Psychological Assessment

Peak-End Bias in Retrospective Recall of Depressive Symptoms on the PHQ-9

Adam G. Horwitz, Zhuo Zhao, and Srijan Sen

Online First Publication, February 9, 2023. https://dx.doi.org/10.1037/pas0001219

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Assessment

ISSN: 1040-3590

https://doi.org/10.1037/pas0001219

BRIEF REPORT

Peak-End Bias in Retrospective Recall of Depressive Symptoms on the PHQ-9

Adam G. Horwitz[1], Zhuo Zhao[2], and Srijan Sen[1, 2]

[1] Department of Psychiatry, University of Michigan Medical School

[2] Michigan Neuroscience Institute, The University of Michigan Medical School

Mental health care is built around patient recall and report of clinical symptoms. However, memories of events and experiences rely on cognitive heuristics that influence our recall. The peak-end bias, which refers to the tendency for the most intense and proximate aspects of an experience to disproportionately influence our memory, has been understudied in the context of mental health symptoms and may unduly influence self-reported symptoms, even in the context of standardized assessments. To determine whether the peak-end bias applies to the report of depressive symptoms on the standardized Patient Health Questionnaire–9 (PHQ-9) assessment, we compared two scores from daily mood assessments collected over a 2-week period from 4,322 medical interns (56% women; 60% non-Hispanic White). The peak-end-mood score, which averaged the single lowest and most recent mood scores over 2 weeks had a significantly stronger correlation with the PHQ-9 than the mean-mood score, which averaged all mood scores during the 2 weeks. Likelihood ratio tests and fit statistics provided further support that the peak-end-mood score was a significantly better predictor of depression than the mean-mood score. Results were consistent when limiting the sample to those with mild-to-severe depressive symptoms, and when only examining the two primary mood items as the dependent variable. These findings provide evidence for a modest peak-end recall bias for mood and depressive symptoms. There may be benefits to implementing intermittent assessment strategies to support clinical decision-making.

Public Significance Statement

The findings from this study suggest that there is a systematic bias in the recall of depressive symptoms that overemphasizes the peak (worst) and end (current) mood states. Implementing brief, intermittent assessments may be a useful tool for overcoming this peak-end bias and providing a more accurate picture of between-session symptomatology for mental health care providers and systems applying measurement-based care practices.

Keywords: peak-end bias, depression, mood, intensive longitudinal assessment, medical interns

For mental health disorders, much of clinical care is built around the recall and report of clinical symptoms. For example, when an individual with major depression returns to their provider 6 weeks after starting a new medication, the person's report of changes in depressive symptoms typically drives the decision to maintain, adjust, or stop the dose of the medication. To capture more accurate reports of clinical symptoms, there have been significant efforts to institute measurement-based care (MBC) practices in behavioral health settings, with standardized clinical symptom assessments used to inform treatment decisions (e.g., Lewis et al., 2019). MBC has demonstrated improvement in clinical costs and outcomes, leading to large implementation efforts to expand MBC by psychologists (e.g., Wright et al., 2020), psychiatrists (e.g., Aboraya et al., 2018), and health care systems like the Veterans Health Administration (e.g., Resnick & Hoff, 2020).

Adam G. Horwitz https://orcid.org/0000-0002-6087-7950

This study was supported by the National Institute of Mental Health (R01 MH101459) to Srijan Sen. Adam G. Horwitz received funding from the National Center for Advancing Translational Sciences (KL2TR002241) and the National Institute of Mental Health (K23MH131761). Funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the article; and decision to submit the article for publication.

Adam G. Horwitz played lead role in writing of original draft, supporting role in formal analysis, investigation and methodology and equal role in conceptualization. Zhuo Zhao played lead role in formal analysis and supporting role in conceptualization, investigation, methodology and writing of original draft. Srijan Sen played lead role in funding acquisition, investigation and methodology, supporting role in writing of original draft and equal role in conceptualization.

This study was not preregistered. For access to data and study materials, please contact the corresponding author.

