daytime continuity across noon as appropriate to each protocol. Separate DATA1 models used recorded experimental night, awakening number and elapsed time from lights-off (Table C5).

The experience-report secondary logistic model used 1,841 records in 324 clusters, with ten datasets supplying stage information. Its REM-versus-NREM OR was 2.63 (95% participant-cluster CR1 normal CI 1.87 to 3.69). Dataset-cluster CR1 and CV3 t-reference CIs (9 df) were 1.58 to 4.38 and 1.46 to 4.73. The observed statistic was $t = 4.291$ ($G = 10$), with linearised bootstrap $P = 0.037$. Participant-stratified regression gave OR 3.54 (conditional-score sandwich CI 2.22 to 5.65) on 902 records in 58 informative clusters.

The conditional-content secondary logistic model compared E with EWR in 914 experience reports, of which 547 in six datasets supplied stage information. The REM-versus-NREM OR was 7.20 (95% participant-cluster CR1 normal CI 2.55 to 20.32). Dataset-cluster CR1 and CV3 t-reference intervals (5 df) were 0.95 to 54.61 and 0.56 to 93.15. The observed statistic was $t = 2.504$ ($G = 6$), with Webb six-point linearised bootstrap $P = 0.250$. Participant-stratified regression gave OR 7.88 (sandwich CI

2.55 to 24.38) on 261 records in 26 informative clusters. Tables C2–C5 specify the samples and interval methods.

The joint and secondary ORs describe different targets. The six-dataset joint hierarchy uses 766 E/EWR/NE records and participant random intercepts, with the hierarchical mean defined on the dataset log-OR scale. Binary common-slope models use the broader samples in Tables C2–C5 and no participant random intercept. Cluster corrections address uncertainty under the common-slope estimand. Conditional logistic models instead condition on participant clusters and retain only those varying in stage and outcome. SCAN and DATA1 have conditional-content medians of 11.01 and 12.04, but TWC USA, Sleep Talking and Aamodt evening have medians of 0.73, 0.33 and 0.63 (Table C1). Their inclusion is consistent with the smaller uncertain hierarchical mean of 1.83.

Supplementary Table C1 Dataset-specific stage odds ratios from the six-dataset primary fit

Dataset	Experience reporting OR	Conditional content OR
LODE	3.54 (1.53 to 7.95)	2.48 (0.77 to 8.95)
TWC USA	5.83 (1.66 to 28.81)	0.73 (0.02 to 8.29)
SCAN	6.66 (2.53 to 24.70)	11.01 (2.36 to 87.27)
Sleep Talking	3.96 (0.60 to 13.37)	0.33 (0.01 to 4.81)
DATA1	8.27 (3.55 to 28.15)	12.04 (3.53 to 62.52)
Aamodt evening	3.74 (0.30 to 16.09)	0.63 (0.05 to 6.51)

Posterior medians and 95% equal-tailed CrIs for REM-versus-NREM odds ratios, conditional on participant effects.

Tables C2–C5 give secondary logistic models under the documented eligibility rules. Both outcomes are coded positively: E or EWR versus NE for experience report, and E versus EWR for conditional content. Content models were fitted with EWR as the event. Complementing the event and reversing the stage contrast each invert the odds ratio, so together they leave its value unchanged. Experience reporting estimates are reciprocated when the stage contrast is reversed, with confidence limits reversed into ascending order. Two-sided tests are unchanged. All intervals in Tables C3–C5 are frequentist confidence intervals.

Supplementary Table C2 Samples for secondary logistic sensitivity analyses

Outcome	Restriction	Model records / events / clusters	Stage-informative datasets	Informative records / events
Experience reporting	Documented participant-explicit	1,716 / 1,224 / 283	LODE, TWC USA, Children dreaming, SCAN, Multiple awakenings, Sleep Talking, Young adults, DATA1, Aamodt evening	1,295 / 913

Outcome	Restriction	Model records / events / clusters	Stage-informative datasets	Informative records / events
Experience reporting	Remove DATA1	1,555 / 1,160 / 314	LODE, TWC USA, Children dreaming, SCAN, Multiple awakenings, Sleep Talking, Young adults, Aamodt evening, Aamodt morning	1,097 / 818
Experience reporting	Clock available unadjusted	1,555 / 1,160 / 314	LODE, TWC USA, Children dreaming, SCAN, Multiple awakenings, Sleep Talking, Young adults, Aamodt evening, Aamodt morning	1,097 / 818
Experience reporting	REM versus N2	1,505 / 1,117 / 311	LODE, TWC USA, Children dreaming, Multiple awakenings, Sleep Talking, Young adults, DATA1, Aamodt evening, Aamodt morning	1,015 / 743
Conditional content	Documented participant-explicit	883 / 703 / 163	LODE, TWC USA, SCAN, Sleep Talking, DATA1, Aamodt evening	547 / 434
Conditional content	Remove DATA1	744 / 619 / 177	LODE, TWC USA, SCAN, Sleep Talking, Aamodt evening	377 / 327
Conditional content	Clock available unadjusted	719 / 597 / 176	LODE, TWC USA, SCAN, Sleep Talking, Aamodt evening	377 / 327

Outcome	Restriction	Model records / events / clusters	Stage-informative datasets	Informative records / events
Conditional content	REM versus N2	665 / 535 / 156	LODE, TWC USA, Sleep Talking, DATA1, Aamodt evening	335 / 276

Events are E or EWR for experience report and E for conditional content. Clusters are dataset–participant combinations. Informative datasets vary in both stage and outcome. DATA1 exclusion and clock availability use the same 377 informative content records. Their total eligible samples differ.

Supplementary Table C3 Stage differences in secondary logistic models

Stage contrast	Experience reporting OR	Conditional content OR
REM versus N1	2.23 (0.95 to 5.21)	3.49 (0.74 to 16.46)
REM versus N2	2.29 (1.60 to 3.28)	7.08 (2.27 to 22.02)
REM versus N3	4.46 (2.74 to 7.28)	7.57 (2.78 to 20.62)
N3 versus N2 within NREM	0.51 (0.30 to 0.89)	0.93 (0.52 to 1.67)

ORs and 95% participant-cluster CR1 normal confidence intervals from binary models with dataset intercepts. REM is the numerator in contrasts involving REM. The final row is a separate N3-versus-N2 stage difference within NREM, using the coefficient difference and its covariance. No multiplicity correction was applied. The conditional-content interval spans substantial associations in both directions. The point estimate has the same direction as the derived E:EWR depth trend in Wong et al. (2025a), Supplementary pp. 11–13.

Supplementary Table C4 Interval methods for participant-stratified logistic models

Outcome	OR	Sandwich CI	Hessian CI	Profile CI	Records / clusters
Experience reporting	3.54	2.22 to 5.65	2.38 to 5.26	2.40 to 5.32	902/58
Conditional content	7.88	2.55 to 24.38	2.94 to 21.13	3.19 to 23.86	261/26

ORs compare REM with NREM. All intervals are 95% confidence intervals. The sandwich uses participant scores and a normal reference. The Hessian and ordinary profile methods use model-based uncertainty. Only clusters varying in both stage and outcome contribute.

Supplementary Table C5 DATA1 timing and sequence adjustments in secondary logistic models

Outcome	Records / events / clusters	Unadjusted OR	Night and order OR	Elapsed plus night and order OR
Experience reporting	286/170/10	8.06 (3.38 to 19.23)	10.11 (4.05 to 25.26)	10.49 (4.49 to 24.49)
Conditional content	170 / 107 / 10	32.77 (5.58 to 192.54)	35.85 (5.72 to 224.46)	34.68 (5.60 to 214.82)

REM-versus-NREM odds ratios and 95% participant-cluster CR1 normal confidence intervals on identical samples. Values are calculated from saved unrounded estimates and reported to two decimal places. Elapsed time is standardised from lights-off in native units. Its absolute unit is unspecified. The content sample contains one REM EWR record. Joint-model timing summaries appear in Supplementary Section D5.

Supplementary Table C6 DATA1 conditional white-dream proportions in N2 and N3

Stage	E/EWR/NE	Classified records	E + EWR	EWR / (E + EWR), % (95% bootstrap interval)
N2	24/26/25	75	50	52.0 (36.4 to 69.1)
N3	46/36/87	169	82	43.9 (22.2 to 63.7)

The N3-minus-N2 conditional white-dream difference was −8.1 pp (95% percentile bootstrap interval −25.0 to 5.1). The conditional denominators are 50 and 82 experience reports. Both stages contain classified records from all ten participants, with eight contributing N2 experience reports and nine N3 experience reports. Resampling drew ten participant clusters with replacement, retaining each selected participant's complete records, nights and stages together, for 20,000 paired-stage resamples (NumPy default_rng seed 2026101601). Intervals are the 2.5th and 97.5th percentiles of ratios of pooled counts and their paired difference. No resample had an empty denominator or was discarded.

Inputs were native DATA1 Records18.csv and Reports.csv linked by Case ID, checked against all 324 highest-amendment DATA1 cases in the version 9 export. Filenames, participant identifiers, category and stage fields agreed after documented label mappings. Combined classifications, Unknown and two blank native stages mapped to deposited Wake were excluded. The N2/N3 comparison retained 244 classified records and 132 experience reports. This post hoc comparison uses only ten participant clusters and is unadjusted for night or awakening order. No sleep-inertia measure was available. With so few clusters, percentile bootstrap intervals may be too narrow and have less than nominal coverage.

Supplementary Material D Joint model and sensitivity specifications

D1 Likelihood and hierarchy

For record i, let d(i) identify its dataset and j(i) its dataset–participant cluster. The stage indicator x_i is zero for REM and one for NREM. The three categories enter a categorical likelihood in the order E, EWR, NE:

$$Y_i \sim \ Categorical\left(a_i c_i, a_i\left(1\ -c_i\right),\ 1\ -a_i\right)$$

$$logit\left(a_i\right)\ =\alpha_{A,d(i)}+\beta_{A,d(i)}x_i+b_{A,j(i)}$$

$$logit\left(c_i\right)\ =\alpha_{C,d(i)}+\beta_{C,d(i)}x_i+b_{C,j(i)}$$

For contribution k in {A, C} and dataset d, the stage slopes are independent draws from the contribution-specific normal distribution, conditional on its mean and standard deviation. Dataset intercepts have independent priors. In the normal distributions below, the second argument is a variance. The Half-Normal arguments are standard-deviation scales.

$$\alpha_{k,d} \sim \ Normal(0, 2.5^2)$$

$$\beta_{k,d} \sim \ Normal(\mu_k, \tau_k^{\ 2})$$

$$\mu_k \sim \ Normal(0, 1.5^2), \tau_k \sim \ Half\text{-}Normal(\lambda)$$

The primary fit and all six leave-one-dataset-out refits use λ = 1. The narrow and wide prior sensitivities use λ = 0.5 and λ = 2, respectively. Participant effects have a common covariance matrix Σ across datasets, with independent effects for distinct retained cluster keys:

$$(b_{A,j}, b_{C,j})^T \sim \ Normal_2(0, \ \Sigma)$$

$$\Sigma \ = \ diag(\sigma_A, \sigma_C) \ \Omega \ diag(\sigma_A, \sigma_C)$$

$$\sigma_k \sim \ Half\text{-}Normal(1.5), \quad \Omega \ \sim \ LKJ_2(2)$$

Ω is a 2 × 2 correlation matrix. The implementation represents its correlation ρ as 2u − 1, with u ∼ Beta(2, 2), equivalent to a two-dimensional LKJ(2) prior. Records are independent conditional on the predictors and random effects in the six-dataset primary fit. Section D5 specifies the DATA1 timing terms; Section D6 specifies the additional stage, clock and history models and the predictive targets.

D2 Standardisation and aggregation

For each posterior draw, $p_{i,k}(s)$ is the predicted probability of report category k for record i after setting the stage indicator to s and retaining its fitted dataset and participant effects. Joint probabilities are averaged within dataset over all n_d retained records, including repeated records from a participant:

$$\pi_{d,k,s} = \frac{\sum_{i:d(i)=d} p_{i,k}(s)}{n_d}$$

The summation is over records with d(i) = d. Within each dataset and stage, the reporting probabilities are derived from these joint averages:

$$a_{d,s} = \pi_{d,E,s} + \pi_{d,EWR,s}$$

$$c_{d,s} = \frac{\pi_{d,E,s}}{\pi_{d,E,s} + \pi_{d,EWR,s}}$$

The contributions in Section 2.1 use $a_{d,R}$, $a_{d,N}$, $c_{d,R}$ and $c_{d,N}$ within each dataset. Record weights w_d are proportional to the dataset count n_d, divided by 766, 593 or 307 for the six-, four- or three-dataset fits, respectively. Equal weights are 1/6, 1/4 or 1/3. Identical weights apply at both stages, so the summed contributions reproduce the weighted gap in every draw. Leave-one-dataset-out summaries use the retained records of each refit (576, 734, 681, 744, 480 or 615) or equal weights of 1/5. Summaries using full-model predictions identify their subset and re-normalised weights separately.

Pooled probabilities are obtained by summing the joint category probabilities with the same weights, then forming the E/(E + EWR) ratio. The decomposition in this paper is always calculated within each dataset before aggregation. All transformations precede averaging over posterior draws.

The six-dataset equal-weight mean is the primary reported target, and the four-dataset equal-weight mean is the post hoc sensitivity analysis. Record, participant-cluster and experience-report overlap weights are secondary targets. This emphasis was chosen after the analyses had been run. Supplementary Table D1 describes the participant-cluster stage support.

Supplementary Table D1 Stage sample sizes within participant clusters

Dataset	Clusters	Both stages	REM only	NREM only
LODE	28	23	1	4
TWC USA	19	5	7	7
SCAN	18	16	2	0
Sleep Talking	12	5	0	7
DATA1	10	10	0	0
Aamodt evening	25	2	0	23
Total	112	61	10	41

Dataset–participant counts in the 766-record sample. Both stages means at least one retained record from each stage. Predictions for the 51 clusters observed in only one stage require model-based extrapolation.

D3 Analysis chronology and model checks

D3.1 Analysis chronology

The analysis chronology is summarised in Table D2, including journal-review and AI-simulated-feedback triggers. Dated changes to documentary judgments are recorded in Table B3. Main-text Section 2.6 describes the audit's outcome-awareness and the exploratory status of the analyses.

The boundary-sensitivity specification was saved on 18 October 2026, after earlier results were known and before the deterministic transformations. It fixed the two-stage grid and reporting paths.

The auxiliary comparison was extended on 20 October 2026 with participant-cluster diagnostics and on 21 October with three new joint hierarchical fits. The model specification was saved before these fits but after the earlier DREAM results, descriptive contrasts and external source-table values were known. External exploratory arithmetic was authorised after the source review and is reported with its remaining conditions in Section F.

Supplementary Table D2 lists the analysis families, their earliest documented dates and the recorded reason for each. A dated specification records the planned calculation. An analysis snapshot establishes that the calculation already existed by that date.

Supplementary Table D2 Recoverable analysis chronology

Analysis or decision	Earliest documented date	Documented motivation or trigger
Eligibility audit and revised inclusion	Formalised after frequency inspection; audit changes dated 1 and 9 Oct 2026 in Table B3	Recover contrasts from questions and coding. The initial audit began before analysis results were known; later native-file checks changed some judgments.
Common-slope logistic and participant-stratified models; CR1, CV3 and wild cluster checks	Present by the 14 Oct 2026 analysis snapshot	Binary and participant-stratified logistic models; joint models and dataset-cluster checks were prompted by journal peer review (Section D3.1).
Six-dataset joint hierarchy and record/equal summaries	Main model specification dated 9 Oct 2026	Model the three categories jointly, allow dataset-specific stage slopes and compare absolute contributions.
Slope-prior sensitivities and participant stage slopes	Present by 14 Oct 2026	Assess dependence on slope variation priors and participant heterogeneity.
Four-dataset subset	Present by 14 Oct 2026; explicitly post hoc	Reduce extreme stage extrapolation from Sleep Talking and Aamodt evening.
Three-dataset subset; DATA1 and SCAN exclusions; prediction-only subset summaries	Present by the 14 Oct 2026 analysis snapshot	Examine DATA1/SCAN influence and distinguish changing the averaging target from refitting.
REM–N2 restriction; secondary within-NREM contrasts	Present by the 14 Oct 2026 analysis snapshot	Examine depth while noting that the restriction also removes SCAN.
Clock, DATA1 night/order/elapsed and report-history adjustments; predictive checks	Present by the 14 Oct 2026 analysis snapshot	Assess recorded timing and dependence between consecutive reports from the same participant and night.
Participant-cluster and experience-report overlap weights	Present by the 14 Oct 2026 analysis snapshot	Compare averaging targets under uneven record and stage support.
Raw benchmark decomposition, bootstrap intervals and planning example	Present by the 14 Oct 2026 analysis snapshot	Compare scales and protocols; illustrate rare REM EWR sampling.
Common-slope anchors, paired share comparisons and selection bounds	Present by the 14 Oct 2026 analysis snapshot	Separate baseline dependence from stage associations; examine latent interpretations under explicit assumptions.
Same-scale E/EWR/NE comparison with Stucky	Present by 17 Oct 2026	Compare category redistribution rather than equating it with product contributions.
DATA1 N2/N3 raw white-dream check	Present by 17 Oct 2026; post hoc	Examine the within-protocol depth comparison.

Analysis or decision	Earliest documented date	Documented motivation or trigger
Complete single-dataset omissions, raw category figure, all-six raw/model contributions and category changes	17 Oct 2026	Prompted by AI-simulated feedback on a draft (see main-text Section 7): complete the single-dataset omissions and compare observed with model-based summaries.
Equal-dataset primary summary and four-dataset sensitivity analysis	Reporting emphasis chosen on 17 Oct 2026, after the analyses had been run	Prompted by AI-simulated feedback on a draft (see main-text Section 7): select one primary target and clarify the analysis hierarchy. The primary fit and eligible sample were retained.
NE-minus-EWR category-increase contrast	17 Oct 2026	Prompted by AI-simulated feedback on a draft (see main-text Section 7): quantify uncertainty in the difference between category increases using saved posterior draws; no new model fit.
NE/EWR boundary sensitivity and descriptive contrast comparison	18 Oct 2026 specification and calculation	Quantify dependence of earlier conclusions on specified category shifts; reuse saved draws without MCMC refitting.
Paired descriptive report-contrast uncertainty	20 Oct 2026	Preserve participant clusters and paired contrasts; inspect non-estimable sparse-stage bootstrap draws.
Conditional or native-label external source-table calculations	21 Oct 2026	Author-authorised descriptive analysis of available Stephan and Kumar tables after qualification gaps were identified; no independent-validation claim.
Ten-dataset mixed-resolution model and two sensitivity refits	21 Oct 2026	Author-requested common-target comparison after earlier numerical results were known; respect combined labels, sparse stage support and prior sensitivity.
Review-led checks (Section D12)	22 Oct 2026 specification and fits	Supplied AI-simulated first-round review; earlier results known. Reweighted Δ, mode summaries, documentary restriction, population integration, common-slope predictions and uncertain-answer refits.
Alternative decompositions, category shares, predictive prior sensitivity and uncertain-answer intervals	22 Oct 2026	Following further simulated-review feedback, existing-draw summaries of alternative decompositions, category shares and predictive prior sensitivity; participant-cluster BCa intervals for uncertain-answer rates. No new model fit.

