# AI Review Report

Paper: Searching for Sleep: What Digital Trace Data Reveals About Infant Sleep Difficulty

Paper ID: 5

Review Service: Reviewer3.com

Generated: May 09, 2026 at 05:26 UTC

---

## Round 1 - Completed

*Paper version: v5*

*Submitted: March 17, 2026 at 13:44 UTC*

*Completed: March 17, 2026 at 13:49 UTC*

### reviewer1

The manuscript frequently conflates the proxy measure (parental search volume) with the clinical outcome (infant sleep difficulty). For example, the abstract asks, "When do babies sleep worst?" and the text assumes that search volume reflects "actual sleep disruption." However, search volume is a proxy for *parental concern and information-seeking*, which can be influenced by parental fatigue, changes in parental leave, or shifting expectations. The Daylight Saving Time (DST) validation perfectly illustrates this confound: DST disrupts the circadian rhythms of the parents as well as the infants. Increased searches for "baby sleep" during DST may simply reflect parents having less tolerance for normal infant wakefulness due to their own sleep deprivation. The authors must revise the causal language throughout the manuscript to strictly frame the findings around parental concern rather than objective infant sleep quality.

### reviewer1

The DST natural experiment utilizes a highly restricted and arbitrary panel of only five DST-observing states (CA, TX, NY, FL, IL) and two non-observing states (AZ, HI). Selecting only the most populous states for the treatment group introduces severe demographic, cultural, and geographic confounding variables that perfectly overlap with the DST condition. Because Google Trends data is available for all US states, the authors should include all 48 DST-observing states to form a robust, representative treatment group, or provide a rigorous methodological justification for why this specific convenience sample was chosen.

### reviewer1

There is a direct discrepancy between the data modeled in the text and the data visualized in Figure 3. The text states that the difference-in-differences (DID) regression relies on a panel of 7 states. However, Figure 3 only plots the event-time data for California, Texas, and Arizona, completely omitting New York, Florida, Illinois, and Hawaii. The figure must be updated to represent all states included in the quantitative DID model, or supplementary panels should be provided, as selectively plotting only a subset of the analyzed states obscures the underlying data.

### reviewer1

The authors hypothesize that the 18-month peak is driven by a convergence of specific developmental mechanisms: separation anxiety, nap transitions, and language development. To provide mechanistic support for this descriptive observation, I recommend analyzing the "related queries" for the 18-month search terms, exactly as was done for the 4-month terms. If these developmental milestones are the underlying drivers of the search spike, terms related to "naps", "schedule", or "separation" should be highly enriched in the 18-month related queries. This addition would bridge the gap between the observed search volume and the proposed behavioral mechanisms.

### reviewer1

The leave-one-out (LOO) anomaly detection relies on the assumption that all search term families follow a valid exponential decay baseline. While Figure 1 supports this functional form for the broad search terms, Figure 2A shows

that the experiential term ("K month old not sleeping", green line) is highly erratic, with massive spikes at 2, 6, and 14 months. The authors should statistically justify the use of an exponential decay baseline for this specific term family. If the experiential term does not fundamentally follow an exponential decay, applying this baseline will produce spurious mathematical outliers that do not reflect true deviations in parental behavior.

**reviewer2**

**Small-Cluster Bias in the Difference-in-Differences (DiD) Analysis:** The DST validation relies on a DiD framework with state-clustered standard errors across only 7 states (5 treated, 2 control). It is well-documented in the econometrics literature that cluster-robust standard errors are severely downward biased when the number of clusters is small (typically under 30-40), which inflates the Type I error rate. Given that the reported p-value is already marginal ($p = 0.085$), the true p-value is likely higher. The authors should apply a small-sample correction, such as a wild cluster bootstrap, to calculate the p-value and confidence intervals. Furthermore, the language claiming that DST "validates the search signal" should be tempered to reflect the statistical limitations and marginal significance of this specific analysis.