Correspondence concerning this article should be addressed to Adam G. Horwitz, Department of Psychiatry, University of Michigan Medical School, 4250 Plymouth Road, Ann Arbor, MI 48109, United States. Email: ahor@umich.edu

While MBC has advantages to being used in conjunction with, rather than strictly relying on, verbal reports to inform treatment decisions, standardized measures of clinical symptoms still rely on the retrospective report of symptoms over weeks to months. A large body of work suggests that our memory of events and experiences (autobiographical memory) relies on cognitive shortcuts (heuristics) that systematically bias our recall (e.g., Shiffman et al., 2008). One important heuristic is the peak-end rule, whereby the most intense and proximate aspects of an experience disproportionately influence our memory (Kahneman et al., 1993). The peak-end bias has been well established in the pain literature, where the retrospective recall of pain during a procedure or event, such as a colonoscopy or vaginal delivery, is disproportionately influenced by both the peak level of pain and the level of pain at the conclusion of the event/procedure (e.g., Chajut et al., 2014; Redelmeier et al., 2003).

Despite the established effects in the pain literature, the peak-end bias has not been extensively studied in the context of mental health and emotional states. One team has experimentally demonstrated the peak-end bias in the context of anxiety, whereby participants in one study reported greater overall anxiety when a horror movie clip ended at the most intense scene relative to those who viewed an extended version of the clip ending on a less intense scene (Müller et al., 2019), and participants in a shock paradigm ending in a "high" threat condition reported greater retroactive distress compared to those who ended in a "moderate" threat condition (Müller et al., 2022). With respect to depression, a small study of Japanese undergraduate students, Sato and Kawahara (2011) had participants complete daily mood assessments for 2 weeks and a final assessment to rate their mood as a whole during the preceding 2 weeks, which demonstrated a significant negativity bias driven by the peak and final daily-level scores for anxiety and depression.

Given the centrality of recall to mental health clinical decision-making, it is critical to further our understanding of whether psychiatric symptoms are subject to the peak-end bias. The Patient Health Questionnaire–9 (PHQ-9; Kroenke et al., 2001) is a widely used measure of depression that assesses the frequency of depressive symptoms during the 2 weeks prior to assessment. Yet, while this measure is intended to capture 2 weeks of retrospective data, a study by Aguilera et al. (2015) found that daily mood scores over a 2 week period only correlated to the PHQ-9 when mood scores were restricted to the preceding 1 week (and not when including the full 2 weeks), suggesting a recency bias. Notably, these previous studies comparing daily measures to recall reports have not examined the combined influence of recency and negativity biases. To further this line of research and directly test whether a peak-end bias applies to depressive symptoms, we compared average mood scores assessed daily during the 2 weeks prior to assessment (mean-mood score) to a calculated score combining only the lowest mood score and most recent mood score (peak-end-mood score) over the same 2 weeks for associations with PHQ-9 scores.

Method

Participants, Measures, and Procedure

First-year resident physicians (interns) entering residency programs from 2016 to 2020 at U.S. medical centers were recruited to participate as part of the Intern Health Study (Sen et al., 2010). Participants (N = 4,322; 56% women; 60% non-Hispanic White) downloaded a mobile application that provided daily mood assessment prompts between 5 and 10 p.m. throughout the internship year, "On a scale of 1 (*lowest*) to 10 (*highest*), how was your mood today?" Participants also completed longer survey assessments prior to internship and at the end of the first quarter of the internship year, which included the PHQ-9 (Kroenke et al., 2001), a widely used nine-item scale assessing the frequency of the nine clinical depressive symptoms (scale range: 0–27) over the previous 2 weeks. With regard to depressive symptom severity, PHQ-9 total scores are classified as follows: minimal (0–4), mild (5–9), moderate (10–14), moderately severe (15–19), and severe (20 and above; Kroenke et al., 2001). The study was approved by the institutional review board at the University of Michigan.

Data Analytic Plan

To be included in the analytic sample, participating interns were required to complete the PHQ-9 at the follow-up assessment (offered during the third month of internship) and to have had at least three daily mood ratings submitted in the 14 days preceding the PHQ-9's completion. Retention analyses examined differences between the 4,322 included interns who completed at least three daily mood scores and the 724 interns who were excluded from the analytic sample due to providing fewer than three daily mood scores.[1] There were no differences for sex or age, but interns of Asian descent (80.7%) were significantly less likely to complete the requisite number of daily surveys compared to White (87.4%) and multiracial (87.3%) interns ($\chi^2[7] = 32.55$, $p < .001$). Preinternship scores of depression were also significantly higher among those who did not complete at least three daily mood surveys (PHQ-9 $M[SD]$: 6.28[4.2] versus 5.55[4.1]; $t[5{,}040] = 4.36$, $p < .001$). Mean-mood scores were calculated by averaging all completed daily mood scores during the 2 weeks prior to the PHQ-9 assessment. The peak-end-mood score was calculated by averaging an individual's single worst daily mood score and most recent daily mood score in the 2 weeks prior to the PHQ-9 assessment.