Dates identify saved specifications and analysis snapshots. Where exact initiation dates or individual feedback exchanges were not retained, only the earliest documented snapshot and available trigger are reported. All analyses were exploratory, and the reporting hierarchy was selected after the analyses had been run.

D3.2 Computation and predictive checks

All analyses used Python 3.12.14. Models used PyMC 6.3.2 (Abril-Pla et al., 2023) and ArviZ 1.3.0 (Kumar et al., 2019), with four chains, 2,000 warm-up and 2,000 retained iterations per chain. The participant-slope fit used 4,000 of each. The primary fit and all six leave-one-dataset-out fits used the No-U-Turn sampler with target acceptance 0.99. Auxiliary fits and the fits in Sections D11–D12 used the same chain lengths, target acceptance 0.99 and maximum tree depth 12. The auxiliary fits recorded PyTensor 3.3.4, NumPy 2.3.5, pandas 2.2.3 and SciPy 1.17.0, with the numba NUTS backend. Diagnostics are grouped by model family in Table D3 (primary and sensitivity fits), Table D28 (auxiliary fits) and Table D38 (additional fits). All met R-hat < 1.01, bulk and tail ESS ≥400, zero divergences and minimum energy Bayesian fraction of missing information (E-BFMI) >0.30.

Supplementary Table D3 Sampling diagnostics for fitted models

Fit	Max R-hat	Min bulk ESS	Min tail ESS	Divergences	Min E-BFMI
Six-dataset primary fit	1.003	2292	2173	0	0.75
Slope prior scale 0.5	1.004	2871	3268	0	0.75
Slope prior scale 2	1.003	1575	2083	0	0.75
Exclude DATA1	1.003	1998	2542	0	0.75
Dataset clock terms	1.006	2024	2487	0	0.73
DATA1 elapsed time	1.004	2788	2470	0	0.78
DATA1 report history	1.002	2106	2062	0	0.77
DATA1 night and order	1.003	2235	1745	0	0.75
Exclude SCAN	1.003	2098	2747	0	0.75
Exclude Sleep Talking and Aamodt evening	1.003	2709	2754	0	0.76
N2 versus REM	1.004	2206	2495	0	0.74
Participant stage slopes	1.003	1376	2001	0	0.83
Exclude DATA1, Sleep Talking and Aamodt evening	1.002	1555	1582	0	0.76
Exclude LODE	1.003	2104	3036	0	0.77
Exclude TWC USA	1.003	1964	1998	0	0.72
Exclude Sleep Talking	1.003	2026	1814	0	0.72
Exclude Aamodt evening	1.003	2607	2121	0	0.74

ESS is posterior effective sample size and E-BFMI is the minimum across chains. Labels identify the fitted sample or covariate extension.

Posterior predictive checks examined dataset–stage counts, participant distributions and consecutive DATA1 reports. Deterministic scripts calculate the summaries, scale checks and prior–posterior

displays from saved draws. The analysis package supplies numerical inputs, figure code and checks of record and output identities.

Diagnostics cover dataset parameters, hyperparameters, participant effects and applicable covariates. Primary-fit prior prediction used 500 replications and dataset-by-stage category prediction 1,000. Participant and sequence checks used 2,000 replications (Supplementary Table D4). Additional fits used the same numerical criteria. Prior-predictive checks covered the primary and ten-dataset auxiliary models.

Supplementary Table D4 Participant and sequence posterior predictive checks

Statistic	Observed	95% prediction interval	Two-sided tail
E participants zero	7.0	5.0 to 17.0	0.266
E participants at least 1	105.0	95.0 to 107.0	0.266
E participants at least 3	60.0	55.0 to 66.0	0.883
E participant count variance	13.109	9.974 to 16.404	0.925
E participant rate variance	0.094	0.088 to 0.127	0.192
E maximum count	22.0	16.0 to 26.0	0.814
EWR participants zero	70.0	60.0 to 79.0	1.000
EWR participants at least 1	42.0	33.0 to 52.0	1.000
EWR participants at least 3	13.0	8.0 to 17.0	0.791
EWR participant count variance	6.009	3.303 to 8.987	0.806
EWR participant rate variance	0.022	0.018 to 0.055	0.239
EWR maximum count	18.0	10.0 to 23.0	0.634
NE participants zero	54.0	43.0 to 61.0	0.745
NE participants at least 1	58.0	51.0 to 69.0	0.745
NE participants at least 3	25.0	19.0 to 29.0	0.793
NE participant count variance	14.061	10.914 to 18.552	0.847
NE participant rate variance	0.089	0.071 to 0.123	0.537
NE maximum count	25.0	19.0 to 27.0	0.805
DATA1 consecutive same category	0.548	0.365 to 0.516	0.004
DATA1 EWR lag covariance	0.059	−0.005 to 0.054	0.025
History DATA1 consecutive same category	0.548	0.447 to 0.626	0.808
History DATA1 EWR lag covariance	0.059	0.017 to 0.097	0.755

Each check uses 2,000 posterior predictive replications. Observations outside the interval diagnose a mismatch for that statistic.

All 18 participant distribution statistics lay within their primary-fit predictive intervals. Across 219 adjacent DATA1 report pairs, the observed same-category frequency was 0.548, above the primary-fit

predictive interval of 0.365 to 0.516. EWR lag covariance was also higher than predicted. The history model's recursively generated same-category interval was 0.447 to 0.626 (Supplementary Table D4).

The participant prediction statistics examine zero events, event-threshold counts, variance of participant counts and rates, and the largest count for each category. Sequence checks use only the 219 adjacent retained DATA1 pairs. Primary-fit prediction generates categories conditionally independently. The history model propagates simulated preceding categories in native within-night order. The tail probabilities are descriptive posterior predictive summaries.

D4 Hierarchical stage summaries

New-protocol predictions integrate uncertainty in the hierarchical mean μ and slope SD τ. Absolute contributions summarise the included records. Table D42 shows how the predictive intervals depend on the slope prior.

Supplementary Table D5 Hierarchical means, dataset variation and new-dataset predictions

Datasets	Reporting outcome	Quantity	Log-OR (95% interval)	Slope SD τ (95% CrI)
Six	Experience report	Mean μ	1.52 (0.48 to 2.42)	0.74 (0.05 to 1.82)
Six	Experience report	New dataset	1.52 (−0.73 to 3.48)	—
Six	Conditional content	Mean μ	0.55 (−1.14 to 1.98)	1.65 (0.54 to 2.89)
Six	Conditional content	New dataset	0.55 (−3.55 to 4.25)	—
Four	Experience report	Mean μ	1.73 (0.65 to 2.68)	0.71 (0.06 to 1.79)
Four	Experience report	New dataset	1.73 (−0.41 to 3.71)	—
Four	Conditional content	Mean μ	1.56 (−0.02 to 2.84)	1.02 (0.06 to 2.35)
Four	Conditional content	New dataset	1.56 (−1.59 to 4.18)	—

Intervals for μ are credible intervals. New-dataset intervals are posterior predictive intervals, calculated by integrating the saved Normal(μ, τ^2) mixture as in Section D13.4. Slopes are reported as REM versus NREM.

D5 DATA1 timing measures and predictor extension

The native Records18.csv, Reports.csv and ExperimentalDescription18.txt identify DATA1, Noreika_DATA1. Reports.csv was identical to the version 1 contributor deposit (Noreika, 2023). Files were linked by case, filename and participant identifier. All 324 native cases and report-code mappings matched the highest-amendment DREAM records. Two cases had blank native stages and Wake labels in DREAM. The deposited Wake labels were retained in the source comparison without independent EEG rescoring. Both cases were excluded from the REM–NREM analysis. All 286 retained DATA1 records had the three timing measures: 107 E, 63 EWR and 116 NE in 10 participant clusters.

Experimental night had four levels, with night 1 as the reference. Awakening number was recorded in the reports file and was not inferred from filenames or sorted cases. The numeric Remarks field was documented as elapsed time from session lights-off to the individual awakening. Its absolute unit was

unspecified. The added linear predictor for contribution k was the main predictor plus DATA1-specific timing terms. Covariates were held at their recorded values under both stage predictions.

$$\eta_{k,i} = \alpha_{k,d(i)} + \beta_{k,d(i)} x_i + b_{k,j(i)} + \; I[d(i) \; = \; 16] z_i^{T} \gamma_k$$

For k in {A, C}, z contains indicators for nights 2, 3 and 4 and standardised awakening number. The second specification also contains standardised elapsed-lights-off values. The categorical likelihood and all other priors match the primary fit. For each timing covariate l, the coefficient prior is:

$$\gamma_{k,l} \sim \; Normal(0, 1.5^2)$$

Both contributions used the same scaling sample of 286 records. Awakening number had mean 4.68 and population standard deviation 2.50. Elapsed lights-off had mean 9,285.18 and population standard deviation 6,196.10 in native units. Their standardised correlation was 0.616. The first fit used night indicators and awakening number. The second added elapsed lights-off. Each timing coefficient had an independent, zero-centred Normal prior with standard deviation 1.5, separately for the two contributions. The timing slopes did not vary by participant or interact with stage.

Timing predictors were zero outside DATA1, although refitting the hierarchy could change shared parameters. Both stages occurred on every recorded night with overlapping awakening-order ranges (Supplementary Table D6). Some participant, stage and timing combinations still required extrapolation. The only REM EWR record occurred on night 4. Supplementary Table D7 gives the DATA1 contributions before and after timing adjustment.

Supplementary Table D6 DATA1 category and awakening-order support by night

Night	REM records (E/EWR/NE)	NREM records (E/EWR/NE)	REM order range	NREM order range
1	7 (7/0/0)	57 (16/14/27)	2–9	1–10
2	11 (10/0/1)	59 (17/21/21)	2–9	1–9
3	12 (11/0/1)	67 (21/14/32)	2–10	1–11
4	12 (9/1/2)	61 (16/13/32)	3–10	1–10

Counts and recorded awakening-order ranges for the 286 retained DATA1 records, linked to the native reports file.

Supplementary Table D7 DATA1 contributions before and after timing adjustment

Specification	Contentful-report gap	Experience-report contribution	Content contribution
Main	48.2 (35.1 to 60.0)	22.9 (15.7 to 30.4)	25.2 (15.6 to 33.6)
Night and order	49.4 (36.3 to 61.0)	24.2 (16.7 to 31.6)	25.2 (15.5 to 33.7)
Night, order and elapsed	49.2 (35.8 to 60.9)	24.0 (16.3 to 31.3)	25.2 (15.6 to 33.5)

Cells give posterior means and 95% equal-tailed credible intervals in pp for REM minus NREM, standardised to the same 286 DATA1 records in 10 participant clusters. All models are fitted to the common six-dataset sample. This table extracts the DATA1-specific summaries. Timing values remain fixed at their recorded values under each stage prediction.

D6 Additional specifications and predictive checks

The complete leave-one-dataset-out series omits LODE, TWC USA, SCAN, Sleep Talking, DATA1 and Aamodt evening separately. Each five-dataset sample was refitted with the same hierarchy and priors. Sparse-stage exclusion removes Sleep Talking and Aamodt evening together. The three-dataset sensitivity additionally removes DATA1. N2–REM restriction also removes SCAN, whose NREM records are N3. Participant slopes are independent of random intercepts and each other, with zero-centred normal distributions and Half-Normal(0.5) scales. Tables D8–D9 specify the samples.

Supplementary Table D8 Joint decompositions across analysis choices

Model	Records / clusters summarised	Weight	Contentful-report gap (pp)	Experience-report contribution (pp)	Content contribution (pp)
Without DATA1, full-model predictions	480 / 102	Record	14.1 (−1.2 to 26.0)	12.9 (6.2 to 18.6)	1.2 (−13.1 to 11.0)
Without DATA1, full-model predictions	480 / 102	Equal	16.1 (2.6 to 27.9)	15.3 (7.1 to 22.4)	0.8 (−10.1 to 9.6)
Exclude DATA1 and refit	480 / 102	Record	11.4 (−3.7 to 23.4)	11.5 (4.9 to 17.3)	−0.1 (−13.9 to 9.8)
Exclude DATA1 and refit	480 / 102	Equal	13.0 (−0.5 to 24.7)	13.3 (5.2 to 21.0)	−0.3 (−11.1 to 8.3)
LODE, TWC USA, SCAN, full-model predictions	307 / 65	Record	24.5 (15.6 to 33.0)	17.1 (9.7 to 24.3)	7.4 (2.1 to 13.0)
LODE, TWC USA, SCAN, full-model predictions	307 / 65	Equal	25.0 (15.5 to 34.9)	18.2 (10.4 to 26.5)	6.9 (1.4 to 13.3)
Exclude Sleep Talking and Aamodt evening	593 / 75	Record	36.9 (28.8 to 44.4)	20.6 (15.0 to 25.9)	16.3 (10.8 to 21.4)
Exclude Sleep Talking and Aamodt evening	593 / 75	Equal	32.9 (24.7 to 41.3)	20.2 (13.7 to 27.1)	12.7 (7.6 to 18.6)
Exclude SCAN	681 / 94	Record	24.0 (12.8 to 33.8)	15.8 (10.2 to 20.7)	8.2 (−1.5 to 16.0)
Exclude SCAN	681 / 94	Equal	16.7 (3.5 to 28.1)	15.2 (7.0 to 22.5)	1.5 (−8.8 to 9.8)
N2 versus REM	442 / 90	Record	18.8 (6.9 to 29.4)	12.3 (5.8 to 18.3)	6.5 (−3.7 to 15.1)
N2 versus REM	442 / 90	Equal	15.5 (2.1 to 27.8)	12.8 (4.9 to 20.4)	2.7 (−8.1 to 12.2)
Dataset clock terms	766 / 112	Record	24.6 (14.5 to 33.0)	14.6 (9.1 to 19.8)	10.0 (1.7 to 16.2)
Dataset clock terms	766 / 112	Equal	18.1 (7.1 to 28.1)	13.7 (6.1 to 20.6)	4.4 (−4.1 to 11.2)

Model	Records / clusters summarised	Weight	Contentful-report gap (pp)	Experience-report contribution (pp)	Content contribution (pp)
DATA1 night and order	766 / 112	Record	27.3 (16.9 to 36.1)	17.1 (11.9 to 21.7)	10.2 (1.1 to 17.2)
DATA1 night and order	766 / 112	Equal	21.7 (10.2 to 31.9)	16.8 (9.6 to 23.3)	4.9 (−4.3 to 12.2)
DATA1 elapsed time	766 / 112	Record	27.4 (16.9 to 36.1)	17.1 (12.0 to 21.6)	10.2 (1.0 to 17.4)
DATA1 elapsed time	766 / 112	Equal	21.8 (10.4 to 31.9)	16.9 (9.8 to 23.3)	4.9 (−4.2 to 12.3)
DATA1 report history	766 / 112	Record	27.1 (17.1 to 35.9)	17.1 (12.1 to 21.6)	10.0 (1.1 to 17.0)
DATA1 report history	766 / 112	Equal	21.7 (10.4 to 31.8)	16.9 (9.7 to 23.4)	4.8 (−4.3 to 12.1)
Participant stage slopes	766 / 112	Record	27.1 (16.7 to 35.9)	16.8 (11.8 to 21.4)	10.3 (1.2 to 17.5)
Participant stage slopes	766 / 112	Equal	21.6 (10.2 to 31.9)	16.7 (9.7 to 23.0)	4.9 (−4.2 to 12.5)
Exclude DATA1, Sleep Talking and Aamodt evening	307 / 65	Record	24.2 (14.8 to 33.1)	16.7 (8.9 to 24.4)	7.5 (2.0 to 13.3)
Exclude DATA1, Sleep Talking and Aamodt evening	307 / 65	Equal	24.9 (14.7 to 35.3)	17.5 (9.2 to 26.6)	7.4 (1.7 to 14.4)
LODE, TWC USA predictions, three-dataset fit	222 / 47	Record	21.4 (10.5 to 31.6)	16.3 (6.9 to 25.0)	5.1 (−1.0 to 11.1)
LODE, TWC USA predictions, three-dataset fit	222 / 47	Equal	21.5 (9.4 to 34.0)	17.4 (7.4 to 28.6)	4.1 (−2.5 to 12.2)
Slope prior scale 0.5	766 / 112	Record	28.4 (19.1 to 36.3)	16.8 (11.8 to 21.4)	11.6 (3.9 to 17.5)
Slope prior scale 0.5	766 / 112	Equal	24.7 (14.6 to 33.5)	17.1 (10.7 to 23.1)	7.6 (−0.3 to 13.8)
Slope prior scale 2	766 / 112	Record	26.0 (15.3 to 35.3)	16.6 (11.2 to 21.3)	9.4 (0.1 to 17.1)
Slope prior scale 2	766 / 112	Equal	19.6 (8.1 to 30.3)	16.3 (9.0 to 23.0)	3.3 (−5.7 to 11.1)

Posterior means and 95% equal-tailed CrIs in pp. Records and clusters describe the observations being summarised. Full-model prediction rows retain all 766 records in 112 clusters during estimation and average over only the stated subset. The LODE and TWC USA prediction rows instead reuse the 307-record, 65-cluster

fit to LODE, TWC USA and SCAN, summarising 222 records in 47 clusters. Restricted refits use their retained records for both estimation and summarisation. Covariate and participant-slope fits retain the full sample.