**reviewer2**

**Multiple Testing in Leave-One-Out (LOO) Analysis:** The study evaluates 22 distinct ages (months 3 through 24) to identify significant deviations from the exponential decay baseline. Because the authors are testing each age for a potential anomaly, this constitutes multiple hypothesis testing. The authors should apply a multiple testing correction (e.g., Bonferroni or Benjamini-Hochberg FDR) to the reported p-values (such as the $p < 0.001$ reported for 18 months) to ensure the false positive rate is appropriately controlled across the family of tests.

**reviewer2**

**Normality Assumptions for Z-Scores:** The LOO analysis computes z-scores and corresponding p-values from the residuals of the baseline fit. However, this distribution of residuals contains only 22 data points. Relying on asymptotic normality (z-scores) for such a small sample is statistically risky. The authors should report a normality test (e.g., Shapiro-Wilk) on the LOO residuals to justify this approach. If the residuals are not normally distributed, a non-parametric approach or a t-distribution should be utilized to determine the significance of the deviations.

**reviewer2**

**Retransformation Bias in Log-Linear OLS:** The primary exponential decay model is fit via OLS on log-transformed search volumes. If the percentage deviations and residuals used for the LOO analysis are calculated by comparing the raw observed volume to the exponentiated OLS predictions, this introduces retransformation bias (due to Jensen's inequality), which systematically underestimates the expected raw volume. The authors should clarify if a correction (such as Duan's smearing estimator) was applied when transforming predictions back to the linear scale. Alternatively, the authors could elevate the Nonlinear Least Squares (NLS) specification?which avoids this bias entirely?to be the primary model used for the main text and Figure 1.

**reviewer2**

**Clarification of the DiD Event Window:** The DiD analysis uses a panel of +/- 4 weeks around the DST event. It is unclear from the text whether the post-treatment indicator is coded as a step function (1 for all 4 weeks post-DST) or if it isolates the acute shock. Because circadian disruptions from DST typically resolve within a few days to a week, averaging the effect over a full 4-week post-period could severely dilute the estimated causal effect. The authors should explicitly define how the post-treatment variable is parameterized and justify the 4-week window in the context of acute sleep disruption.

**reviewer3**

The abstract and conclusion strongly assert that 18 months is the "single age" and "one dramatic exception" of elevated search activity. However, Section 3.2 reveals that for the experiential distress term ("K month old not sleeping"), 6 months is actually the largest outlier ($z = 3.44$). The manuscript largely glosses over this finding in the discussion. The authors must reconcile this discrepancy, as the presence of a distinct, highly significant peak at 6 months for the pure distress term directly challenges the core conclusion that 18 months is the sole period of elevated parental sleep concern.

### reviewer3

The use of an exponential decay model to establish the baseline of normal sleep concern lacks theoretical justification in the text. While parental concern generally decreases as infants mature, the choice of an exponential function mathematically dictates the size of the residuals used to identify outliers. The authors should justify why an exponential decay is the appropriate null model for infant sleep development and discuss how sensitive the outlier detection (and the resulting z-scores) is to this specific functional form compared to alternative baseline models.

### reviewer3

The manuscript acknowledges a vocabulary shift where parents increasingly use the term "toddler" instead of "18 month old," suggesting this means the 18-month peak is undercounted. However, if the broader search volume for "K month old" drops systematically after 12 months due to this vocabulary shift, the exponential decay baseline may be artificially steepened by linguistic habits rather than a true decline in sleep concern. The authors should address how this vocabulary shift impacts the validity of the baseline curve itself, as an artificially steep baseline would inflate the relative magnitude of any subsequent peaks.

### reviewer3

The authors contrast their approach with app-based studies, claiming Google Trends provides "massive implicit sample sizes reflecting the behavior of millions of parents with no recruitment-based selection bias." This overlooks inherent selection biases in search engine usage. Information-seeking behavior via Google (versus social media platforms, specialized apps, or direct medical consultation) likely varies significantly by parent age, socioeconomic status, and digital literacy. The authors should temper the claim of lacking selection bias and explicitly discuss the demographic limitations of relying solely on Google Trends data.

### reviewer3

In Section 3.3, the authors attribute the 18-month peak to converging developmental milestones such as separation anxiety, nap transitions, and language development. To properly situate this finding within the literature, the manuscript should discuss whether existing clinical, polysomnography, or actigraphy-based literature corroborates a measurable peak in objective sleep disruptions at 18 months, or if the current literature suggests this phenomenon is strictly limited to parental perception and search behavior.