We used Fisher's r-to-z transformation and computed a Steiger's Z test (Steiger, 1980) to assess whether the correlation of the peak-end-mood score with the PHQ-9 was significantly different from the correlation of the mean-mood score with the PHQ-9, accounting for the nonindependence of the correlations. To assess how PHQ-9 scores corresponded to the mean-mood and peak-end-mood scores over the past 2 weeks, we conducted likelihood ratio tests comparing reduced models to full models, with two reduced models (each testing the effect when the independent variable was removed from the model) nested within the same full model (both mean-mood and peak-end-mood scores predicting PHQ-9 scores). We compared the Akaike information criterion (AIC) of model fit and used a difference of at least 10 AIC units as the threshold for demonstrating a significantly better model fit (Burnham & Anderson, 2004). To extend generalizability to clinical samples, we examined these effects for a subset of interns with PHQ-9 scores of 5 or higher, indicating mild-to-severe depressive symptoms (2,190 interns; 50.7% of the sample). We also examined results

[1] Out of concern that only requiring three responses would reduce potential variability between mean-mood and peak-end-mood scores, a sensitivity analysis examined results restricted to those with 7+ responses. Results did not meaningfully differ with this more selective sample, so our study maintained inclusion for three or more responses.

restricted to primary mood symptom (PHQ-2; low or depressed mood, anhedonia) to control for the potential influence of environmental factors captured by the PHQ-9 that might function independently of mood in the context of medical internship (e.g., sleeping difficulty due to shift work, fatigue from long work hours).

Results

Participants included in the study completed an average of 9.3 (SD = 3.7) daily mood surveys during the 2-week assessment period. Sample descriptives for the analytic sample and clinical subsample are described in Table 1. PHQ-9 depressive symptom scores correlated significantly with both the mean-mood ($r = -.405$, $p < .001$) and peak-end-mood scores ($r = -.458$, $p < .001$),[2] and the mean-mood and peak-end-mood scores were significantly correlated with each other ($r = .765$, $p < .001$). The computed Steiger's Z test indicated that peak-end-mood scores had a significantly stronger correlation to PHQ-9 scores than the mean-mood scores ($Z = 5.79$, $p < .001$). In likelihood ratio tests, the peak-end-mood score was a significantly better predictor of PHQ-9 scores than the mean-mood score (284.97 vs. 38.33, $p < .001$) and had a significantly better model fit according to the AIC (23511.9 vs. 23758.6).[3] Findings were consistent, with a better fit for the peak-end model, when restricting the sample to a subset of interns with mild-to-severe depressive symptoms, and when only examining primary mood symptoms as an outcome (see Table 2).

Discussion

This study demonstrated the presence of a peak-end bias with respect to mood and depressive symptoms in a large sample of medical interns. Our primary findings suggest that symptom reports, even when filtered through standardized assessments, are susceptible to the influence of both the worst mood state over the preceding period and the current mood state. In practice, a provider might increase a particular dose for a drug or treatment in response to an elevated symptom scale, despite this score potentially being influenced by a nonrepresentative day and not necessarily reflecting the general mental health status during the interval period. The average depression scores for the full, nonclinical sample of interns were a bit higher than depression scores found in normative samples of adults (e.g., Kocalevent et al., 2013), and the average depression scores for the clinical subsample of interns were slightly less than those found in clinical outpatient and inpatient settings (e.g., Hansson et al., 2009; Sun et al., 2020). While the peak-end bias remained in effect when restricting our sample to those with at least mild or worse depressive symptoms, additional research is needed in clinical contexts to clarify the magnitude of this effect.