Supplementary Table D9 Complete leave-one-dataset-out refits

Excluded dataset	Records / clusters	Weights	Experience-report contribution (pp)	Content contribution (pp)	Total gap (pp)
LODE	576 / 84	Equal	18.1 (9.6 to 25.7)	4.1 (−6.8 to 13.2)	22.2 (8.2 to 34.7)
LODE	576 / 84	Record	17.8 (12.0 to 23.0)	11.5 (−0.4 to 20.9)	29.4 (15.8 to 40.6)
TWC USA	734 / 93	Equal	15.3 (7.8 to 21.6)	6.9 (−3.5 to 14.6)	22.2 (9.9 to 32.6)
TWC USA	734 / 93	Record	16.2 (10.7 to 21.0)	11.4 (2.2 to 18.3)	27.5 (17.3 to 36.4)
SCAN	681 / 94	Equal	15.2 (7.0 to 22.5)	1.5 (−8.8 to 9.8)	16.7 (3.5 to 28.1)
SCAN	681 / 94	Record	15.8 (10.2 to 20.7)	8.2 (−1.5 to 16.0)	24.0 (12.8 to 33.8)
Sleep Talking	744 / 100	Equal	16.7 (10.8 to 22.6)	9.0 (−0.8 to 16.1)	25.7 (14.5 to 34.8)
Sleep Talking	744 / 100	Record	17.0 (11.8 to 21.7)	12.2 (2.7 to 18.9)	29.2 (18.4 to 37.5)
DATA1	480 / 102	Equal	13.3 (5.2 to 21.0)	−0.3 (−11.1 to 8.3)	13.0 (−0.5 to 24.7)
DATA1	480 / 102	Record	11.5 (4.9 to 17.3)	−0.1 (−13.9 to 9.8)	11.4 (−3.7 to 23.4)
Aamodt evening	615 / 87	Equal	19.8 (12.3 to 27.2)	8.5 (2.5 to 14.2)	28.3 (18.5 to 37.5)
Aamodt evening	615 / 87	Record	20.2 (14.6 to 25.5)	15.3 (10.0 to 20.2)	35.5 (27.2 to 42.9)

Posterior means and 95% CrIs in pp. Each five-dataset sample was refitted with the same hierarchy and priors. Equal weights are primary and record weights secondary, both re-normalised over the retained datasets. These analyses assess dataset influence. Sampling diagnostics are in Table D3.

Clock-time terms are dataset-specific, linear and standardised within LODE, TWC USA, SCAN, Sleep Talking and Aamodt evening. Values below noon are shifted by 24 hours only in SCAN and Aamodt evening, preserving overnight continuity. DATA1 terms are zero in this clock model. Supplementary Table D10 describes stage overlap. In the history model, DATA1 has indicators for the immediately preceding EWR, NE or unavailable report, with E as the reference, indicators for nights 2–4 and standardised awakening order. Numeric order uses mean 4.68 and population standard deviation 2.50. Covariate coefficients in each logit have independent Normal(0, 1.5^2) priors. Standardisation retains the recorded covariates under both stage settings.

Supplementary Table D10 Stage sample sizes for dataset clock terms

Dataset	Stage	Records	Clock range (hours)	Within other stage range
LODE	REM	63	6.00 to 11.14	63
LODE	NREM	127	4.49 to 13.37	118

Dataset	Stage	Records	Clock range (hours)	Within other stage range
TWC USA	REM	20	9.32 to 15.68	18
TWC USA	NREM	12	9.33 to 13.77	12
SCAN	REM	46	25.56 to 33.31	26
SCAN	NREM	39	23.91 to 30.77	19
Sleep Talking	REM	5	5.65 to 6.39	5
Sleep Talking	NREM	17	5.63 to 7.03	9
Aamodt evening	REM	2	23.52 to 23.67	2
Aamodt evening	NREM	149	20.63 to 24.00	8

Clock values use the continuity rule in Supplementary Section D6. Overlap counts summarise dataset-level clock ranges. Elapsed time since sleep onset and within-participant overlap require separate information.

Native files linked all 286 retained DATA1 records to experimental night, awakening order and elapsed time from lights-off. The elapsed measure had an unspecified absolute unit and did not identify sleep onset. The history model used 219 adjacent retained report pairs within participant and night, without bridging missing or excluded records. Covariates remained at their recorded values under both stage predictions.

The record- and equal-weight summaries differed mainly because of two datasets: changing the weights of DATA1 and Aamodt evening accounts for 4.9 of the 5.3 pp difference in the content contribution mean. Stage-overlap weights gave 12.9 pp. About 1.8 pp of their 2.8 pp increase over record weights came from the other four datasets (Supplementary Table D11).

Supplementary Table D11 Fixed weights for the six-dataset summary

Dataset	Record	Equal dataset	Participant-cluster weights	Experience reporting stage overlap
LODE	0.248	0.167	0.250	0.356
TWC USA	0.042	0.167	0.170	0.065
SCAN	0.111	0.167	0.161	0.189
Sleep Talking	0.029	0.167	0.107	0.026
DATA1	0.373	0.167	0.089	0.341
Aamodt evening	0.197	0.167	0.223	0.023

Weights sum to one before rounding. Participant-cluster weights are proportional to cluster counts, and stage-overlap weights to $n_R n_N / (n_R + n_N)$, using counts of E or EWR by stage. Contributions are calculated within datasets before weighting, which changes the relative influence of datasets while preserving record contributions within them. Stage-overlap weights describe observed stage support.

Across the examined choices, dataset inclusion changed the content estimate more than the covariate extensions (Figure D3). Excluding Sleep Talking and Aamodt evening gave 16.3 pp (95% CrI 10.8 to 21.4) with record weights and 12.7 pp (7.6 to 18.6) with equal weights. Excluding DATA1 alone gave −0.1 pp (−13.9 to 9.8) and −0.3 pp (−11.1 to 8.3).

Excluding all three left LODE, TWC USA and SCAN (307 records, 65 participant clusters), where the content contribution was positive: 7.5 pp (2.0 to 13.3) with record weights and 7.4 pp (1.7 to 14.4) with equal weights. Experience-report contributions were 16.7 pp (8.9 to 24.4) and 17.5 pp (9.2 to 26.6). The refit shows that the positive contribution did not require pooling with DATA1. Full-model summaries over the same three datasets were similar and slightly smaller: 7.4 pp (2.1 to 13.0) and 6.9 pp (1.4 to 13.3), respectively (Supplementary Table D8).

SCAN supplied about half of the record-weighted mean (3.9 of 7.5 pp; 51%) and nearly two-thirds of the equal-weighted mean (4.7 of 7.4 pp; 63%). LODE's own contribution was 5.5 pp (−0.9 to 11.7), with an interval including zero. Summarising only LODE and TWC USA from the same fit gave 5.1 pp (−1.0 to 11.1) and 4.1 pp (−2.5 to 12.2), respectively. SCAN was omitted from this summary but retained during estimation. Both summary intervals included zero, so the positive pooled mean outside DATA1 rested largely on SCAN (Supplementary Table D8).

Timing covariates, preceding report categories, participant stage slopes and heterogeneity priors changed the estimates less (Supplementary Figure D3). The N2–REM restriction also removed SCAN, whose NREM records were all N3. Its content contributions (6.5 and 2.7 pp) resembled those from removing SCAN alone (8.2 and 1.5 pp). Because datasets and substages both changed, this reduction cannot be attributed to depth alone. Table D8 gives the full intervals.

Supplementary Table D12 Standardised report probabilities by dataset and stage

Dataset or weights	Stage	E (%)	EWR (%) Observed / model	NE (%)	Experience reporting (%)	Conditional content (%)
LODE	REM	75.3 (65.9 to 83.3)	Observed 3/63 = 4.8% Model 5.7 (1.8 to 11.8)	19.0 (11.8 to 27.7)	81.0 (72.3 to 88.2)	93.0 (85.7 to 97.7)
LODE	NREM	54.4 (46.7 to 61.9)	Observed 11/127 = 8.7% Model 8.6 (4.7 to 13.8)	36.9 (29.7 to 44.2)	63.1 (55.8 to 70.3)	86.3 (78.5 to 92.5)
TWC USA	REM	85.2 (71.4 to 94.6)	Observed 1/20 = 5.0% Model 5.2 (0.5 to 15.5)	9.6 (2.5 to 20.7)	90.4 (79.3 to 97.5)	94.2 (82.9 to 99.4)

Dataset or weights	Stage	E (%)	EWR (%) Observed / model	NE (%)	Experience reporting (%)	Conditional content (%)
TWC USA	NREM	65.3 (43.4 to 84.7)	Observed 0/12 = 0.0% Model 3.8 (0.0 to 16.3)	30.9 (12.4 to 52.2)	69.1 (47.8 to 87.6)	94.4 (76.9 to 99.9)
SCAN	REM	88.9 (80.3 to 95.3)	Observed 0/46 = 0.0% Model 3.1 (0.4 to 8.5)	8.0 (2.7 to 15.3)	92.0 (84.7 to 97.3)	96.6 (90.7 to 99.6)
SCAN	NREM	54.5 (40.3 to 68.0)	Observed 7/39 = 17.9% Model 16.3 (7.0 to 28.6)	29.2 (18.4 to 41.5)	70.8 (58.5 to 81.6)	77.0 (60.1 to 89.9)
Sleep Talking	REM	56.6 (25.6 to 80.9)	Observed 1/5 = 20.0% Model 13.3 (0.9 to 41.2)	30.2 (11.5 to 57.4)	69.8 (42.6 to 88.5)	81.0 (42.8 to 98.6)
Sleep Talking	NREM	44.5 (25.9 to 63.2)	Observed 0/17 = 0.0% Model 4.0 (0.1 to 14.9)	51.5 (32.6 to 70.5)	48.5 (29.5 to 67.4)	91.8 (70.6 to 99.8)
DATA1	REM	78.5 (66.4 to 88.8)	Observed 1/42 = 2.4% Model 7.5 (1.5 to 17.3)	13.9 (6.3 to 22.8)	86.1 (77.2 to 93.7)	91.2 (80.1 to 98.2)
DATA1	NREM	30.4 (25.0 to 35.9)	Observed 62/244 = 25.4% Model 24.4 (19.6 to 29.7)	45.2 (39.8 to 50.8)	54.8 (49.2 to 60.2)	55.4 (47.2 to 63.4)
Aamodt evening	REM	65.1 (21.8 to 92.6)	Observed 1/2 = 50.0% Model 29.4 (3.7 to 73.8)	5.5 (0.9 to 22.4)	94.5 (77.6 to 99.1)	68.9 (23.0 to 96.0)

Dataset or weights	Stage	E (%)	EWR (%) Observed / model	NE (%)	Experience reporting (%)	Conditional content (%)
Aamodt evening	NREM	71.8 (64.9 to 78.3)	Observed 26/149 = 17.4% Model 17.8 (12.2 to 23.9)	10.4 (6.4 to 15.0)	89.6 (85.0 to 93.6)	80.1 (73.4 to 86.3)
Record	REM	75.9 (66.2 to 83.6)	11.0 (4.6 to 20.2)	13.2 (9.0 to 18.3)	86.8 (81.7 to 91.0)	87.4 (76.8 to 94.7)
Record	NREM	49.0 (45.4 to 52.7)	16.9 (14.2 to 19.8)	34.1 (30.8 to 37.4)	65.9 (62.6 to 69.2)	74.4 (70.2 to 78.3)
Equal	REM	74.9 (65.1 to 83.0)	10.7 (4.5 to 19.8)	14.4 (9.4 to 20.7)	85.6 (79.3 to 90.6)	87.5 (76.9 to 94.8)
Equal	NREM	53.5 (47.7 to 59.1)	12.5 (9.8 to 16.0)	34.0 (28.7 to 39.5)	66.0 (60.5 to 71.3)	81.0 (75.9 to 85.1)

Model entries are posterior means and 95% equal-tailed CrIs, standardised over the same retained records within each dataset at both stages. The EWR column also gives the observed count, denominator and percentage for that stage, without standardisation. Conditional content is E/(E + EWR), calculated within each draw. The final record- and equal-weight rows retain the six-dataset standardised target.

The raw pooled REM EWR proportion uses each dataset's share of REM records. The standardised record-weight summary instead uses its share of all retained records at both stages. In the six-dataset sample, Aamodt evening contributes only 2/178 observed REM records but 151/766 of the fixed summary weight. Its raw REM EWR proportion is 1/2. Applying the fixed weights to raw within-dataset proportions gives 12.7%, compared with the raw pooled 3.9% and the model-standardised 11.0% (Table D13). Aamodt evening contributes 53% of the model-standardised value.

The four-dataset discrepancy persists after that cohort is removed: observed pooled and fixed-weight raw REM EWR are both 2.9%, while the standardised model mean is 6.3%. Hierarchical smoothing, nonzero predictions for observed zero cells, participant effects and averaging each participant's predictions over both observed and unobserved stages all contribute to differences from raw stage proportions. DATA1 illustrates the participant-mix issue: most of its records are NREM, but their participant effects also enter its predicted REM probability, 7.5% versus 1/42 (2.4%) observed in the six-dataset fit. These mechanisms jointly contribute to the model predictions.

Supplementary Table D13 Observed pooling and standardisation of EWR

Sample and stage	Raw pooled EWR (%)	Raw EWR with fixed record weights (%)	Model standardised EWR (%)
Six datasets REM	3.9	12.7	11.0 (4.6 to 20.2)

Sample and stage	Raw pooled EWR (%)	Raw EWR with fixed record weights (%)	Model standardised EWR (%)
Six datasets NREM	18.0	17.1	16.9 (14.2 to 19.8)
Four datasets REM	2.9	2.9	6.3 (2.6 to 11.4)
Four datasets NREM	19.0	17.6	17.3 (14.2 to 20.7)

The raw pooled column weights datasets by their stage-specific counts. The next column applies the same total-record dataset weights at both stages to the raw within-dataset proportions. The final column uses the joint model, averaging predictions for the same retained participant/record mixture at both stages. Intervals are 95% CrIs. These diagnostic record-weight comparisons explain why standardised and observed REM EWR proportions differ. The primary manuscript estimates use equal dataset weights.

Supplementary Table D14 Signed experience-report shares of the gap

Dataset or target	Median share (%)	95% CrI (%)	Draws with positive gap (%)
LODE	77.0	46.4 to 113.3	100.0
TWC USA	102.2	34.7 to 279.9	97.9
SCAN	53.9	28.4 to 80.9	100.0
Sleep Talking	105.5	−602.8 to 788.3	79.4
DATA1	47.5	36.5 to 60.2	100.0
Aamodt evening	−0.1	−390.2 to 425.9	42.7
Record average	61.2	46.3 to 93.8	100.0
Equal dataset average	76.1	53.4 to 136.4	100.0

Shares are D_A / Δ_E within each posterior draw. All draws are retained. Opposing contributions can produce shares outside 0–100%, and near-zero gaps make the ratio unstable. Sparse datasets should therefore be read alongside their absolute contributions.

With the primary equal weights, NREM increases in NE and EWR were 19.6 and 1.8 pp across six datasets, and 24.0 and 8.9 pp across four. In the four-dataset equal-weight mean, NE accounts for 73% of the decrease in E. With record weights, the NE/EWR increases are 20.9/5.9 pp and 25.9/11.0 pp, or NE shares of 78% and 70%. These percentages are ratios of posterior mean category changes. Supplementary Table D16 compares the NE and EWR increases using matched posterior draws.

Supplementary Table D15 Standardised category probabilities and stage differences

Fit and weights	Report	REM (%)	NREM (%)	REM − NREM (pp)
Six datasets; record	E	75.9 (66.2 to 83.6)	49.0 (45.4 to 52.7)	26.8 (16.4 to 35.7)

Fit and weights	Report	REM (%)	NREM (%)	REM − NREM (pp)
Six datasets; record	EWR	11.0 (4.6 to 20.2)	16.9 (14.2 to 19.8)	−5.9 (−13.1 to 3.5)
Six datasets; record	NE	13.2 (9.0 to 18.3)	34.1 (30.8 to 37.4)	−20.9 (−26.5 to −14.9)
Six datasets; equal	E	74.9 (65.1 to 83.0)	53.5 (47.7 to 59.1)	21.4 (10.2 to 31.6)
Six datasets; equal	EWR	10.7 (4.5 to 19.8)	12.5 (9.8 to 16.0)	−1.8 (−9.2 to 7.8)
Six datasets; equal	NE	14.4 (9.4 to 20.7)	34.0 (28.7 to 39.5)	−19.6 (−26.8 to −11.5)
Four datasets; record	E	79.8 (73.1 to 85.6)	42.9 (38.6 to 47.4)	36.9 (28.8 to 44.4)
Four datasets; record	EWR	6.3 (2.6 to 11.4)	17.3 (14.2 to 20.7)	−11.0 (−16.1 to −5.1)
Four datasets; record	NE	13.9 (9.4 to 19.1)	39.8 (35.6 to 44.1)	−25.9 (−32.4 to −19.1)
Four datasets; equal	E	82.8 (77.2 to 87.6)	49.8 (42.9 to 56.6)	32.9 (24.7 to 41.3)
Four datasets; equal	EWR	5.2 (2.4 to 8.8)	14.1 (10.4 to 18.9)	−8.9 (−14.4 to −3.9)
Four datasets; equal	NE	12.1 (8.2 to 16.7)	36.1 (29.8 to 42.9)	−24.0 (−31.7 to −16.6)

Posterior means and 95% CrIs. Probabilities are percentages, and REM − NREM differences are in pp. The table includes both record and equal weights in the six- and four-dataset fits. Category changes use the same matched posterior draws as Table D16.

Supplementary Table D16 Difference between NE and EWR increases after NREM awakening

Fit and weights	NE increase (pp)	EWR increase (pp)	NE increase minus EWR increase (pp)
Six datasets; equal	19.6 (11.5 to 26.8)	1.8 (−7.8 to 9.2)	17.8 (6.0 to 30.2)
Four datasets; equal	24.0 (16.6 to 31.7)	8.9 (3.9 to 14.4)	15.1 (5.1 to 25.2)

Posterior means and 95% equal-tailed CrIs. Increases are NREM minus REM. Within each of the 8,000 saved posterior draws, the NE-minus-EWR difference was computed within dataset and then averaged with equal dataset weights. This paired calculation preserves the posterior dependence between the category changes.

D7 Information about dataset slope variation

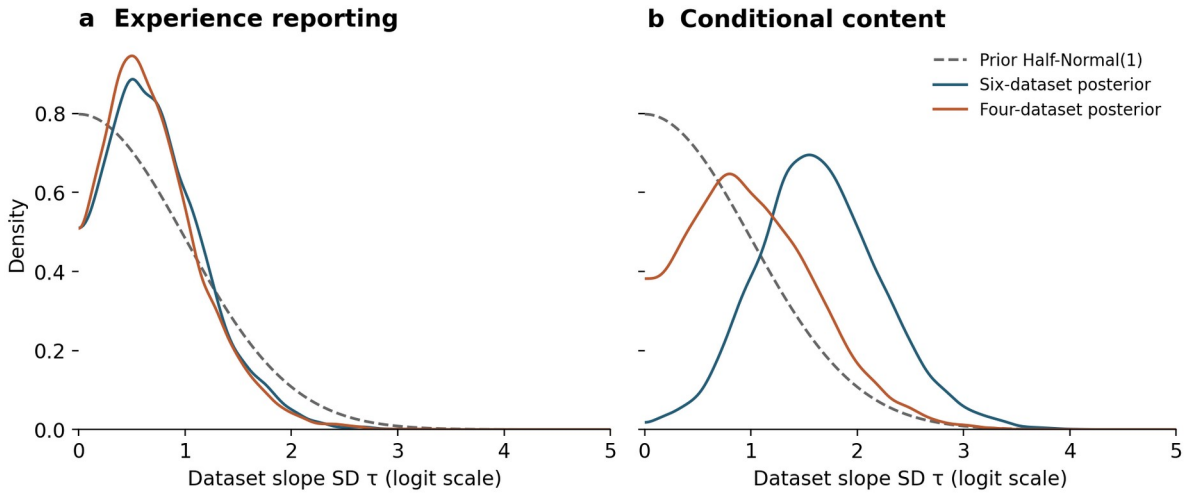
Supplementary Figure D1 compares the analytical Half-Normal(1) prior with posterior draws from the six-dataset model and four-dataset subset. Each posterior has 8,000 draws. Density estimates use a Gaussian kernel reflected at zero. The plotted range is 0–5. Supplementary Table D17 supplies numerical summaries.