---

## Round 2 - Approved

*Paper version: v6*
*Submitted: March 17, 2026 at 20:51 UTC*
*Completed: March 17, 2026 at 20:51 UTC*

### Comment 1

**Score: 4/4 - Fully addressed**

**Original comment:**

The manuscript frequently conflates the proxy measure (parental search volume) with the clinical outcome (infant sleep difficulty). For example, the abstract asks, "When do babies sleep worst?" and the text assumes that search volume reflects "actual sleep disruption." However, search volume is a proxy for *parental concern and information-seeking*, which can be influenced by parental fatigue, changes in parental leave, or shifting expectations. The Daylight Saving Time (DST) validation perfectly illustrates this confound: DST disrupts the circadian rhythms of the parents as well as the infants. Increased searches for "baby sleep" during DST may simply reflect parents having less tolerance for normal infant wakefulness due to their own sleep deprivation. The authors must revise the causal language throughout the manuscript to strictly frame the findings around parental concern rather than objective infant sleep quality.

**Author response:**

We agree this was a significant framing problem.
We have revised the language

throughout the manuscript:
? The title now reads ?Parental Sleep Concern in Infancy? rather than ?Infant Sleep Difficulty.?
? The abstract asks ?When are parents most concerned?? rather than ?When do babies sleep worst??
? A new paragraph in the Introduction explicitly states: ?search volume measures parental concern and information-seeking, not objective infant sleep quality.?
? The DST section now acknowledges that DST disrupts parental sleep as well, so increased searches may reflect reduced parental tolerance rather than (only) increased infant disruption. We relabel the DST section ?Suggestive Evidence for Search Signal Validity.?
? The Discussion emphasizes that the 18-month peak could reflect genuinely worse sleep, a content gap, or heightened parental uncertainty, and that search data alone cannot distinguish these.

**Reviewer assessment:**

The author has comprehensively addressed this crucial conceptual point. The manuscript's title, abstract, introduction, and discussion have all been systematically revised to precisely frame the findings in terms of "parental concern" rather than objective sleep quality. A new paragraph was added to the introduction to explicitly state this distinction, and the DST section now acknowledges the parental confound. These changes fully resolve the initial concern about conflating the proxy measure with the outcome.

## Comment 2

**Score: 3/4 - Well acknowledged**

**Original comment:**

The DST natural experiment utilizes a highly restricted and arbitrary panel of only five DST-observing states (CA, TX, NY, FL, IL) and two non-observing states (AZ, HI). Selecting only the most populous states for the treatment group introduces severe demographic, cultural, and geographic confounding variables that perfectly overlap with the DST condition. Because Google Trends data is available for all US states, the authors should include all 48 DST-observing states to form a robust, representative treatment group, or provide a rigorous methodological justification for why this specific convenience sample was chosen.

**Author response:**

We now provide an explicit justification for the state selection: Google Trends returns unreliable (zero or highly volatile) weekly data for less populous states. We selected the most populous states in each DST category to ensure data quality. We acknowledge this convenience sample as a limitation and note that future work should expand the panel.

**Reviewer assessment:**

The author did not expand the state panel as suggested, but they have now provided a clear methodological justification for using a convenience sample of populous states, citing data quality issues in smaller states. They have also added a sentence to the Data section explicitly acknowledging this as a limitation. While not the ideal solution, this is a reasonable and transparent response to the reviewer's concern.

## Comment 3

**Score: 3/4 - Well acknowledged**

**Original comment:**

There is a direct discrepancy between the data modeled in the text and the data visualized in Figure 3. The text states that the difference-in-differences (DID) regression relies on a panel of 7 states. However, Figure 3 only plots the event-time data for California, Texas, and Arizona, completely omitting New York, Florida, Illinois, and Hawaii. The figure must be updated to represent all states included in the quantitative DID model, or supplementary panels should be provided, as selectively plotting only a subset of the analyzed states obscures the underlying data.