These findings have implications for the use of standardized assessment measures in clinical practice and MBC initiatives, particularly with respect to longitudinal monitoring of symptoms. While standardized assessments may assist in screening for a range of symptoms, assist with diagnostics at an intake appointment, and provide a useful endpoint for treatment, caution may be needed with respect to drawing conclusions about symptoms between sessions. Previous studies have demonstrated that even for the 2 week interval of the PHQ-9, correlations with daily mood scores are more reflective of the past week than 2 weeks (Aguilera et al., 2015). While the prospect of regular assessment between sessions may initially seem burdensome, the administration of single items, on a simple scale, delivered through mobile applications or text messages is quite feasible (e.g., Porras-Segovia et al., 2020), and may provide a less biased perspective of mental health functioning between sessions and support clinical decision-making in conjunction with standardized measures.

Table 1
Sample Characteristics

Demographics	Full sample *n* (%)	Clinical subsample *n* (%)
N	4,322 (100)	2,190 (100)
Sex		
Male	1,912 (44.0)	860 (39.3)
Female	2,410 (56.0)	1,330 (60.7)
Race/ethnicity		
White	2,597 (60.2)	1,311 (59.9)
Asian	870 (20.2)	431 (19.7)
African American	193 (4.5)	110 (5.0)
Latino	177 (4.1)	89 (4.1)
Arab/Middle Eastern	64 (1.5)	23 (1.1)
Native American	5 (0.1)	3 (0.1)
Multiracial	399 (9.2)	217 (9.9)
Other	17 (0.4)	6 (0.3)
Age, *M* (*SD*)	27.60 (2.72)	27.67 (2.79)
Clinical	***M* (*SD*)**	***M* (*SD*)**
PHQ-9 total score	5.55 (4.14)	8.76 (3.37)
Mean-mood score	7.48 (1.25)	7.06 (1.29)
Peak-end-mood score	6.42 (1.52)	5.86 (1.51)

Note. PHQ-9 = Patient Health Questionnaire–9.

Table 2
Model Comparisons

Models	Full sample (n = 4,332) LRT[a]	Full sample (n = 4,332) AIC	Clinical subsample (n = 2,190) LRT[a]	Clinical subsample (n = 2,190) AIC
Predicting PHQ-9				
Mean-mood	38.33	23758.6	7.70	11371.6
Peak-end-mood	284.97	23511.9	85.72	11293.6
Predicting PHQ-2				
Mean-mood	60.76	7295.7	26.41	3517.3
Peak-end-mood	274.18	7082.3	85.02	3456.7

Note. LRT = Likelihood ratio test; AIC = Akaike information criterion (lower scores indicate better fit); PHQ-9 = Patient Health Questionnaire–9 total depressive symptom score (range: 0–27); PHQ-2 = Patient Health Questionnaire–2 depressive symptoms score for the first two PHQ-9 primary mood items (i.e., low or depressed mood, anhedonia; range: 0–6). Clinical sample = mild-to-severe symptoms of depression (PHQ-9 total scores ≥5).
[a] LRT signifies the value when the independent variable was left out of the model.

[2] Post hoc analyses examined the "end" and "peak" scores individually to ensure the correlation strength was not being unduly influenced by one of the two items making up the peak-end score. The correlations for end-only ($r = -.396$) and peak-only ($r = .438$) scores had weaker correlations with the PHQ-9 than the combined peak-end-mood score.

[3] When requiring at least seven daily mood responses, AIC values were 17068.7 (peak-end) versus 17123.0 (mean-mood).

While this study has several notable strengths, findings should be understood within the context of its limitations. Even with sample stratification for depression severity, medical interns are educationally and occupationally homogenous, which may limit generalizability. Since our study used a nonclinical sample, our daily mood item was not symptom-focused, which may explain a smaller than expected correlation between daily mood and depressive symptoms. Future studies may wish to test the peak-end bias more directly in clinical samples by specifically assessing depressed mood daily rather than the mood in general. Baseline depression scores were significantly higher among interns who were excluded from analyses due to low survey adherence, and interns of Asian descent were also less likely to complete the daily surveys, which may have contributed to bias within the analytic sample. We did not find that increasing the threshold for number of daily mood responses (i.e., seven or more daily mood scores) to alter patterns associated with peak-end-mood and average-mood models, and so maintained the threshold at three responses to maximize inclusion, though we acknowledge that missingness may contribute to potential measurement error.