Supplementary Table D17 Prior and posterior summaries of slope variation

Distribution	Mean	Median	95% interval
Prior, both contributions	0.80	0.67	0.03 to 2.24
Six-dataset experience reporting	0.74	0.68	0.05 to 1.82
Four-dataset experience reporting	0.71	0.64	0.06 to 1.79
Six-dataset conditional content	1.65	1.62	0.54 to 2.89
Four-dataset conditional content	1.02	0.95	0.06 to 2.35

The six-dataset experience-slope SD posterior closely resembles its Half-Normal(1) prior, providing little information about cross-protocol variation. The content-slope posterior favours more variation than that prior, with an imprecise magnitude. Prior scales 0.5 and 2 yield posterior mean SDs of 0.86 and 2.48 for conditional content, and 0.46 and 0.96 for experience reporting. Section D12.4 compares common and varying content slopes with the same conditional count-prediction check.

Supplementary Figure D1 Prior and posterior distributions of between-dataset slope variation



a, Experience reporting. b, Conditional content. Dashed grey: Half-Normal(1) prior. Blue: six-dataset posterior. Orange: four-dataset posterior. Both parameters are dataset stage-slope standard deviations on the logit scale. Their calculation is implemented in the accompanying figure script.

D8 Uncertainty and baseline dependence

All calculations retain the fitted draws. Within each draw, gap shares are the experience-report contribution divided by the contentful-report gap. Pairwise differences use matched draws. Medians are used because near-zero denominators can produce long tails. Supplementary Table D18 gives paired differences in gap shares; Supplementary Table D19 compares observed and model-based contributions. The analysis archive (file pairwise_differences.csv) also gives differences in conditional stage log-odds ratios and residual share differences.

Supplementary Table D18 Posterior differences in experience-report gap shares

Fit	Dataset difference	Median difference (95% CrI)	P(difference > 0)
Six	LODE − TWC USA	−25.1 (−208.0 to 57.6)	0.189
Six	LODE − SCAN	23.1 (−18.2 to 67.1)	0.878
Six	LODE − Sleep Talking	−28.9 (−711.4 to 698.3)	0.264
Six	LODE − DATA1	29.5 (−3.4 to 67.1)	0.963
Six	LODE − Aamodt evening	71.3 (−343.5 to 470.0)	0.846
Six	TWC USA − SCAN	48.3 (−30.1 to 226.7)	0.936
Six	TWC USA − Sleep Talking	−6.6 (−856.7 to 902.3)	0.441
Six	TWC USA − DATA1	54.5 (−16.3 to 233.7)	0.958
Six	TWC USA − Aamodt evening	98.8 (−433.1 to 677.7)	0.867
Six	SCAN − Sleep Talking	−52.6 (−742.6 to 661.3)	0.199
Six	SCAN − DATA1	6.4 (−20.7 to 34.1)	0.670
Six	SCAN − Aamodt evening	48.1 (−378.8 to 441.5)	0.774
Six	Sleep Talking − DATA1	57.9 (−646.9 to 743.9)	0.816
Six	Sleep Talking − Aamodt evening	93.9 (−1061.9 to 1227.3)	0.775
Six	DATA1 − Aamodt evening	46.4 (−379.6 to 436.8)	0.750
Four	LODE − TWC USA	−12.7 (−87.6 to 40.5)	0.281
Four	LODE − SCAN	16.1 (−20.3 to 55.1)	0.813
Four	LODE − DATA1	23.4 (−8.1 to 54.7)	0.940
Four	TWC USA − SCAN	28.7 (−25.5 to 105.5)	0.868
Four	TWC USA − DATA1	36.5 (−15.1 to 111.8)	0.926

Fit	Dataset difference	Median difference (95% CrI)	P(difference > 0)
Four	SCAN − DATA1	7.0 (−19.2 to 32.1)	0.701

Differences are percentage points of the signed experience-report share. They are descriptive posterior comparisons, without multiplicity-adjusted decision thresholds. Unstable ratios in sparsely supported datasets should be read with their absolute contributions and the gap-sign probabilities in Supplementary Table D14.

Observed experience-report shares are D_A / Δ_E. Among datasets with at least 20 records per stage, estimates and 95% bootstrap intervals were LODE 68.7% (18.9 to 146.7), SCAN 53.4% (23.6 to 80.4), DATA1 46.0% (30.8 to 66.7) and Siclari experiment 2 34.5% (28.0 to 41.7). The Siclari experiment 2 counts are REM E/EWR/NE = 131/31/8 and NREM = 193/225/168 (Siclari et al., 2017, Supplementary Table 2). Mean conditional content across stages is [(131/162) + (193/418)]/2 = 0.635, rounded to 0.64 in the main text.

Supplementary Table D19 Observed and model-based contributions in all six datasets

Dataset and stage counts	Estimate	Experience-report contribution (pp)	Content contribution (pp)	Total gap (pp)
LODE REM/NREM 63/127	Observed	11.2 (1.8 to 21.2)	5.1 (−4.0 to 14.2)	16.3 (4.9 to 28.1)
LODE REM/NREM 63/127	Model	16.1 (5.9 to 25.1)	4.8 (−1.9 to 11.4)	20.9 (8.9 to 31.7)
TWC USA REM/NREM 20/12	Observed	27.6 (6.8 to 54.3)	−4.3 (−15.0 to 0.0)	23.3 (−2.2 to 51.6)
TWC USA REM/NREM 20/12	Model	20.0 (4.5 to 39.5)	−0.2 (−10.1 to 12.8)	19.8 (0.5 to 41.6)
SCAN REM/NREM 46/39	Observed	25.1 (9.2 to 41.9)	21.9 (8.6 to 35.4)	46.9 (31.5 to 61.2)
SCAN REM/NREM 46/39	Model	18.4 (8.7 to 30.0)	16.0 (5.1 to 30.1)	34.3 (20.3 to 49.9)
Sleep Talking REM/NREM 5/17	Observed	5.9 (−32.9 to 52.6)	−18.8 (−60.8 to 0.0)	−12.9 (−70.0 to 47.1)
Sleep Talking REM/NREM 5/17	Model	18.4 (−7.7 to 37.6)	−6.3 (−30.1 to 7.8)	12.1 (−20.5 to 37.7)
DATA1 REM/NREM 42/244	Observed	27.4 (20.1 to 36.1)	32.1 (17.5 to 46.6)	59.4 (50.8 to 68.3)
DATA1 REM/NREM 42/244	Model	22.9 (15.7 to 30.4)	25.2 (15.6 to 33.6)	48.2 (35.1 to 60.0)
Aamodt evening REM/NREM 2/149	Observed	6.6 (1.6 to 14.4)	−29.1 (−82.9 to 24.2)	−22.5 (−80.0 to 35.3)
Aamodt evening REM/NREM 2/149	Model	3.6 (−8.7 to 8.8)	−10.3 (−53.1 to 15.4)	−6.7 (−50.3 to 21.7)

Dataset and stage counts	Estimate	Experience-report contribution (pp)	Content contribution (pp)	Total gap (pp)
Siclari experiment 2 REM/NREM: 170/586	Observed	15.2 (12.0 to 18.4)	28.9 (22.4 to 35.3)	44.1 (36.8 to 51.4)

Observed rows give point estimates and 95% participant-cluster percentile bootstrap intervals. Model rows give posterior means and 95% CrIs from the six-dataset primary fit. DREAM stage counts are in main Table 2 and Figure 2. The LODE, SCAN and DATA1 contribution intervals use 20,000 resamples from a generator initialised with seed 2026101312. The aggregate Siclari benchmark follows these datasets in the same generator sequence and uses 20,000 stage-specific multinomial resamples without participant clustering. TWC USA, Sleep Talking and Aamodt evening use 20,000 paired-stage participant-cluster resamples from seed 2026101701, with the generator traversing all six datasets in Set ID order and drawing complete participant records with replacement. A contribution is estimable only when both stages contain at least one experience report. Valid contribution resamples were 19,998 for TWC USA, 19,366 for Sleep Talking and 17,569 for Aamodt evening. Missing-denominator draws are counted rather than filled with pseudocounts. Gap intervals use every resample with both stages represented. The raw contribution and category-change resampling procedures in Tables D19–D20 differ when a stage has no experience reports. Sparse-cell percentile intervals can be unstable and should not be read as calibrated evidence against a zero cell.

Category changes in Table D20 use seed 2026101701 for all six datasets. A change requires both stages to be represented, but not a positive experience denominator. Valid resamples numbered 20,000 each for LODE, TWC USA, SCAN and DATA1, 19,969 for Sleep Talking (31 lacked REM) and 17,569 for Aamodt evening (2,431 lacked REM). Sleep Talking had a further 603 resamples with a zero experience denominator for its contribution calculation; TWC USA had two.

The two-proportion planning example uses a conventional normal approximation with two-sided $\alpha = .05$ and 80% power: 274.25 is rounded upward to 275 independent experience reports per stage. The illustrative 6% and 13% are not an observed stage pair: 6% rounds the REM upper value of 6.1%, and 13% is a scenario near the NREM lower value of 12.0%, among cells with ≥20 records (Table B2). The calculation omits clustering, unequal allocation and dataset heterogeneity. The external experiment 2 REM proportion, 31/162 (19.1%), shows why the 6% scenario is not universal.

Benchmark protocol. Experiment 2 of Siclari et al. (2017) provides REM and NREM counts under an explicit three-category interview. Seven trained participants completed 51 serial-awakening nights, five to ten each, after two weeks of home questionnaire practice. The original experiments 1 and 2 used different participants (Methods, “Study participants in experiment 2”). Siclari serial reuses the experiment 1 protocol. Overlap with experiment 2 would require the expanded deposit to include that experiment’s participants, which its case assignments do not establish. The benchmark is a descriptive comparison across these protocols.

Scale also changes descriptive ordering. From raw counts, the conditional-content/experience-report log-OR ratio is 1.38 for LODE, 1.67 for DATA1 and 0.76 for the Siclari benchmark. Their absolute content shares are 31%, 54% and 66%. The benchmark has the smallest slope ratio and largest absolute content share among these three. SCAN has no REM EWR record, making its uncorrected content log-OR infinite. These point calculations illustrate scale dependence rather than test dataset differences. The calculations are available in the analysis archive (file scale_comparison.csv).

Supplementary Table D20 Category changes in every modelled dataset

Dataset	Estimate	NE increase NREM minus REM (pp)	EWR increase NREM minus REM (pp)
LODE	Observed	12.4 (1.7 to 24.1)	3.9 (−5.9 to 13.1)
LODE	Model	17.9 (6.5 to 28.0)	2.9 (−4.2 to 9.4)
TWC USA	Observed	28.3 (7.0 to 55.6)	−5.0 (−17.6 to 0.0)
TWC USA	Model	21.2 (4.8 to 41.8)	−1.4 (−12.0 to 10.1)
SCAN	Observed	29.0 (10.8 to 47.1)	17.9 (6.5 to 31.6)
SCAN	Model	21.2 (10.0 to 34.2)	13.2 (3.3 to 25.9)
Sleep Talking	Observed	7.1 (−52.1 to 57.9)	−20.0 (−66.7 to 0.0)
Sleep Talking	Model	21.4 (−8.8 to 42.0)	−9.3 (−37.3 to 5.1)
DATA1	Observed	36.4 (28.6 to 44.1)	23.0 (10.6 to 36.7)
DATA1	Model	31.3 (21.7 to 40.8)	16.9 (5.9 to 25.2)
Aamodt evening	Observed	10.1 (2.7 to 18.6)	−32.6 (−88.0 to 23.1)
Aamodt evening	Model	4.9 (−11.8 to 11.2)	−11.6 (−56.2 to 14.6)
Siclari experiment 2	Observed	24.0 (19.0 to 28.7)	20.2 (13.1 to 27.0)

Positive values mean more NE or EWR after NREM awakening. Observed rows use participant-cluster percentile intervals. Model rows use the six-dataset posterior and 95% CrIs. The NE and EWR increases sum to the fall in E. Table D19 and the text above specify bootstrap denominators, seeds and sparse-cell limitations.

The transformations replace each dataset-specific stage slope b with its hierarchical mean μ, for each reporting logit and draw. Let the fitted predictor be $\eta(s) = \alpha + b{\cdot}s + u$, where REM is s = 0 and NREM is s = 1. For anchor t, the transformed predictor is $\eta^*(s) = \alpha + b{\cdot}t + u + \mu{\cdot}(s - t)$. Here μ is the model's NREM-versus-REM mean stage coefficient. The reported REM-versus-NREM mean associations have the opposite sign. We use t = 0 to preserve REM, t = 1 to preserve NREM, and t = 0.5 to preserve the midpoint of the fitted logits. The midpoint is on the logit scale, not the probability scale. Participant effects and standardisation records are unchanged. Each table row compares fitted and transformed predictions from the same six- or four-dataset posterior. Fitted reconstruction matches saved probabilities to a maximum absolute error of 1.4×10^{-15}. The code checks that REM- and NREM-anchored transformations preserve the respective stage probabilities. The transformations illustrate how fitted baselines affect the decomposition. Supplementary Table D21 and Figure D2 show the paired anchoring comparisons; Supplementary Table D22 gives the four-dataset selection lower endpoints and their weighted contributions.

Supplementary Table D21 Fitted shares and common slopes under three anchors

Fit	Dataset	Fitted share (95% CrI)	REM fixed (95% CrI)	NREM fixed (95% CrI)	Logit midpoint fixed (95% CrI)
Six	LODE	77.0 (46.4 to 113.3)	89.7 (42.0 to 127.2)	84.5 (42.3 to 255.3)	87.1 (50.0 to 168.0)

Fit	Dataset	Fitted share (95% CrI)	REM fixed (95% CrI)	NREM fixed (95% CrI)	Logit midpoint fixed (95% CrI)
Six	TWC USA	102.2 (34.7 to 279.9)	88.0 (25.4 to 142.0)	96.1 (51.0 to 129.0)	92.8 (44.3 to 131.9)
Six	SCAN	53.9 (28.4 to 80.9)	90.8 (34.1 to 124.5)	69.4 (−173.7 to 304.2)	81.8 (32.1 to 199.1)
Six	Sleep Talking	105.5 (−602.8 to 788.3)	85.8 (27.2 to 182.2)	96.0 (60.8 to 128.4)	91.6 (49.1 to 143.6)
Six	DATA1	47.5 (36.5 to 60.2)	83.3 (25.8 to 144.5)	62.1 (−209.4 to 369.8)	72.7 (22.1 to 250.2)
Six	Aamodt evening	−0.1 (−390.2 to 425.9)	31.2 (−261.6 to 340.9)	41.3 (−250.2 to 328.3)	35.3 (−253.4 to 336.7)
Four	LODE	72.3 (43.5 to 101.1)	74.7 (37.2 to 101.8)	74.1 (52.7 to 103.3)	74.3 (46.1 to 102.9)
Four	TWC USA	86.4 (33.7 to 157.1)	70.3 (22.4 to 100.2)	81.8 (37.2 to 100.1)	76.6 (32.6 to 100.2)
Four	SCAN	55.9 (31.6 to 80.2)	67.7 (23.7 to 100.1)	59.2 (31.4 to 97.8)	63.6 (30.3 to 100.1)
Four	DATA1	48.8 (38.2 to 60.8)	55.5 (19.0 to 99.4)	51.0 (32.3 to 92.7)	53.0 (26.5 to 98.2)

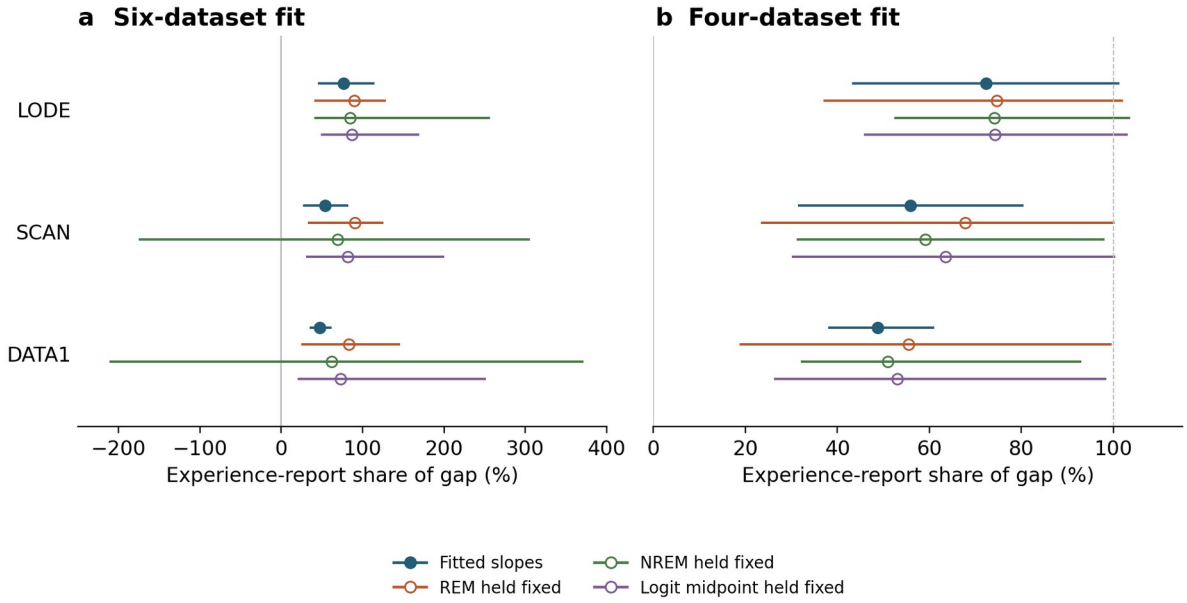
Posterior medians and 95% CrIs in percent. The fitted share and all three transformations in a row use the same posterior. The first transformation fixes fitted REM logits, the second fixes NREM logits, and the third fixes the midpoint of the two logits. Paired fitted-minus-transformed residuals and their uncertainty are supplied for every anchor in the analysis archive (file common_slope_anchors.csv).