**Author response:**

We now acknowledge this directly in the figure caption, which states: ?for three originally collected states (CA, TX, AZ). The DID regression in the text uses the expanded 7-state

panel (adding NY, FL, IL, HI).? The event-time plot predates the expansion of the DID panel. The regression coefficients are estimated on the full 7-state panel; the figure shows the subset for which the original visualization was produced. All state-level data is available in the code repository.

**Reviewer assessment:**

The author has addressed the discrepancy by adding a clarification to the caption of Figure 3. The caption now explicitly states that the plot shows a subset of the data, while the regression analysis uses the full 7-state panel. While updating the figure would have been preferable, this textual clarification resolves the potential for misinterpretation and makes the manuscript more transparent.

## Comment 4

**Score: 4/4 - Fully addressed**

**Original comment:**

The authors hypothesize that the 18-month peak is driven by a convergence of specific developmental mechanisms: separation anxiety, nap transitions, and language development. To provide mechanistic support for this descriptive observation, I recommend analyzing the "related queries" for the 18-month search terms, exactly as was done for the 4-month terms. If these developmental milestones are the underlying drivers of the search spike, terms related to "naps", "schedule", or "separation" should be highly enriched in the 18-month related queries. This addition would bridge the gap between the observed search volume and the proposed behavioral mechanisms.

**Author response:**

Done. We collected related queries for both ?18 month old sleep? (broad) and ?18 month sleep? (ksleep). The results are now reported in Section 3.3:
? For ?18 month old sleep?: the top related queries are ?18 month old sleep training? and ?how much should an 18 month old sleep,? reflecting practical guidance-seeking rather than regression content.
? For ?18 month sleep?: 67% of top queries mention ?regression,? consistent with the content ecosystem having co-opted this age, but the signal persists in the regression-free phrasing.
This supports the interpretation that 18-month searches reflect genuine parental concern rather than content browsing.

**Reviewer assessment:**

The author has fully addressed this comment by performing the requested analysis. They collected and analyzed the "related queries" for the 18-month search terms and integrated the findings into Section 3.3. This new evidence provides valuable mechanistic support for their interpretation of the 18-month peak and substantially strengthens the paper.

## Comment 5

**Score: 4/4 - Fully addressed**

**Original comment:**

The leave-one-out (LOO) anomaly detection relies on the assumption that all search term families follow a valid exponential decay baseline. While Figure 1 supports this functional form for the broad search terms, Figure 2A shows that the experiential term ("K month old not sleeping", green line) is highly erratic, with massive spikes at 2, 6, and 14 months. The authors should statistically justify the use of an exponential decay baseline for this specific term family. If the experiential term does not fundamentally follow an exponential decay, applying this baseline will produce spurious mathematical outliers that do not reflect true deviations in parental behavior.

**Author response:**

We agree and now address this directly in a new paragraph (?The 6-month experiential peak,? Section 3.2). We note that the experiential terms have much lower search volume, making them substantially noisier; several ages show large swings. We explicitly state that ?an exponential decay baseline may not be the right null model for these terms? and that we ?report it transparently but do not build our main conclusions on the experiential family.?

**Reviewer assessment:**

The author has fully addressed this concern by adding a new paragraph in Section 3.2 that explicitly discusses the

limitations of applying the exponential decay model to the noisy experiential search terms. They correctly note that the baseline may not be appropriate for this term family and state that they do not build their main conclusions on this part of the analysis. This is a transparent and scientifically sound way to handle the issue.

## Comment 6

**Score: 4/4 - Fully addressed**

**Original comment:**

**Small-Cluster Bias in the Difference-in-Differences (DiD) Analysis:** The DST validation relies on a DiD framework with state-clustered standard errors across only 7 states (5 treated, 2 control). It is well-documented in the econometrics literature that cluster-robust standard errors are severely downward biased when the number of clusters is small (typically under 30-40), which inflates the Type I error rate. Given that the reported p-value is already marginal (p = 0.085), the true p-value is likely higher. The authors should apply a small-sample correction, such as a wild cluster bootstrap, to calculate the p-value and confidence intervals. Furthermore, the language claiming that DST "validates the search signal" should be tempered to reflect the statistical limitations and marginal significance of this specific analysis.