Supplementary Table D22 Selection lower endpoints in the four-dataset fit

Dataset or target	Record weight	Lower endpoint (95% CrI), pp	Weighted contribution to lower endpoint, pp
LODE	32.0%	−8.3 (−20.1 to 2.8)	−2.7
TWC USA	5.4%	−3.2 (−19.9 to 17.1)	−0.2
SCAN	14.3%	11.8 (−2.1 to 28.5)	1.7
DATA1	48.2%	22.5 (8.7 to 35.7)	10.8
LODE, TWC USA, SCAN Same fit, reweighted	Renormalised	−2.2 (−10.8 to 6.6)	—

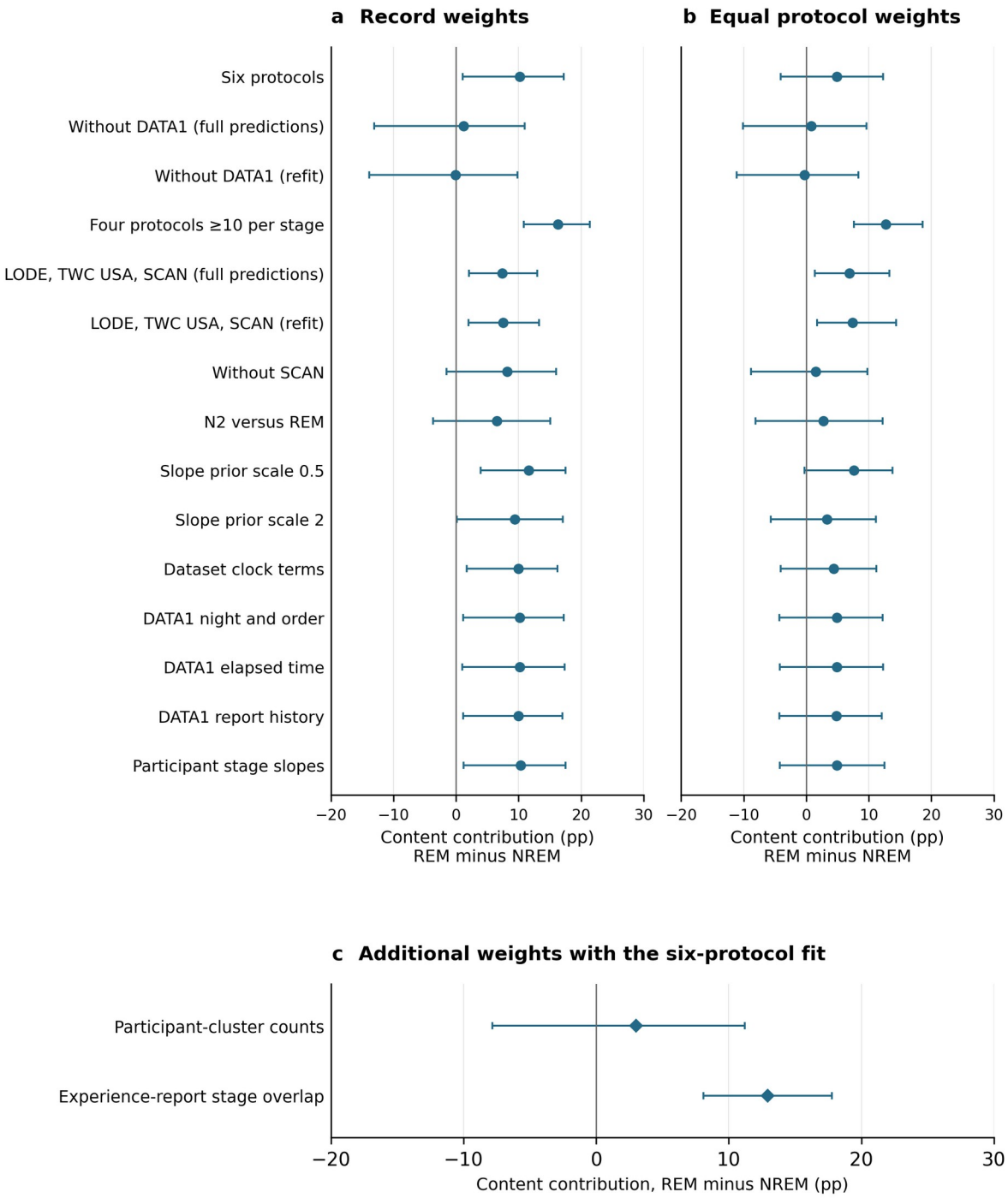
Posterior means and 95% CrIs from the four-dataset posterior. DATA1 contributes 10.8 pp to the 9.7 pp record-weighted lower endpoint, and the other datasets jointly contribute −1.1 pp. The final row omits DATA1 from averaging and renormalises the remaining record weights, while retaining its contribution to model estimation.

Supplementary Figure D2 Fitted shares and three anchoring choices



a, Six-dataset fit. b, Four-dataset fit. Points are medians and lines 95% CrIs. The displayed datasets have ≥20 records per stage. Within each panel, every estimate uses that panel's posterior. Blue shows fitted dataset slopes. Orange, green and purple replace slopes with the hierarchical mean while holding REM, NREM or the logit midpoint fixed. Panel ranges differ to display the wider six-dataset uncertainty. All intervals are shown in Table D21.

Supplementary Figure D3 Secondary specifications and weighting targets



Panels a–b show the content contribution in pp (REM minus NREM) under record and equal weights for sample restrictions, prediction-only summaries and model extensions. Panel c shows the same contribution under participant-cluster and experience-report overlap weights from the six-dataset fit. All are secondary summaries. Numerical values and sample definitions are in Tables D8 and D11. Complete single-dataset exclusions are in Table D9 and main Figures 3–4.

D9 Sensitivity to the boundary between negative reports and white dreams

This post hoc analysis (Table D2) used the primary fit (766 records, 112 participant clusters) and the four-dataset subset (593 records, 75 clusters). Each supplied 8,000 posterior draws standardised over the same retained records at both stages (Section D2). Boundary scenarios transform these fitted probabilities without refitting.

Let e, w and n denote E, EWR and NE probabilities, with R and N indicating REM and NREM. For a specified fraction r at stage s, the transformation was

$$n'_s = (1 - r_s)n_s; w'_s = w_s + r_s n_s; e'_s = e_s$$

G is the NREM-minus-REM increase in NE minus the corresponding increase in EWR. Therefore

$$G' = G - 2(r_N n_N - r_R n_R)$$

Each draw was transformed within each dataset before computing experience reporting a = 1 − n, conditional content c = e/a, and the symmetric contributions defined in the main text. Fixed dataset weights were applied after decomposition. For both fits we retained equal weights, weights proportional to each dataset's total retained records, and each single-dataset target: 14 targets, 441 scenarios each, and 6,174 result rows. Parameters ranged independently from 0 to 1 in steps of 0.05. The two fixed paths moved only NREM NE, or the same fraction of NE at both stages. Fractions specify the share of NE probability transferred within each stage.

Supplementary Table D23 Selected boundary scenarios in the six dataset equal weight summary

Scenario	rN / rR	G, pp (95% CrI)	DA, pp (95% CrI)	DA − DC, pp (95% CrI)
Stored categories	0.00 / 0.00	17.8 (6.0 to 30.2)	16.6 (9.6 to 22.9)	11.7 (1.6 to 22.1)
NREM only	0.10 / 0.00	11.0 (−0.3 to 23.0)	13.3 (6.8 to 18.9)	5.1 (−4.4 to 15.2)
NREM only	0.25 / 0.00	0.8 (−10.0 to 12.3)	8.7 (2.9 to 13.6)	−4.0 (−12.8 to 6.0)
Same fraction	0.10 / 0.10	13.9 (3.3 to 25.5)	14.3 (8.3 to 19.7)	7.2 (−1.7 to 17.0)
Same fraction	0.25 / 0.25	8.0 (−0.9 to 18.5)	11.3 (6.6 to 15.4)	1.2 (−6.6 to 10.3)
Same fraction	0.50 / 0.50	−1.8 (−9.2 to 7.8)	7.0 (4.1 to 9.5)	−7.5 (−15.0 to 1.6)

Entries are posterior means and 95% equal-tailed CrIs. DA > 0 concerns a positive experience-report contribution; DA − DC > 0 concerns its being the larger contribution; G > 0 concerns a greater NE than EWR increase. These are distinct statements. The E gap is unchanged in every scenario. Full results retain all datasets, fits, weights and grid points.

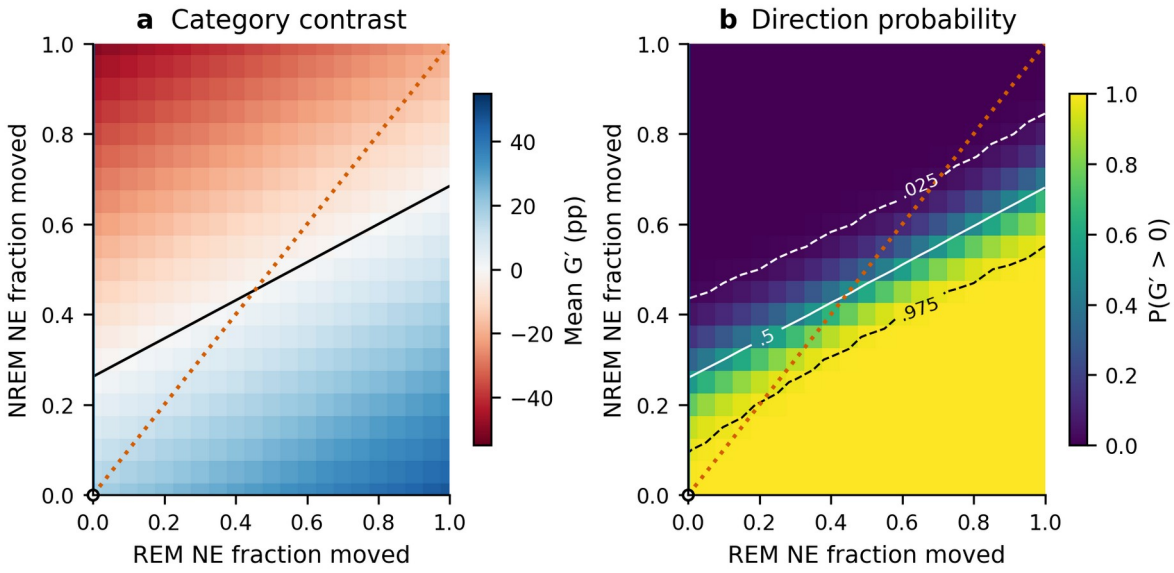
Supplementary Table D24 Moved fractions at which the posterior mean of G reaches zero

Fit and fixed weights	NREM only	Same fraction at both stages
Six datasets, equal	26.2%	45.4%
Six datasets, record	22.1%	35.9%
Four datasets, equal	20.9%	31.5%
Four datasets, record	18.7%	28.7%

These roots use unrounded posterior means. They are not thresholds for interval inclusion of zero. We separately retained every draw's algebraic root and status: already nonpositive at the origin, crossing within 0–1, decreasing without a crossing by 1, nondecreasing, or undefined denominator. In the six-dataset equal-weight summary, 17/8,000 draws were already nonpositive. On the same-fraction path one further draw did not cross within 0–1. Root quantiles conditional on crossing are labelled as conditional and do not represent an unconditional posterior interval.

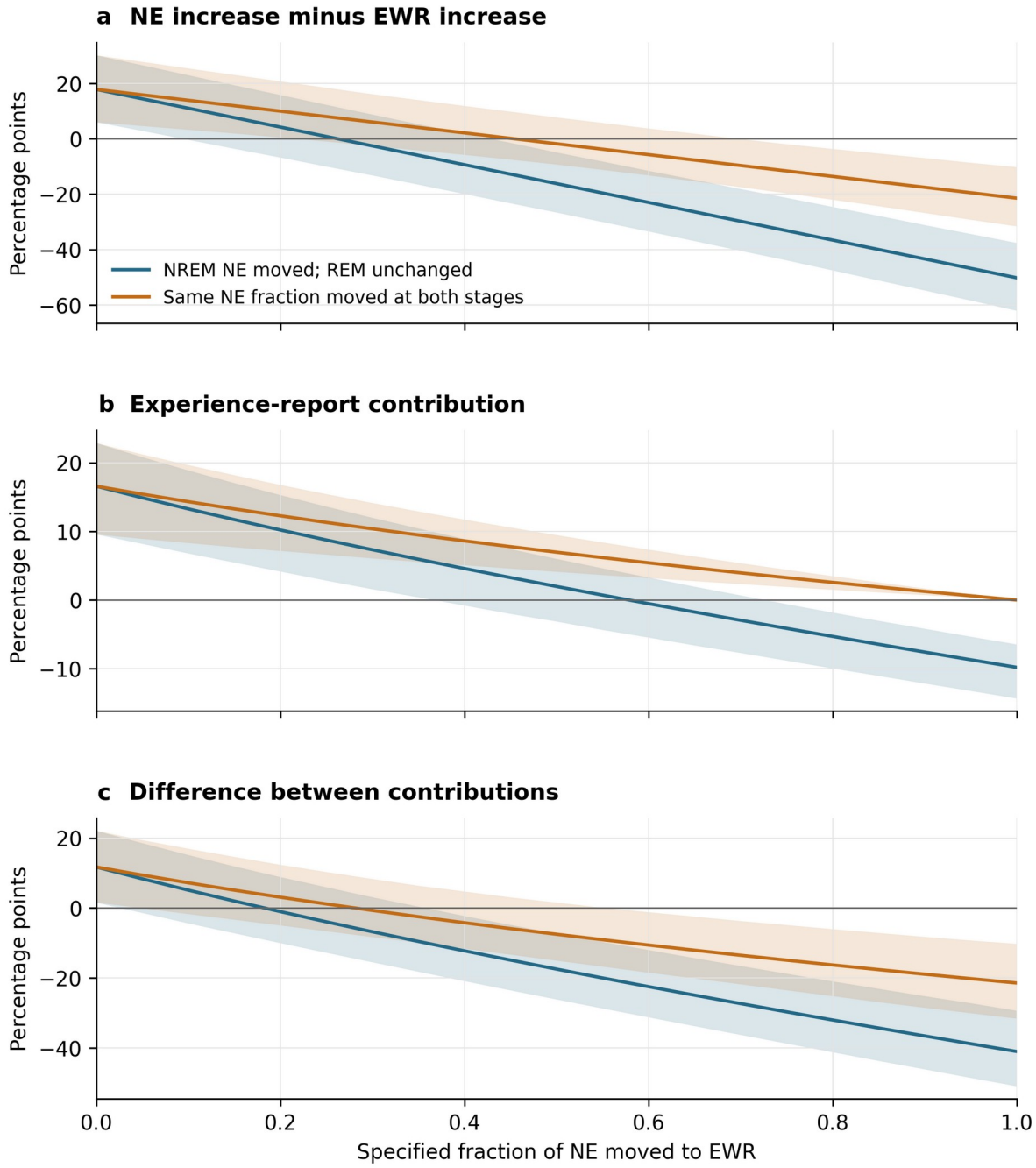
At 0.05 grid resolution, G's CrI first included zero at 0.10 on the NREM-only path and 0.25 on the same-fraction path. Its mean first became negative at 0.30 and 0.50. These grid locations describe different summaries. Recalculation checked probability sums, the G identity, additive decomposition and the invariant E gap, reproducing all 6,174 rows within 5×10^{-13}.

Supplementary Figure D4 Boundary grid for the six dataset equal weight summary



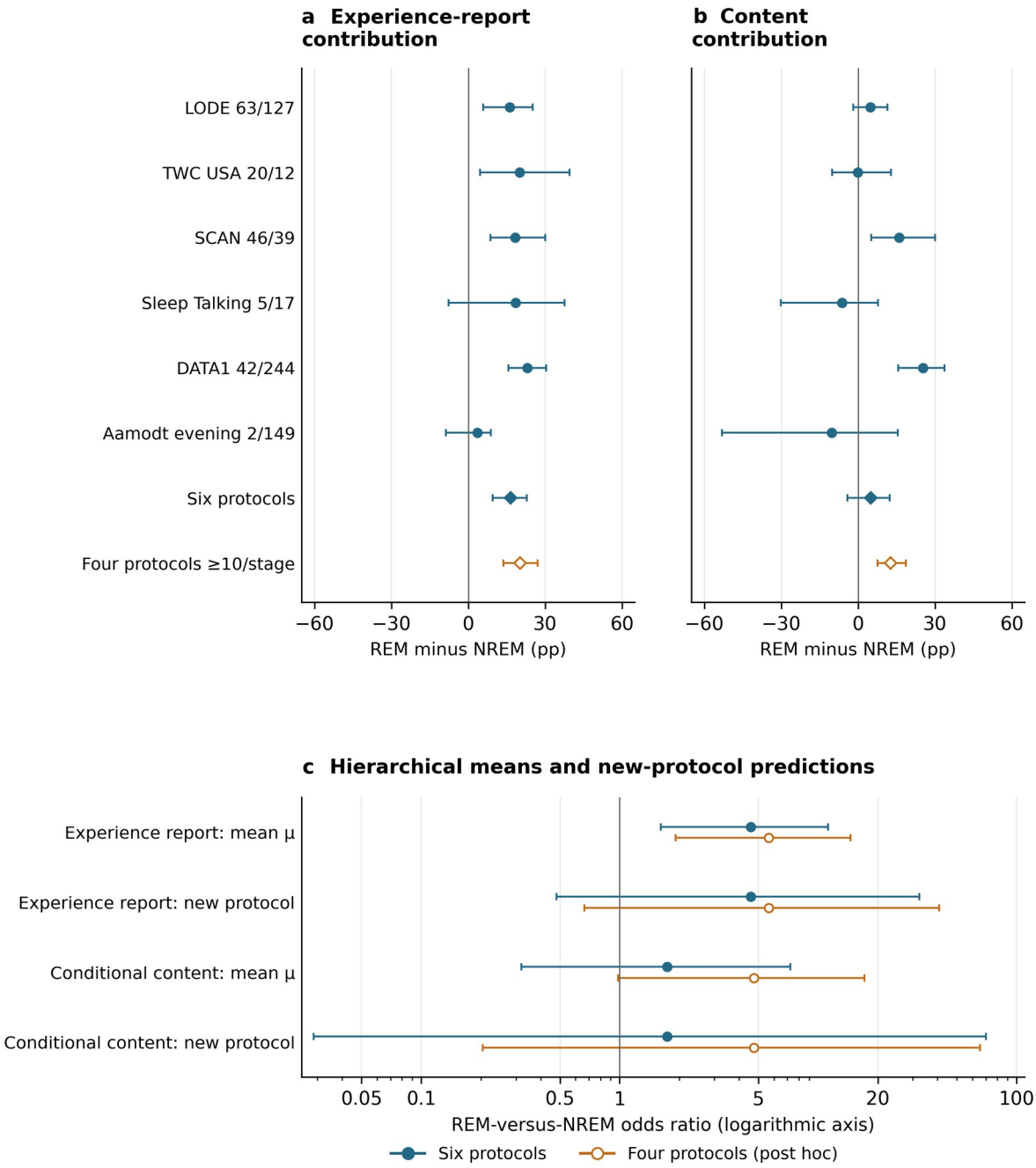
a, Posterior mean of G′, with a black contour at zero. b, P(G′ > 0), with contours at .025, .5 and .975. The blue edge transfers only NREM NE, the orange diagonal transfers the same fraction at both stages, and the open circle denotes deposited categories. Contours interpolate a grid with 0.05 spacing. The .025 to .975 region corresponds to an equal-tailed interval including zero.