**Author response:**

We now apply the Cameron, Gelbach, and Miller (2008) correction, using the t(G?2) =
t(5) distribution rather than the normal distribution for inference. This yields p = 0.094 (vs. 0.085
with normal approximation) and wider CIs [?2.6, 23.5]. We cite Cameron et al. (2008) and relabel
the DST result as ?suggestive? throughout. We also temper the language from ?validates? to
?provides suggestive evidence.?

**Reviewer assessment:**

The author has fully addressed this important statistical point. They correctly applied a small-sample correction for the clustered standard errors by using a t-distribution, which is an appropriate response. They have reported the updated (and more conservative) p-value and confidence intervals, and have also tempered the language throughout the manuscript to describe the evidence as "suggestive." This represents a significant methodological improvement.

## Comment 7

**Score: 4/4 - Fully addressed**

**Original comment:**

**Multiple Testing in Leave-One-Out (LOO) Analysis:** The study evaluates 22 distinct ages (months 3 through 24) to identify significant deviations from the exponential decay baseline. Because the authors are testing each age for a potential anomaly, this constitutes multiple hypothesis testing. The authors should apply a multiple testing correction (e.g., Bonferroni or Benjamini-Hochberg FDR) to the reported p-values (such as the $p < 0.001$ reported for 18 months) to ensure the false positive rate is appropriately controlled across the family of tests.

**Author response:**

Done. We now apply Benjamini-Hochberg FDR correction across all 22 tested ages.
After correction, 18 months remains significant (padj = 0.027). No other age is significant. This is
reported in the abstract and Section 3.1.

**Reviewer assessment:**

The author has fully addressed this comment by applying the recommended Benjamini-Hochberg FDR correction to account for multiple hypothesis testing. The revised manuscript now reports the adjusted p-value in both the abstract and the main text (Section 3.1), confirming that the key finding for 18 months remains significant. This correction makes the statistical analysis more rigorous.

## Comment 8

**Score: 4/4 - Fully addressed**

**Original comment:**

**Normality Assumptions for Z-Scores:** The LOO analysis computes z-scores and corresponding p-values from the residuals of the baseline fit. However, this distribution of residuals contains only 22 data points. Relying on asymptotic normality (z-scores) for such a small sample is statistically risky. The authors should report a normality test (e.g.,

Shapiro-Wilk) on the LOO residuals to justify this approach. If the residuals are not normally distributed, a non-parametric approach or a t-distribution should be utilized to determine the significance of the deviations.

**Author response:**

Done.
A Shapiro-Wilk test rejects normality of the LOO residuals ($W = 0.74$, $p < 0.001$). We therefore compute p-values using the t-distribution with df = 21 rather than the normal distribution. This is reported in Section 3.1.

**Reviewer assessment:**

The author has fully addressed this statistical concern. As requested, they performed a Shapiro-Wilk test, found that the residuals were not normally distributed, and correctly switched from using a normal distribution (z-scores) to a t-distribution for calculating p-values. This change, reported in Section 3.1, makes the inference more appropriate for the small sample size and non-normal data.

## Comment 9

**Score: 4/4 - Fully addressed**

**Original comment:**

**Retransformation Bias in Log-Linear OLS:** The primary exponential decay model is fit via OLS on log-transformed search volumes. If the percentage deviations and residuals used for the LOO analysis are calculated by comparing the raw observed volume to the exponentiated OLS predictions, this introduces retransformation bias (due to Jensen's inequality), which systematically underestimates the expected raw volume. The authors should clarify if a correction (such as Duan's smearing estimator) was applied when transforming predictions back to the linear scale. Alternatively, the authors could elevate the Nonlinear Least Squares (NLS) specification?which avoids this bias entirely?to be the primary model used for the main text and Figure 1.

**Author response:**

Done. NLS is now the primary specification throughout the paper (Section 3.1, Figure 1, abstract). We explain the rationale: NLS avoids the retransformation bias inherent in OLS-on-log when predictions are transformed back to the original scale. The OLS-on-log specification is retained in the robustness table (Table 1) for comparison. We explicitly note that we do not apply Duan?s smearing estimator to the OLS specification.

**Reviewer assessment:**

The author has fully addressed this comment by elevating the Nonlinear Least Squares (NLS) specification to be the primary model, as suggested. This change, reflected in the abstract, Section 3.1, Figure 1, and Table 1, effectively avoids the retransformation bias issue. The author also added text explaining the rationale for this choice, which improves the methodological transparency of the paper.

## Comment 10

**Score: 4/4 - Fully addressed**

**Original comment:**

**Clarification of the DiD Event Window:** The DiD analysis uses a panel of +/- 4 weeks around the DST event. It is unclear from the text whether the post-treatment indicator is coded as a step function (1 for all 4 weeks post-DST) or if it isolates the acute shock. Because circadian disruptions from DST typically resolve within a few days to a week, averaging the effect over a full 4-week post-period could severely dilute the estimated causal effect. The authors should explicitly define how the post-treatment variable is parameterized and justify the 4-week window in the context of acute sleep disruption.

**Author response:**

We now explicitly state that PostDSTt = 1 for all weeks at or after the DST date within the ±4-week window (i.e., a step function). We acknowledge that this is conservative: ?because circadian disruptions typically resolve within days, the acute effect is likely larger but diluted by averaging over the full post-window.?

**Reviewer assessment:**

The author has fully addressed this comment by clarifying the parameterization of the post-treatment variable in Section 3.4. They now explicitly state that it is a step function over the four-week window. Furthermore, they thoughtfully acknowledge that this approach is conservative and likely dilutes the acute effect, which adds important context for the reader.

## Comment 11

**Score: 4/4 - Fully addressed**

**Original comment:**

The abstract and conclusion strongly assert that 18 months is the "single age" and "one dramatic exception" of elevated search activity. However, Section 3.2 reveals that for the experiential distress term ("K month old not sleeping"), 6 months is actually the largest outlier (z = 3.44). The manuscript largely glosses over this finding in the discussion. The authors must reconcile this discrepancy, as the presence of a distinct, highly significant peak at 6 months for the pure distress term directly challenges the core conclusion that 18 months is the sole period of elevated parental sleep concern.

**Author response:**

We now discuss this explicitly in a new paragraph (?The 6-month experiential peak,? Section 3.2). We interpret the result cautiously: the experiential terms have low volume and are noisy, the exponential baseline may not be appropriate for them, and the 6-month signal could reflect genuine distress (teething, solid food introduction) or noise. We do not claim 18 months is the sole peak across all term families; rather, it is the sole significant outlier in the broad term family that our main analysis is built on.

**Reviewer assessment:**

The author has fully addressed this concern. A new paragraph has been added in Section 3.2 to explicitly discuss the 6-month peak found in the experiential search terms. The author provides a cautious interpretation, noting the data's noisiness and the potential inappropriateness of the baseline model for this specific term family. They correctly clarify that their main conclusion about the 18-month peak applies specifically to the primary (broad) search term analysis, thus reconciling the apparent contradiction.

## Comment 12

**Score: 4/4 - Fully addressed**

**Original comment:**

The use of an exponential decay model to establish the baseline of normal sleep concern lacks theoretical justification in the text. While parental concern generally decreases as infants mature, the choice of an exponential function mathematically dictates the size of the residuals used to identify outliers. The authors should justify why an exponential decay is the appropriate null model for infant sleep development and discuss how sensitive the outlier detection (and the resulting z-scores) is to this specific functional form compared to alternative baseline models.

**Author response:**

We now explicitly state (Section 3.1): ?We choose the exponential as a parsimonious two-parameter model for monotone decline; it is not derived from a theory of sleep development.? We note that the robustness table shows results are qualitatively unchanged under both NLS and OLS-on-log fitting, and with/without ages 1-2, which involve different effective functional forms. The 18-month outlier is robust across these specifications.

**Reviewer assessment:**

The author has fully addressed this comment by adding a sentence in Section 3.1 that clarifies the rationale for their model choice. They now state that the exponential function was chosen for its parsimony in modeling monotone decline, rather than being derived from a specific theory. This clarification on the model's atheoretical nature is sufficient and transparent.