Supplementary Figure D5 Separate consequences along the two specified paths



a, G′. b, Experience-report contribution. c, Experience-report contribution minus content contribution. Lines show posterior means and shading 95% CrIs. Blue transfers only NREM NE, and orange transfers the same NE fraction at both stages. The connected grid specifies hypothetical transfer fractions.

Supplementary Figure D6 Dataset contributions and hierarchical stage associations



a–b show posterior means and 95% CrIs for the absolute contributions from the six-dataset fit. Dataset labels give REM/NREM records. Diamonds show equal-protocol means for the six- and four-dataset fits. c exponentiates the updated log-OR means and interval endpoints in Table D5. New-protocol quantiles integrate the saved normal-hierarchy mixture; Table D42 reports prior sensitivity. Both reporting outcomes are coded positive. Main Figures 3–4 compare observed and modelled dataset contributions.

D10 Comparison of stored and aligned report contrasts

D10.1 Descriptive comparison and cluster diagnostics

We used the same base of 1,383 classified sleep records from ten datasets supporting experience reporting with REM and NREM (LODE, TWC USA, Children dreaming, SCAN, Multiple awakenings, Sleep Talking, Young adults, DATA1, Aamodt evening and Aamodt morning). The stored E-versus-NE contrast omitted 132 separately stored EWR records and retained 1,251 records. The experience-report contrast used E + EWR versus NE and retained all 1,383. In Children dreaming, Multiple awakenings and Young adults, the stored positive category did not preserve the E/EWR boundary and was retained as combined experience reporting. Their positive labels remain combined.

A paired bootstrap resampled whole participants with all observed-stage records (20,000 replications). Paired risk differences were estimable in 7,988 resamples and non-estimable in 12,012. Table D25 therefore reports point estimates. The model-based comparison in Section D10.2 provides paired uncertainty while retaining sparse REM support.

Supplementary Table D25 Observed stage differences under two report contrasts

Fixed weights	Stored E vs NE, pp	Experience reporting vs NE, pp	Second minus first, pp
Equal dataset weights	17.9	15.8	−2.1
Fixed total-record weights	22.9	18.6	−4.3

Values in Table D25 are fixed-weight averages of observed within-dataset REM-minus-NREM positive-report probabilities. Dataset weights are identical across contrasts. Record weights use each dataset’s share of the 1,383 base records.

D10.2 Joint model and common target

The auxiliary model used 1,383 records from ten datasets and 242 dataset–participant clusters. For the seven sources with separate labels (LODE, TWC USA, SCAN, Sleep Talking, DATA1, Aamodt evening and Aamodt morning), the categorical likelihood factors into a binomial likelihood for positive reports versus NE and a conditional binomial likelihood for E versus EWR. Children dreaming, Multiple awakenings and Young adults contribute only the former because their positive labels are merged. Aamodt morning enters the exploratory native-label comparison while its broad conditional-content eligibility remains indeterminate.

Both logits used dataset intercepts, hierarchical dataset NREM slopes and participant intercepts nested within datasets. Participant intercepts were correlated where both branches were measured. Dataset intercepts had Normal(0, 2.5^2) priors, slope hypermeans Normal(0, 1.5^2), dataset slope standard deviations Half-Normal(1), and participant standard deviations Half-Normal(1.5). The participant correlation was $2u - 1$, with u distributed Beta(2, 2), equivalent to a two-dimensional LKJ(2) prior. Records from all nights were conditionally independent given these predictors and participant effects. The auxiliary model omitted night effects, report-history terms and participant stage slopes.

Within each draw, all base records were predicted at both stages with fitted participant effects retained. Joint category probabilities were averaged within each dataset before contrasts were formed. For separate-label sources, stored means E/(E + NE) and aligned means E + EWR, with each ratio formed after averaging. For merged-positive sources the contrasts coincide by construction. The same fixed dataset weights apply at both stages and to both contrasts. Δ is the aligned stage gap minus the stored stage gap in each draw. The primary auxiliary summary averages the seven sources storing EWR separately (Section D12.5). The ten-dataset mean is secondary.

The eight-dataset refit retained 1,144 records in 201 clusters, excluding Aamodt evening and Aamodt morning under the post hoc rule of at least five participants observed at each stage. A separate ten-dataset auxiliary fit widened only the dataset slope-SD prior from Half-Normal(1) to Half-Normal(2). Computation settings and diagnostic locations are given in Section D3.2.

Supplementary Table D26 Paired stage differences in the auxiliary joint models

Fit	Stored stage difference (pp)	Aligned stage difference (pp)	Paired difference Δ (pp)	P(Δ < 0)
Ten datasets	17.23 (9.63 to 23.79)	15.37 (8.66 to 21.46)	−1.87 (−3.36 to 0.70)	0.947
Eight-dataset refit	21.07 (13.50 to 28.31)	18.60 (11.38 to 25.53)	−2.48 (−3.97 to −0.48)	0.988
Ten datasets wide prior	16.53 (7.97 to 23.31)	15.10 (7.45 to 21.21)	−1.43 (−3.18 to 1.76)	0.882

Posterior means and 95% equal-tailed CrIs. Stored denotes E/(E + NE) and aligned denotes E + EWR, with stage gaps calculated as REM minus NREM. Here Δ is aligned minus stored, the opposite of the omission effect reported in the main text. The ten- and eight-dataset targets include three sources whose Δ is zero by construction. Table D36 gives the primary auxiliary summaries over the seven and five sources storing EWR separately. Posterior sign probabilities describe the fraction of draws in each direction and are presented with full intervals.

Supplementary Table D27 Weighting targets and marginal odds ratio comparisons

Fit and within-dataset target	Dataset weights	Paired Δ (pp)	Median aligned/stored OR ratio
Ten datasets Record distribution	Equal datasets	−1.87 (−3.36 to 0.70)	0.94 (0.88 to 1.09)
Ten datasets Record distribution	Fixed base records	−3.30 (−4.54 to −1.45)	0.88 (0.83 to 0.99)
Ten datasets Equal participants	Equal datasets	−1.87 (−3.39 to 0.77)	0.94 (0.88 to 1.10)
Ten datasets Equal participants	Fixed base records	−3.29 (−4.55 to −1.41)	0.88 (0.84 to 1.00)
Eight-dataset refit Record distribution	Equal datasets	−2.48 (−3.97 to −0.48)	0.91 (0.86 to 1.02)
Eight-dataset refit Record distribution	Fixed base records	−4.08 (−5.25 to −2.98)	0.86 (0.81 to 0.90)
Eight-dataset refit Equal participants	Equal datasets	−2.46 (−3.95 to −0.46)	0.91 (0.86 to 1.02)

Fit and within-dataset target	Dataset weights	Paired Δ (pp)	Median aligned/stored OR ratio
Eight-dataset refit Equal participants	Fixed base records	−4.05 (−5.22 to −2.95)	0.86 (0.81 to 0.90)
Ten datasets wide prior Record distribution	Equal datasets	−1.43 (−3.18 to 1.76)	0.96 (0.89 to 1.15)
Ten datasets wide prior Record distribution	Fixed base records	−3.13 (−4.55 to −0.62)	0.89 (0.83 to 1.05)
Ten datasets wide prior Equal participants	Equal datasets	−1.43 (−3.20 to 1.84)	0.96 (0.89 to 1.16)
Ten datasets wide prior Equal participants	Fixed base records	−3.12 (−4.55 to −0.64)	0.89 (0.83 to 1.05)

Δ uses posterior means; OR ratios use posterior medians, both with 95% equal-tailed CrIs. Marginal odds ratios are formed from the aggregated stage probabilities. Record and participant targets retain identical weights across stages and contrasts. Stored, aligned and the sign of Δ follow Table D26. With equal dataset weights, the ten-dataset Δ is 0.7 times the seven-source Δ because the three merged-label sources contribute zero. This scaling does not apply to record-weighted summaries. The ten-dataset auxiliary posterior re-averaged over only the eight supported datasets gave Δ = −2.31 pp (−3.69 to −0.32), versus −2.48 (−3.97 to −0.48) in the actual eight-dataset refit. The re-averaged target retains Aamodt evening and Aamodt morning in estimation.

Supplementary Table D28 Sampling diagnostics for the auxiliary models

Fit	Max R-hat	Min bulk / tail ESS	Divergences	Min E-BFMI	Max tree depth
Ten datasets	1.0027	2070 / 2013	0	0.733	7
Eight-dataset refit	1.0045	2160 / 2368	0	0.733	7
Ten datasets wide prior	1.0049	1231 / 1374	0	0.705	7

Diagnostics cover parameters and paired Δ. No fit reached tree depth 12. Coherent category-count predictions used 1,000 replications per fit, drawing content conditional on simulated positive counts. Counts outside marginal 95% prediction intervals were 0/54, 0/42 and 0/54 for ten datasets, eight datasets and ten datasets with wider prior. Ten-dataset prior prediction used 500 draws. Saved prediction reconstruction agreed within 10^{-12} and verified likelihood factorisation and weights.

Supplementary Table D29 Observed stage support for the auxiliary ten-dataset sample

Dataset	Stored resolution	REM counts; participants k	NREM counts; participants k	Both-stage participants
LODE	E/EWR/NE	46/3/14; k=24	72/11/44; k=27	23
TWC USA	E/EWR/NE	18/1/1; k=12	8/0/4; k=12	5
Children dreaming	Positive/NE	2/3; k=5	11/14; k=25	0
SCAN	E/EWR/NE	44/0/2; k=18	19/7/13; k=16	16
Multiple awakenings	Positive/NE	125/29; k=19	188/92; k=19	19
Sleep Talking	E/EWR/NE	2/1/2; k=5	9/0/8; k=12	5
Young adults	Positive/NE	20/10; k=30	20/15; k=35	0

Dataset	Stored resolution	REM counts; participants k	NREM counts; participants k	Both-stage participants
DATA1	E/EWR/NE	37/1/4; k=10	70/62/112; k=10	10
Aamodt evening	E/EWR/NE	1/1/0; k=2	108/26/15; k=25	2
Aamodt morning	E/EWR/NE	1/0/0; k=1	55/19/13; k=16	1

Separated sources list E/EWR/NE and combined sources list positive/NE. Participants are dataset–subject keys. Children dreaming and Young adults have no participants observed at both stages. Aamodt evening and Aamodt morning have only two and one REM participants and no REM negatives. NREM retains the source's substage mixture.

D11 Observed-stage restriction and uncertain-answer inclusion checks

These post hoc checks are recorded in Table D2.

Table D1 lists 61 of 112 clusters observed at both stages: LODE 23/28, TWC USA 5/19, SCAN 16/18, Sleep Talking 5/12, DATA1 10/10 and Aamodt evening 2/25. Their 583 records include 164 REM and 419 NREM records, and six REM EWR reports from five clusters. Restriction removes wholly unobserved participant-stage predictions while changing the participant mixture.

The hierarchy in Section D1 was refitted with the same priors and an algebraically factored likelihood (seed 2026102201). Computation settings are given in Section D3.2. Restricted records were predicted at both stages. Joint probabilities were averaged within dataset and decomposed before aggregation, with record weights based on the restricted totals.

Maximum rank-normalised R-hat was 1.0032, minimum bulk and tail ESS 2391 and 2577, with no divergences or depth-limit hits. Minimum E-BFMI was 0.731. Likelihood factorisation agreed within 7.2×10^{-15}.

Supplementary Table D30 Same-model refit restricted to participants observed at both stages

Dataset weights	Total gap (pp)	Experience contribution (pp)	Content contribution (pp)	G (pp)
Equal dataset	26.5 (14.0 to 37.5)	17.2 (10.1 to 24.5)	9.3 (−1.2 to 18.5)	15.5 (1.6 to 29.9)
Restricted record	35.9 (27.8 to 43.4)	19.8 (14.3 to 25.2)	16.1 (10.3 to 21.3)	14.4 (4.6 to 24.2)

Posterior means and 95% equal-tailed credible intervals. Both targets retain 583 records and 61 clusters. The equal-dataset experience-report contribution is positive and the content interval includes zero. Relative to the primary fit, the check changes participant composition, stage support and hierarchical estimation.

DATA1 excluded 34 joint NE/EWR sleep records (one REM, 33 NREM), and Multiple awakenings excluded 21 (one REM, 20 NREM). For each source, uncertain answers at each stage were assigned wholly to NE or EWR, giving four assignments. Expanded denominators are 43/277 REM/NREM in DATA1 and 155/300 in Multiple awakenings. Wake, Unknown and missing-stage records stay excluded. Table D31 gives raw arithmetic; Section D12.6 adds hierarchical refits.

Supplementary Table D31 Expanded-denominator scenarios for documented uncertain answers

Uncertain-answer assignment	Experience-report stage gap (pp)	Experience contribution (pp)	Content contribution (pp)	G (pp)
DATA1 REM → NE; NREM → NE	40.7	30.6	30.2	20.7
DATA1 REM → NE; NREM → EWR	28.8	20.1	40.6	−3.2
DATA1 REM → EWR; NREM → NE	43.0	31.8	28.9	25.3
DATA1 REM → EWR; NREM → EWR	31.1	21.4	39.4	1.5
Multiple awakenings REM → NE; NREM → NE	18.0	Not identified	Not identified	Not identified
Multiple awakenings REM → NE; NREM → EWR	11.3	Not identified	Not identified	Not identified
Multiple awakenings REM → EWR; NREM → NE	18.6	Not identified	Not identified	Not identified
Multiple awakenings REM → EWR; NREM → EWR	12.0	Not identified	Not identified	Not identified

Under these raw assignments, DATA1's experience-report gap ranges from 28.8 to 43.0 pp, its experience contribution from 20.1 to 31.8 pp, and G from −3.2 to 25.3 pp. In Multiple awakenings, the experience-report gap ranges from 11.3 to 18.6 pp. Its positive E/EWR split is unmeasured, leaving content contributions and G unidentified. These are extrema over the specified assignments.

These assignments add previously excluded answers and enlarge the denominators. Section D9 transforms retained NE probability with fixed denominators. In an expanded all-NE baseline, uncertain answers form 20.0% and 22.8% of DATA1's REM and NREM negatives, and 3.3% and 17.9% of those in Multiple awakenings. The percentages describe the excluded joint-answer subset.

D12 Additional estimand and predictive checks

These post hoc analyses (Table D2) use the primary six-dataset posterior unless stated otherwise. Additional fits retain the likelihood and priors in Section D1. Specifications, record identities, seeds and posterior parameters accompany the code.

D12.1 Participant-stage prediction support

For a prediction at either stage, the unobserved-stage fraction is the total-record weight of participants observed only at the other stage. The two-stage column averages these two fractions. The fraction counts only participants with no observation at the predicted stage. Sparse observations or missing timing and covariate support can require additional extrapolation.

Supplementary Table D32 Standardisation weight from wholly unobserved participant stages

Dataset	REM / NREM participants	Both / all	REM target (%)	NREM target (%)	Two-stage mean (%)
LODE	24/27	23/28	6.3	1.1	3.7
TWC USA	12/12	5/19	21.9	28.1	25.0
SCAN	18/16	16/18	0.0	3.5	1.8
Sleep Talking	5/12	5/12	54.5	0.0	27.3
DATA1	10/10	10/10	0.0	0.0	0.0
Aamodt evening	2/25	2/25	91.4	0.0	45.7

D12.2 Awakening mode

Protocols are grouped as experimenter-induced (SCAN, DATA1 and Aamodt evening), final-morning (LODE and Sleep Talking), and mixed (TWC USA). LODE is spontaneous; Sleep Talking's induction method is unspecified, and TWC USA has only a protocol-level mode description. Raw and model contributions are calculated within each protocol before equal averaging within groups. Model intervals reuse the six-dataset posterior.

Supplementary Table D33 Contributions grouped by awakening procedure

Procedure group	Raw experience (pp)	Raw content (pp)	Model experience (pp)	Model content (pp)
Experimenter-induced SCAN, DATA1 and Aamodt evening	19.7	8.3	15.0 (9.6 to 20.0)	10.3 (−4.3 to 21.0)
Final-morning LODE and Sleep Talking	8.6	−6.9	17.3 (2.4 to 28.8)	−0.8 (−13.0 to 7.5)
Mixed TWC USA	27.6	−4.3	20.0 (4.5 to 39.5)	−0.2 (−10.1 to 12.8)

D12.3 Documentary restriction and participant-distribution standardisation

The documentary restriction excludes TWC USA and Sleep Talking for broadened NE or an unspecified reference period. It retains LODE, SCAN, DATA1 and Aamodt evening: 712 records in 81 clusters, including five REM EWR reports from four clusters. The four-dataset subset instead retains LODE, TWC USA, SCAN and DATA1. Equal and record weights use the retained sample.

A supplementary refit of the six-dataset model supplied participant-effect parameters for integration and predictive checks. Dataset-specific contribution means differed from the primary fit by at most 0.21 pp. Population integration averages joint probabilities over the model's correlated bivariate normal participant distribution within the same six protocols. The 40-node-per-dimension Gauss–Hermite result differed from 24 nodes by at most 0.0023 pp in any category probability.

Supplementary Table D34 Additional model and standardisation targets

Target or model	Experience contribution (pp)	Content contribution (pp)	Total gap (pp)
Six-dataset supplementary refit; equal	16.6 (9.7 to 22.9)	4.8 (−4.6 to 12.2)	21.4 (10.0 to 31.4)
Integrated participant effects; equal	16.3 (9.4 to 22.9)	4.6 (−4.8 to 12.3)	20.9 (9.4 to 31.2)
Integrated participant effects; records	16.9 (11.2 to 22.5)	9.5 (0.1 to 17.2)	26.4 (15.8 to 35.7)
Exclude TWC USA and Sleep Talking; equal	15.2 (9.3 to 20.3)	12.1 (1.5 to 18.8)	27.3 (16.0 to 36.0)
Exclude TWC USA and Sleep Talking; records	16.4 (10.9 to 21.2)	13.8 (4.7 to 20.0)	30.3 (20.1 to 38.5)
Common content slope; equal	16.6 (9.4 to 23.0)	10.3 (5.8 to 15.2)	26.9 (18.3 to 34.5)

The supplementary six-dataset refit estimated the shared participant-intercept correlation as 0.04 (95% CrI −0.47 to 0.52). The broad interval permits either direction of association between stable experience-report and content-report tendencies.

D12.4 Common and varying content slopes

The comparison model retains dataset intercepts, hierarchical experience-report slopes and correlated participant intercepts, but constrains all six content slopes to a common zero-centred normal coefficient with standard deviation 1.5. The varying-slope fit uses Section D1. For each posterior draw, white-dream counts are predicted conditional on observed participant-stage experience-report totals. Predictions are aggregated into the 12 dataset–stage cells. The discrepancy sums squared differences between observed and expected white-dream counts, divided by the summed conditional binomial variance in each cell. Four thousand predictive replications use evenly spaced posterior draws and recorded seeds.

The replicated discrepancy exceeded the observed discrepancy in 0.025 of common-slope replications and 0.362 of varying-slope replications. These are conditional, in-sample posterior predictive checks of the observed white-dream count pattern. Together with Table D17’s prior comparison, they support allowing protocol variation while leaving its magnitude imprecise.

Supplementary Table D35 Conditional white-dream count predictions

Dataset / stage	Observed EWR / experience reports	Common-slope predicted EWR	Varying-slope predicted EWR
LODE REM	3 / 49	2.5 (0.0 to 7.0)	3.6 (0.0 to 9.0)
LODE NREM	11 / 83	12.0 (5.0 to 21.0)	11.0 (4.0 to 20.0)
TWC USA REM	1 / 19	0.7 (0.0 to 3.0)	1.1 (0.0 to 4.0)

Dataset / stage	Observed EWR / experience reports	Common-slope predicted EWR	Varying-slope predicted EWR
TWC USA NREM	0 / 8	0.9 (0.0 to 3.0)	0.4 (0.0 to 3.0)
SCAN REM	0 / 44	2.5 (0.0 to 7.0)	1.5 (0.0 to 5.0)
SCAN NREM	7 / 26	5.0 (1.0 to 10.0)	6.2 (1.0 to 12.0)
Sleep Talking REM	1 / 3	0.2 (0.0 to 1.0)	0.6 (0.0 to 2.0)
Sleep Talking NREM	0 / 9	1.4 (0.0 to 4.0)	0.7 (0.0 to 3.0)
DATA1 REM	1 / 38	4.7 (1.0 to 10.0)	2.4 (0.0 to 7.0)
DATA1 NREM	62 / 132	58.6 (44.0 to 73.0)	61.1 (46.0 to 76.0)
Aamodt evening REM	1 / 2	0.2 (0.0 to 1.0)	0.7 (0.0 to 2.0)
Aamodt evening NREM	26 / 134	27.1 (16.0 to 39.0)	26.4 (15.0 to 39.0)

Entries are predictive means and 95% equal-tailed intervals for white-dream counts. Both models condition on the same experience-report denominators and include the sparse cells without continuity correction.

D12.5 Contrast differences where EWR is stored separately

The primary auxiliary target reuses the ten-dataset auxiliary model and averages its seven sources storing EWR separately (LODE, TWC USA, SCAN, Sleep Talking, DATA1, Aamodt evening and Aamodt morning). The eight-dataset refit includes five such sources (LODE, TWC USA, SCAN, Sleep Talking and DATA1). The remaining three sources have identical stored and aligned probabilities by construction. At equal weights, every ten-dataset Δ draw is exactly 0.7 times its seven-source counterpart, scaling its magnitude and interval width while preserving its sign probability and zero inclusion.

Supplementary Table D36 Paired contrasts among sources storing EWR separately

Fit and target	Protocol weights	Stored gap (pp)	Aligned gap (pp)	Paired Δ (pp)
Ten-dataset auxiliary model; seven-source target	Equal	18.17 (10.09 to 25.58)	15.51 (8.81 to 21.97)	−2.67 (−4.80 to 1.01)
Ten-dataset auxiliary model; seven-source target	Base records	22.87 (15.85 to 29.13)	17.52 (11.70 to 23.05)	−5.35 (−7.36 to −2.34)
Wider prior; seven-source target	Equal	17.42 (8.43 to 24.92)	15.38 (8.18 to 21.84)	−2.04 (−4.55 to 2.52)

Fit and target	Protocol weights	Stored gap (pp)	Aligned gap (pp)	Paired Δ (pp)
Wider prior; seven-source target	Base records	22.60 (14.99 to 29.15)	17.53 (11.43 to 23.06)	−5.07 (−7.37 to −1.01)
Eight-dataset refit; five-source target	Equal	24.26 (15.81 to 33.15)	20.29 (12.53 to 28.44)	−3.96 (−6.35 to −0.76)
Eight-dataset refit; five-source target	Base records	30.49 (22.51 to 38.11)	22.90 (15.51 to 29.79)	−7.59 (−9.77 to −5.55)

Posterior means and 95% CrIs. Gaps are REM minus NREM, and Δ is aligned minus stored. The seven-source primary auxiliary result is −2.67 pp (−4.80 to 1.01). The wider-prior and support-restricted targets remain explicit. Equal-participant versions and all ten-dataset summaries are retained in the numerical output and Tables D26–D27. Aamodt morning's unresolved broad conditional-content eligibility is unchanged.

D12.6 Hierarchical refits including uncertain answers

The 55 joint NE/EWR sleep records comprise 34 in DATA1 and 21 in Multiple awakenings. Adding them to the auxiliary sample gives 1,438 records in 242 existing clusters. Four fits independently assign the uncertain answers at REM and NREM wholly to NE or wholly to EWR. The source-specific positive code remains merged in Multiple awakenings: assigning its uncertain answer to experience adds a positive observation without predicting E or EWR separately. All resolved observations, priors and hierarchical structures match the ten-dataset auxiliary baseline. Refits update participant effects and partial pooling, unlike Section D9's fixed-prediction transforms. Standardisation uses each fit's expanded within-dataset record mixture.

Assigning all uncertain DATA1 answers to NE gives G = 21.4 pp and assigning them all to EWR gives 0.0 pp, compared with 12.0 pp for resolved answers only. The experience-report contribution remains positive in all four assignments. Raw uncertain-answer frequencies are 33/277 NREM versus 1/43 REM in DATA1 and 20/300 versus 1/155 in Multiple awakenings. Table D39 gives BCa intervals for these rates and their stage differences, based on resampling participants with both stages kept together.

Supplementary Table D37 Uncertain-answer refits and their resolved-record baseline

Source and assignment	Experience-report gap (pp)	Experience contribution (pp)	Content contribution (pp)	G (pp)
DATA1 Resolved baseline	28.9 (17.2 to 39.8)	21.2 (12.4 to 29.5)	24.6 (14.8 to 33.2)	12.0 (−4.1 to 28.9)
DATA1 REM → NE; NREM → NE	35.4 (22.7 to 46.2)	26.0 (16.5 to 34.3)	23.5 (14.1 to 31.4)	21.4 (4.8 to 38.0)
DATA1 REM → NE; NREM → EWR	23.1 (13.2 to 33.4)	15.7 (8.9 to 23.0)	34.5 (24.9 to 42.5)	−4.0 (−18.3 to 12.1)
DATA1 REM → EWR; NREM → NE	37.6 (24.4 to 48.1)	27.0 (17.4 to 35.1)	22.5 (12.8 to 30.9)	25.7 (7.7 to 42.7)
DATA1 REM → EWR; NREM → EWR	25.0 (14.8 to 35.1)	16.7 (9.8 to 23.9)	33.3 (23.1 to 42.0)	0.0 (−15.2 to 16.8)

Source and assignment	Experience-report gap (pp)	Experience contribution (pp)	Content contribution (pp)	G (pp)
Multiple awakenings Resolved baseline	15.0 (7.1 to 22.3)	Not identified	Not identified	Not identified
Multiple awakenings REM → NE; NREM → NE	18.9 (11.1 to 26.2)	Not identified	Not identified	Not identified
Multiple awakenings REM → NE; NREM → EWR	12.9 (5.1 to 19.8)	Not identified	Not identified	Not identified
Multiple awakenings REM → EWR; NREM → NE	19.5 (11.6 to 26.8)	Not identified	Not identified	Not identified
Multiple awakenings REM → EWR; NREM → EWR	13.4 (5.7 to 20.3)	Not identified	Not identified	Not identified

Posterior means and 95% CrIs from the auxiliary model under four assignments of uncertain answers, alongside the resolved-answer baseline. G is the NREM-minus-REM increase in NE minus the increase in EWR.

DATA1's five-level certainty item (item N1 in the source questionnaire) precedes the white-dream item (Wong et al., 2020, Supplementary Document 2, p. 2). Case-linked certainty ratings were unavailable in the supplied files, and depositing them would support stage-specific confidence analyses. Signal-detection or meta-d′ estimation also requires suitable accuracy information or latent-state assumptions (Maniscalco & Lau, 2012).

D12.7 Computation

Computation settings are given in Section D3.2, and Table D38 reports the seven additional fits, including the supplementary refit. Factorised and categorical log likelihoods agreed within 10^{-10}. Model parameters and input identities are saved for each assignment.

Supplementary Table D38 Diagnostics for the additional fits

Fit	Max R-hat	Min bulk / tail ESS	Divergences	Min E-BFMI	Max tree depth
Six-dataset supplementary refit	1.0033	2433 / 2257	0	0.750	7
Common content slope	1.0032	2964 / 3155	0	0.751	7
Exclude TWC USA and Sleep Talking	1.0025	2272 / 2263	0	0.705	7
Uncertain answers: NE / NE	1.0032	1958 / 2237	0	0.716	7
Uncertain answers: NE / EWR	1.0025	1467 / 1693	0	0.691	7
Uncertain answers: EWR / NE	1.0031	1535 / 1542	0	0.688	7
Uncertain answers: EWR / EWR	1.0060	1600 / 1990	0	0.672	8

D13 Report contrasts and uncertainty checks

These post hoc checks (Table D2) use existing records and saved posterior draws.

D13.1 Uncertain answer rates

Uncertain answers are the deposited joint NE/EWR category. Denominators include resolved E, EWR and NE sleep records plus these answers, and exclude Unknown, Wake and missing-stage records. Whole participants are resampled with all their REM and NREM records together. We use 100,000 resamples and bias-corrected and accelerated (BCa) 95% intervals, with acceleration from a delete-participant jackknife (Efron, 1987). Seeds are 20261038 for DATA1 and 20261031 for Multiple awakenings. Every resample retained both stages. The calculation is exploratory: few clusters and a single uncertain REM answer in each source limit rare-event interval calibration. Percentile intervals are retained in the numerical output to expose method dependence.

Supplementary Table D39 Uncertain answers by stage and participant cluster

Dataset	Participants	REM uncertain / total Rate (%), 95% BCa interval	NREM uncertain / total Rate (%), 95% BCa interval	NREM − REM (pp) 95% cluster BCa interval
DATA1	10	1/43 2.3 (0.0 to 11.1)	33/277 11.9 (8.0 to 15.3)	9.6 (−0.1 to 14.5)
Multiple awakenings	19	1/155 0.6 (0.0 to 3.3)	20/300 6.7 (3.4 to 11.8)	6.0 (3.1 to 10.1)

DATA1's BCa difference interval includes zero, whereas its simple percentile interval is 1.6 to 15.2 pp. For Multiple awakenings the percentile interval is 2.9 to 9.7 pp. The method dependence reflects the sparse uncertainty events. Table D37 separately estimates how assigning these answers changes the model-based category balance.

D13.2 Alternative decompositions

The symmetric experience-report contribution multiplies the stage difference in experience reporting by average conditional content. The NREM-content decomposition uses the NREM content probability instead, leaving the content difference weighted by REM experience reporting. The REM-content decomposition reverses these baselines. We calculated both within each of the 8,000 primary posterior draws and each dataset, then averaged contributions equally across protocols. Each decomposition sums exactly to the same contentful-report gap. Shares divide the matched-draw contribution by that gap.

Supplementary Table D40 Contributions and shares under three decompositions

Dataset / target	Content baseline for experience term	Experience contribution (pp)	Content contribution (pp)	Experience share (%)
Six equal	NREM content	15.4 (8.3 to 21.7)	6.0 (−3.8 to 14.1)	70.8 (46.6 to 131.8)
Six equal	Symmetric	16.6 (9.6 to 22.9)	4.9 (−4.1 to 12.3)	76.1 (53.4 to 136.4)
Six equal	REM content	17.7 (10.7 to 24.3)	3.7 (−4.6 to 10.5)	81.2 (60.1 to 139.9)
LODE	NREM content	15.5 (5.6 to 24.4)	5.4 (−2.1 to 12.8)	73.9 (42.6 to 115.1)

Dataset / target	Content baseline for experience term	Experience contribution (pp)	Content contribution (pp)	Experience share (%)
LODE	Symmetric	16.1 (5.9 to 25.1)	4.8 (−1.9 to 11.4)	77.0 (46.4 to 113.3)
LODE	REM content	16.7 (6.1 to 26.1)	4.2 (−1.6 to 10.1)	80.1 (49.9 to 111.7)
SCAN	NREM content	16.3 (7.6 to 27.4)	18.1 (5.8 to 33.9)	47.9 (22.7 to 77.8)
SCAN	Symmetric	18.4 (8.7 to 30.0)	16.0 (5.1 to 30.1)	53.9 (28.4 to 80.9)
SCAN	REM content	20.4 (9.6 to 33.0)	13.9 (4.4 to 26.3)	60.2 (33.8 to 84.2)
DATA1	NREM content	17.3 (11.6 to 23.5)	30.8 (18.9 to 41.6)	35.7 (25.4 to 50.1)
DATA1	Symmetric	22.9 (15.7 to 30.4)	25.2 (15.6 to 33.6)	47.5 (36.5 to 60.2)
DATA1	REM content	28.5 (19.3 to 37.9)	19.6 (12.1 to 26.4)	59.3 (47.3 to 70.5)

Contributions are posterior means and shares posterior medians, both with 95% CrIs. The equal-protocol mean experience contribution is larger under every decomposition. Under the NREM-content decomposition its 95% share interval crosses 50%, allowing a larger content contribution within posterior uncertainty. The analysis archive contains all six dataset summaries.

D13.3 Negative report shares

Within each primary posterior draw, the negative-report share is the NREM-minus-REM increase in NE divided by the REM-minus-NREM decrease in E. It describes category redistribution, whereas product decomposition weights each probability change by the other report probability. We report the median and equal-tailed CrI alongside the ratio of posterior mean category changes (Section D8).

Supplementary Table D41 Negative report shares of the fall in contentful reports

Dataset / target	Ratio of means (%)	Median share (%) 95% CrI	Draws with positive E gap (%)
LODE	85.9	85.5 (52.5 to 128.5)	100.0
TWC USA	106.9	105.6 (40.5 to 301.7)	97.9
SCAN	61.6	62.1 (34.9 to 88.2)	100.0
Sleep Talking	176.5	114.8 (−782.0 to 1022.1)	79.4
DATA1	64.9	64.5 (50.2 to 84.7)	100.0
Aamodt evening	−72.4	−0.1 (−508.1 to 530.2)	42.7
Six equal	91.6	89.6 (64.0 to 165.8)	100.0

The published Siclari experiment 2 counts give a descriptive negative-report share of 54.0%. Its uncertainty depends on the benchmark’s participant clustering and any cohort overlap (Section D8).

D13.4 Prior sensitivity of new protocol predictions

We retrieved the primary six-dataset posterior hyperparameters for Half-Normal slope-SD scales 0.5, 1 and 2, the same fitted models summarised in Table D8. For each reporting outcome, the predictive log-odds distribution is the equally weighted mixture of Normal(μ, τ²) over 8,000 saved draws. Numerical inversion of its averaged normal cumulative distribution yields the .025, .5 and .975 quantiles, avoiding a further random prediction draw. Exponentiation gives the odds-ratio quantiles. Table D5 and Figure D6 use the same deterministic integration.

Supplementary Table D42 New protocol predictions across slope priors

Half-Normal scale	Reporting outcome	Slope SD: mean 95% CrI	New-protocol OR: median 95% predictive interval
0.5	Experience reporting	0.46 (0.03 to 1.13)	5.09 (1.16 to 19.08)
0.5	Conditional content	0.86 (0.10 to 1.61)	2.93 (0.22 to 20.98)
1	Experience reporting	0.74 (0.05 to 1.82)	4.85 (0.48 to 32.41)
1	Conditional content	1.65 (0.54 to 2.89)	1.93 (0.03 to 70.16)
2	Experience reporting	0.96 (0.06 to 2.56)	4.74 (0.23 to 51.11)
2	Conditional content	2.48 (0.94 to 4.56)	1.45 (0.003 to 338.68)
1; four datasets	Experience reporting	0.71 (0.06 to 1.79)	5.80 (0.66 to 40.97)
1; four datasets	Conditional content	1.02 (0.06 to 2.35)	5.34 (0.20 to 65.19)

The primary experience-slope SD posterior mean is 0.74 (95% CrI 0.05 to 1.82), close to the Half-Normal(1) prior mean 0.80 (0.03 to 2.24). The alternative scales give means 0.46 and 0.96. The six protocols provide little information about experience-report heterogeneity. The predictive interval for a new protocol has a lower limit above or below one depending on the prior scale.

D13.5 Odds identity and probability scale

At stage s, let a denote the experience-report probability and c conditional content, as in Figure 1. With EWR omitted, the odds of E versus NE are:

$$Odds_s(E:NE) = \frac{a_s c_s}{1 - a_s} = \frac{a_s}{1 - a_s} c_s$$

Taking the REM-versus-NREM ratio gives:

$$\mathrm{OR}_{E:NE} = \mathrm{OR}_{experience} \times \frac{c_{REM}}{c_{NREM}}$$

The second factor is a ratio of conditional content probabilities, not their odds ratio. The identity holds for each coherent three-category distribution, including the standardised within-dataset probabilities. Marginalising across datasets does not preserve a product of averaged odds ratios. For merged positive labels, both targets are already E + EWR versus NE.

A larger odds ratio also need not give a larger absolute stage gap. For example, experience-report probabilities of 0.60/0.30 and conditional content of 0.20/0.10 at REM/NREM give an experience-report OR of 3.5 and E-versus-NE OR of 7.0. Yet the respective probability gaps are 30.0 and 19.0 pp. Table D36 reports matched-draw probability-gap differences for the specified baseline rates and averaging targets.

Supplementary Material E Verification and reproduction

The author completed item-level verification of sources, code, numerical results and key inferences, including the BCa bootstrap method and its citation (Efron, 1987), with evidence, dates and conclusions recorded. Re-running the analysis package in a separate computing environment reproduced the reported outputs. The author confirmed the reported results on 25 October 2026. Main-text Section 7 discloses AI assistance.