## Comment 13

**Score: 4/4 - Fully addressed**

**Original comment:**

The manuscript acknowledges a vocabulary shift where parents increasingly use the term "toddler" instead of "18 month old," suggesting this means the 18-month peak is undercounted. However, if the broader search volume for "K month old" drops systematically after 12 months due to this vocabulary shift, the exponential decay baseline may be artificially steepened by linguistic habits rather than a true decline in sleep concern. The authors should address how this vocabulary shift impacts the validity of the baseline curve itself, as an artificially steep baseline would inflate the relative magnitude of any subsequent peaks.

**Author response:**

We now address this directly in a new paragraph (Section 3.5, ?Vocabulary shift and baseline validity?). We acknowledge the concern: if second-year volumes are undercounted, the decay rate is overestimated and 18-month predictions are too low. However, the direct (unchained) validation, which compares 18-month volume to other ages within the same query (not relative to a fitted baseline), still shows +198% elevation. This suggests the signal is not purely a baseline artifact, though the magnitude is uncertain.

**Reviewer assessment:**

The author has fully addressed this nuanced methodological point. A new paragraph in Section 3.5 now explicitly discusses the possibility that a vocabulary shift could artificially steepen the baseline. The author thoughtfully counters this concern by pointing to the direct (unchained) validation results, which are not dependent on the fitted baseline and still show a large peak at 18 months. This is an excellent response that uses other evidence from the paper to address a potential confound.

## Comment 14

**Score: 4/4 - Fully addressed**

**Original comment:**

The authors contrast their approach with app-based studies, claiming Google Trends provides "massive implicit sample sizes reflecting the behavior of millions of parents with no recruitment-based selection bias." This overlooks inherent selection biases in search engine usage. Information-seeking behavior via Google (versus social media platforms, specialized apps, or direct medical consultation) likely varies significantly by parent age, socioeconomic status, and digital literacy. The authors should temper the claim of lacking selection bias and explicitly discuss the demographic limitations of relying solely on Google Trends data.

**Author response:**

We now acknowledge this explicitly in the Introduction: ?Google search data carries its own selection biases: information-seeking via search engines varies by parent age, socioeconomic status, and digital literacy, and the population of parents who search Google may differ systematically from those who consult pediatricians, use specialized apps, or seek advice on social media.? In the Discussion, we state that ?neither data source [GT nor app-based tracking] is representative on its own.?

**Reviewer assessment:**

The author has fully addressed this comment by removing the overly strong claim and adding explicit acknowledgements of the selection biases inherent in Google Trends data. New text in both the Introduction and Discussion sections now clearly states the limitations related to user demographics and information-seeking preferences, providing a more balanced and accurate description of the data source.

## Comment 15

**Score: 4/4 - Fully addressed**

**Original comment:**

In Section 3.3, the authors attribute the 18-month peak to converging developmental milestones such as separation anxiety, nap transitions, and language development. To properly situate this finding within the literature, the manuscript should discuss whether existing clinical, polysomnography, or actigraphy-based literature corroborates a measurable

peak in objective sleep disruptions at 18 months, or if the current literature suggests this phenomenon is strictly limited to parental perception and search behavior.

**Author response:**

We now discuss this explicitly in Section 3.3. We state: ?To our knowledge, no existing polysomnography or actigraphy study has documented a specific, discrete peak in objective sleep disruption at 18 months. Longitudinal studies show that night wakings generally decrease through the second year (Sadeh et al., 2009), and the clinical literature on ?sleep regressions? at specific ages remains thin. The 18-month peak in our data may reflect a gap between parental expectations and developmental reality rather than an objective nadir in sleep quality.?

**Reviewer assessment:**

The author has fully addressed this comment by adding a new paragraph to Section 3.3 that situates their findings within the existing clinical literature. They now explicitly state that, to their knowledge, no objective sleep studies have documented a similar peak at 18 months. This important addition provides crucial context and correctly frames their finding as a potential discrepancy between parental concern and objective sleep measures, strengthening the paper's contribution.