A presentation revision following simulated-review feedback on 22 October 2026 reordered questions, findings and implications and reformatted figures and tables using existing summaries.

Supplementary Material F External source availability and exploratory comparisons

External quantitative comparison requires comparable REM/NREM E/EWR/NE definitions, verifiable report denominators and participant or night structure. Table F1 records the materials examined and the requirements still unresolved in the exploratory search. The Europe PMC search retained 1,244 unique retrieval records; full-text eligibility screening, other database coverage and citation tracing remained incomplete. Cohort overlap with DREAM also requires verification before cross-source synthesis.

Supplementary Table F1 Examples of unresolved requirements for external quantitative synthesis

Source	Material examined	Remaining issue	Evidence locator
Schoch et al. (2019)	Actual published S1 obtained and read	The supplement specifies selection, awakenings and immediate reporting, but not separate EWR/NE counts; the published overall and stage-specific awakening totals remain unaligned.	Published S1; Results 3.2
Martin et al. (2020)	Published PDF, Figure 4 and actual S1 read	The PIRS definitions distinguish no experience and white dreams, but published N2 percentages sum above 100% and disagree with the figure. A self-consistent stage table and operational coding remain unresolved.	S1; Figure 4; PDF p. 9
Ataei et al. (2026)	Methods and data-access statement read	The EEG cohort uses a retrospective frequency outcome. Morning category outcomes in the other cohorts use Fitbit stages; Dataset 3 combines the final-awakening items with other dreams. Raw final-awakening items would require separate assessment.	Methods pp. 6–10; data statement p. 28
Known DREAM source reports	Article and deposited dataset links traced	A separate article is not automatically a new cohort. Exact participant, experiment and version correspondence is needed before adding independent evidence.	DREAM version 9 sources in Section B1
Stephan et al. (2021)	Formal methods, Figure S1 and actual Data S1/S2 read	Data S1 contains 58 illustrative report rows; Data S2 contains 27 published model tables. Conditional integer compatibility does not establish the source aggregation rule, complete report flow, target G covariance or independence from existing sources.	Figure S1A/C; STAR Methods; Data S1 A-D; Data S2 Table 8
Kumar et al. (2025)	Formal methods, Table 2 and supplied supplement read	Original dream labels and stage integers available. Broad experience mapping, response flow, protocol comparability and participant-level dependence remain unresolved. Nights 2/3 are the same cohort.	Main PDF pp. 3-4 and 7, Table 2; supplied supplementary material
Feige et al. (2018)	Formal paper and supplied supplementary figure read	The published report uses binary mentation and clarity measures without a separate target EWR/NE distinction. Retained as binary auxiliary evidence, not a three-category primary input.	Methods pp. 3-5; Table 3; recruitment supplement

A later search identified 13 further candidate articles. Two full texts were obtained, and the other 11 were not assessed. The descriptive calculations were made after the DREAM results were known.

F1 Conditional reconstruction from Stephan et al. 2021

For Stephan’s healthy group of 20 participants, Figure S1 lists 408 NREM (N2/N3) and 126 REM classified reports with one-decimal category percentages. If those are pooled, unweighted proportions

with common denominators rounded to 0.1 percentage point, compatible CE/CEWR/NE triplets are 163/158/87 and 80/33/13. The reconstructed integers depend on these assumptions. A full awakening flow and participant dependence are needed for sampling inference; Data S1 supplies illustrative reports and Data S2 supplies repeated-measures models.

Supplementary Table F2 Listed or conditionally compatible counts in the exploratory sources

Source and stage	Original label sequence and counts	Classified denominator
Stephan healthy NREM	CE/CEWR/NE: 163/158/87	408 (conditional)
Stephan healthy REM	CE/CEWR/NE: 80/33/13	126 (conditional)
Kumar night 2 N2	DR/DNR/ND: 56/32/25	113
Kumar night 2 N3	DR/DNR/ND: 7/1/8	16
Kumar night 2 REM	DR/DNR/ND: 73/12/8	93
Kumar night 3 N2	DR/DNR/ND: 62/22/39	123
Kumar night 3 N3	DR/DNR/ND: 3/4/11	18
Kumar night 3 REM	DR/DNR/ND: 75/9/15	99

CE, conscious experience with content; CEWR, conscious experience without recalled content. DR, Dream Recall; DNR, Dream No Recall; ND, No Dream. Kumar's original dream-focused labels are analysed on their own terms. Stephan's integers retain all stated reconstruction conditions.

Conditional Stephan counts give NREM increases of 11.01 pp in NE and 12.54 pp in CEWR, hence G = −1.53 pp. The symmetric experience-report and content contributions are 6.69 and 16.85 pp, whose difference is −10.16 pp; aligned-minus-stored Δ is −9.82 pp. Allowing only display rounding gives a G range of −1.70 to −1.30 pp. Sampling intervals require the unresolved report flow, aggregation and participant dependence. Adding three unlisted REM CEWR reports or six NREM NE reports would reverse the conditional point sign; these are hypothetical sensitivity counts.

F2 Native dream label comparisons from Kumar et al. 2025

Kumar's published counts use Dream Recall, Dream No Recall and No Dream, separately by night and N2/N3. Table F3 compares each NREM group with its same-night REM group. The comparison retains these dream-focused labels; translation to broad E/EWR/NE requires additional definitional evidence. The two nights involve the same 29 participants.

Supplementary Table F3 Descriptive differences between original label increments

Source and contrast	Negative-label increase (pp)	Unrecalled-label increase (pp)	First minus second (pp)
Stephan NREM minus REM (conditional)	11.01	12.54	−1.53
Kumar night 2 N2 minus REM	13.52	15.42	−1.89
Kumar night 2 N3 minus REM	41.40	−6.65	48.05
Kumar night 3 N2 minus REM	16.56	8.80	7.76
Kumar night 3 N3 minus REM	45.96	13.13	32.83

For Stephan, the final column compares the NE and CEWR increases and is conditional G. For Kumar, it compares No Dream and Dream No Recall under the original dream-focused labels. The tables retain the source-specific point calculations. Sampling inference and cross-source synthesis require resolution of the qualification and dependence issues above.

Supplementary references

Aamodt, A., Nilsen, A. S., Thürer, B., Hasanzadeh Moghadam, F., Kauppi, N., Juel, B. E., & Storm, J. F. (2021). EEG signal diversity varies with sleep stage and aspects of dream experience. Frontiers in Psychology, 12, 655884. https://doi.org/10.3389/fpsyg.2021.655884

Aamodt, A., Sevenius Nilsen, A., Markhus, R., Kusztor, A., HasanzadehMoghadam, F., Kauppi, N., Thürer, B., Storm, J. F., & Juel, B. E. (2023). EEG Lempel-Ziv complexity varies with sleep stage, but does not seem to track dream experience. Frontiers in Human Neuroscience, 16, 987714. https://doi.org/10.3389/fnhum.2022.987714

Abril-Pla, O., Andreani, V., Carroll, C., Dong, L., Fonnesbeck, C. J., Kochurov, M., Kumar, R., Lao, J., Luhmann, C. C., Martin, O. A., Osthege, M., Vieira, R., Wiecki, T., & Zinkov, R. (2023). PyMC: A modern, and comprehensive probabilistic programming framework in Python. PeerJ Computer Science, 9, e1516. https://doi.org/10.7717/peerj-cs.1516

Ataei, S., Jafarzadeh Esfahani, M., Axmacher, N., Dresler, M., & Schoch, S. F. (2026). How awakenings shape dream recall: A multilevel study (Version 1) [Preprint]. bioRxiv. https://doi.org/10.64898/2026.05.20.722678

Das Gupta, P. (1993). Standardization and decomposition of rates: A user's manual (Current Population Reports, P23-186). U.S. Bureau of the Census. https://www2.census.gov/library/publications/1993/demographics/p23-186.pdf

Efron, B. (1987). Better bootstrap confidence intervals. Journal of the American Statistical Association, 82(397), 171–185. https://doi.org/10.1080/01621459.1987.10478410

Eichenlaub, J.-B., van Rijn, E., Gaskell, M. G., Lewis, P. A., Maby, E., Malinowski, J. E., Walker, M. P., Boy, F., & Blagrove, M. (2018). Incorporation of recent waking-life experiences in dreams correlates with frontal theta activity in REM sleep. Social Cognitive and Affective Neuroscience, 13(6), 637–647. https://doi.org/10.1093/scan/nsy041

Elce, V., Bergamo, D., Avvenuti, G., Bellesi, M., & Bernardi, G. (2023). LODE [Data set]. figshare. https://doi.org/10.6084/m9.figshare.22147085

Elce, V., Bergamo, D., Bontempi, G., Pedreschi, B., Bellesi, M., Handjaras, G., & Bernardi, G. (2025). The individual determinants of morning dream recall. Communications Psychology, 3, 25. https://doi.org/10.1038/s44271-025-00191-z

Feige, B., Nanovska, S., Baglioni, C., Bier, B., Cabrera, L., Diemers, S., Quellmalz, M., Siegel, M., Xeni, I., Szentkiralyi, A., Doerr, J.-P., & Riemann, D. (2018). Insomnia—perchance a dream? Results from a NREM/REM sleep awakening study in good sleepers and patients with insomnia. Sleep, 41(5), zsy032. https://doi.org/10.1093/sleep/zsy032

Hernán, M. A., Hernández-Díaz, S., & Robins, J. M. (2004). A structural approach to selection bias. Epidemiology, 15(5), 615–625. https://doi.org/10.1097/01.ede.0000135174.63482.43

Kitagawa, E. M. (1955). Components of a difference between two rates. Journal of the American Statistical Association, 50(272), 1168–1194. https://doi.org/10.1080/01621459.1955.10501299

Konkoly, K. R., Appel, K., Chabani, E., Mangiaruga, A., Gott, J., Mallett, R., Caughran, B., Witkowski, S., Whitmore, N. W., Mazurek, C. Y., Berent, J. B., Weber, F. D., Türker, B., Leu-Semenescu, S., Maranci, J.-B., Pipa, G., Arnulf, I., Oudiette, D., Dresler, M., & Paller, K. A. (2021). Real-time dialogue between experimenters and dreamers during REM sleep. Current Biology, 31(7), 1417–1427.e6. https://doi.org/10.1016/j.cub.2021.01.026

Kumar, G., Naaz, S., Jabin, N., Sasidharan, A., Nagendra, R. P., Yadav, R., & Kutty, B. M. (2025). Neurophysiological features of dream recall and the phenomenology of dreams: Auditory stimulation impacts dream experiences. Consciousness and Cognition, 132, 103869. https://doi.org/10.1016/j.concog.2025.103869

Kumar, R., Carroll, C., Hartikainen, A., & Martin, O. (2019). ArviZ a unified library for exploratory analysis of Bayesian models in Python. Journal of Open Source Software, 4(33), 1143. https://doi.org/10.21105/joss.01143

Kumral, D., Palmieri, J., Gais, S., & Schönauer, M. (2025). Pre-sleep experiences shape neural activity and dream content in the sleeping brain. iScience, 28(8), 113032. https://doi.org/10.1016/j.isci.2025.113032

MacKinnon, J. G., Nielsen, M. Ø., & Webb, M. D. (2025). Cluster-robust jackknife and bootstrap inference for logistic regression models (Version 2) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2406.00650

Mangiaruga, A., D'Atri, A., Scarpelli, S., Alfonsi, V., Camaioni, M., Annarumma, L., Gorgoni, M., Pazzaglia, M., & De Gennaro, L. (2022). Sleep talking versus sleep moaning: electrophysiological patterns preceding linguistic vocalizations during sleep. Sleep, 45(5), zsab284. https://doi.org/10.1093/sleep/zsab284

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Consciousness and Cognition, 21(1), 422–430.

Manski, C. F. (2003). Partial identification of probability distributions. Springer.

Martin, J. M., Andriano, D. W., Mota, N. B., Mota-Rolim, S. A., Araújo, J. F., Solms, M., & Ribeiro, S. (2020). Structural differences between REM and non-REM dream reports assessed by graph analysis. PLOS ONE, 15(7), e0228903. https://doi.org/10.1371/journal.pone.0228903

Marzano, C., Ferrara, M., Mauro, F., Moroni, F., Gorgoni, M., Tempesta, D., Cipolli, C., & De Gennaro, L. (2011). Recalling and forgetting dreams: Theta and alpha oscillations during sleep predict subsequent dream recall. Journal of Neuroscience, 31(18), 6674–6683. https://doi.org/10.1523/JNEUROSCI.0412-11.2011

McNeish, D. (2026). A practical guide to selecting and blending approaches for clustered data: Clustered errors, multilevel models, and fixed-effect models. Psychological Methods, 31(2), 225–251. https://doi.org/10.1037/met0000620

Noreika, V. (2023). Noreika_DATA1 (Version 1) [Data set]. figshare. https://doi.org/10.6084/m9.figshare.24058740.v1

Noreika, V., Valli, K., Lahtela, H., & Revonsuo, A. (2009). Early-night serial awakenings as a new paradigm for studies on NREM dreaming. International Journal of Psychophysiology, 74(1), 14–18. https://doi.org/10.1016/j.ijpsycho.2009.06.002

Noreika, V., Windt, J. M., Kern, M., Valli, K., Salonen, T., Parkkola, R., Revonsuo, A., Karim, A. A., Ball, T., & Lenggenhager, B. (2020). Modulating dream experience: Noninvasive brain stimulation over the sensorimotor cortex reduces dream movement. Scientific Reports, 10(1), 6735. https://doi.org/10.1038/s41598-020-63479-6

Palhano-Fontes, F., Barreto, D., Onias, H., Andrade, K. C., Novaes, M. M., Pessoa, J. A., Mota-Rolim, S. A., Osório, F. L., Sanches, R., dos Santos, R. G., Tófoli, L. F., de Oliveira Silveira, G., Yonamine, M., Riba, J., Santos, F. R., Silva-Junior, A. A., Alchieri, J. C., Galvão-Coelho, N. L., Lobão-Soares, B., … Araújo, D. B. (2019). Rapid antidepressant effects of the psychedelic ayahuasca in treatment-resistant depression: a randomized placebo-controlled trial. Psychological Medicine, 49(4), 655–663. https://doi.org/10.1017/S0033291718001356

Reda, F., Gorgoni, M., D'Atri, A., Scarpelli, S., Carpi, M., Di Cola, E., Menghini, D., Vicari, S., Stella, G., & De Gennaro, L. (2021). Sleep-related declarative memory consolidation in children and adolescents with developmental dyslexia. Brain Sciences, 11(1), 73. https://doi.org/10.3390/brainsci11010073

Scarpelli, S., Bartolacci, C., D'Atri, A., Camaioni, M., Annarumma, L., Gorgoni, M., Cloos, C., Ferrara, M., & De Gennaro, L. (2020). Electrophysiological correlates of dream recall during REM sleep: Evidence from multiple awakenings and within-subjects design. Nature and Science of Sleep, 12, 1043–1052. https://doi.org/10.2147/NSS.S279786

Scarpelli, S., D'Atri, A., Bartolacci, C., Gorgoni, M., Mangiaruga, A., Ferrara, M., & De Gennaro, L. (2020). Dream recall upon awakening from non-rapid eye movement sleep in older adults: Electrophysiological pattern and qualitative features. Brain Sciences, 10(6), 343. https://doi.org/10.3390/brainsci10060343

Scarpelli, S., D'Atri, A., Bartolacci, C., Mangiaruga, A., Gorgoni, M., & De Gennaro, L. (2019). Oscillatory EEG activity during REM sleep in elderly people predicts subsequent dream recall after awakenings. Frontiers in Neurology, 10, 985. https://doi.org/10.3389/fneur.2019.00985

Schoch, S. F., Cordi, M. J., Schredl, M., & Rasch, B. (2019). The effect of dream report collection and dream incorporation on memory consolidation during sleep. Journal of Sleep Research, 28(1), e12754. https://doi.org/10.1111/jsr.12754

Siclari, F., Baird, B., Perogamvros, L., Bernardi, G., LaRocque, J. J., Riedner, B., Boly, M., Postle, B. R., & Tononi, G. (2017). The neural correlates of dreaming. Nature Neuroscience, 20(6), 872–878. https://doi.org/10.1038/nn.4545

Siclari, F., Bernardi, G., Cataldi, J., & Tononi, G. (2018). Dreaming in NREM sleep: A high-density EEG study of slow waves and spindles. Journal of Neuroscience, 38(43), 9175–9185. https://doi.org/10.1523/JNEUROSCI.0855-18.2018

Sikka, P., Revonsuo, A., Noreika, V., & Valli, K. (2019). EEG frontal alpha asymmetry and dream affect: Alpha oscillations over the right frontal cortex during REM sleep and presleep wakefulness predict anger in REM sleep dreams. Journal of Neuroscience, 39(24), 4775–4784. https://doi.org/10.1523/JNEUROSCI.2884-18.2019

Stephan, A. M., Lecci, S., Cataldi, J., & Siclari, F. (2021). Conscious experiences and high-density EEG patterns predicting subjective sleep depth. Current Biology, 31, 5487–5500. https://doi.org/10.1016/j.cub.2021.10.012

Webb, M. D. (2023). Reworking wild bootstrap-based inference for clustered errors. Canadian Journal of Economics/Revue canadienne d'économique, 56(3), 839–858. https://doi.org/10.1111/caje.12661

Wong, W., Andrillon, T., Decat, N., Herzog, R., Noreika, V., Valli, K., Windt, J., & Tsuchiya, N. (2025b). The DREAM database (Version 9) [Data set]. Monash University. https://doi.org/10.26180/22133105.v9

Wong, W., Herzog, R., Andrade, K. C., Andrillon, T., de Araujo, D. B., Arnulf, I., Ataei, S., Avvenuti, G., Baird, B., Bellesi, M., Bergamo, D., Bernardi, G., Blagrove, M., Decat, N., Demirel, Ç., Dresler, M., Eichenlaub, J.-B., Elce, V., Gais, S., … Tsuchiya, N. (2025a). A dream EEG and mentation database. Nature Communications, 16, 7495. https://doi.org/10.1038/s41467-025-61945-1

Wong, W., Noreika, V., Móró, L., Revonsuo, A., Windt, J., Valli, K., & Tsuchiya, N. (2020). The Dream Catcher experiment: Blinded analyses failed to detect markers of dreaming consciousness in EEG spectral power. Neuroscience of Consciousness, 2020(1), niaa006. https://doi.org/10.1093/nc/niaa006

Zhang, J., & Wamsley, E. J. (2019). EEG predictors of dreaming outside of REM sleep. Psychophysiology, 56(7), e13368. https://doi.org/10.1111/psyp.13368