Despite these limitations, our findings demonstrate the potential presence of the peak-end bias in the report of depressive symptoms and highlight the need for additional research into this phenomenon, particularly in clinical samples. While the overall effect of the peak-end bias was relatively modest, there may be individual differences based on personality, clinical symptoms, or sociodemographic factors that suggest some individuals may be more prone to this reporting bias than others. Additional research into these nuances would help clarify the clinical significance associated with this peak-end bias in practice. Nevertheless, with the increasing ease of mobile assessments through applications or texting, mental health care providers and systems may be able to improve upon MBC practices by implementing brief, intermittent assessments rather than relying strictly on retrospective recall on the day of appointments to gather a summary of the preceding time period.

References

Aboraya, A., Nasrallah, H. A., Elswick, D. E., Elshazly, A., Estephan, N., Aboraya, D., Berzingi, S., Chumbers, J., Berzingi, S., Justice, J., Zafar, J., & Justice, J. (2018). Measurement-based care in psychiatry: Past, present, and future. *Innovations in Clinical Neuroscience*, *15*(11–12), 13–26. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6380611/

Aguilera, A., Schueller, S. M., & Leykin, Y. (2015). Daily mood ratings via text message as a proxy for clinic based depression assessment. *Journal of Affective Disorders*, *175*, 471–474. https://doi.org/10.1016/j.jad.2015.01.033

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Chajut, E., Caspi, A., Chen, R., Hod, M., & Ariely, D. (2014). In pain thou shalt bring forth children: The peak-and-end rule in recall of labor pain. *Psychological Science*, *25*(12), 2266–2271. https://doi.org/10.1177/0956797614551004

Hansson, M., Chotai, J., Nordstöm, A., & Bodlund, O. (2009). Comparison of two self-rating scales to detect depression: HADS and PHQ-9. *The British Journal of General Practice*, *59*(566), e283–e288. https://doi.org/10.3399/bjgp09X454070

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, *4*(6), 401–405. https://doi.org/10.1111/j.1467-9280.1993.tb00589.x

Kocalevent, R.-D., Hinz, A., & Brähler, E. (2013). Standardization of the depression screener patient health questionnaire (PHQ-9) in the general population. *General Hospital Psychiatry*, *35*(5), 551–555. https://doi.org/10.1016/j.genhosppsych.2013.04.006

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Lewis, C. C., Boyd, M., Puspitasari, A., Navarro, E., Howard, J., Kassab, H., Hoffman, M., Scott, K., Lyon, A., Douglas, S., Simon, G., & Kroenke, K. (2019). Implementing measurement-based care in behavioral health: A review. *JAMA Psychiatry*, *76*(3), 324–335. https://doi.org/10.1001/jamapsychiatry.2018.3329

Müller, U. W. D., Gerdes, A. B. M., & Alpers, G. W. (2022). Time is a great healer: Peak-end memory bias in anxiety—Induced by threat of shock. *Behaviour Research and Therapy*, *159*, Article 104206. https://doi.org/10.1016/j.brat.2022.104206

Müller, U. W. D., Witteman, C. L. M., Spijker, J., & Alpers, G. W. (2019). All's bad that ends bad: There is a peak-end memory bias in anxiety. *Frontiers in Psychology*, *10*, Article 1272. https://doi.org/10.3389/fpsyg.2019.01272

Porras-Segovia, A., Molina-Madueño, R. M., Berrouiguet, S., López-Castroman, J., Barrigón, M. L., Pérez-Rodríguez, M. S., Marco, J. H., Díaz-Oliván, I., de León, S., Courtet, P., Artés-Rodríguez, A., & Baca-García, E. (2020). Smartphone-based ecological momentary assessment (EMA) in psychiatric patients and student controls: A real-world feasibility study. *Journal of Affective Disorders*, *274*, 733–741. https://doi.org/10.1016/j.jad.2020.05.067

Redelmeier, D. A., Katz, J., & Kahneman, D. (2003). Memories of colonoscopy: A randomized trial. *Pain*, *104*(1–2), 187–194. https://doi.org/10.1016/S0304-3959(03)00003-4

Resnick, S. G., & Hoff, R. A. (2020). Observations from the national implementation of Measurement Based Care in Mental Health in the Department of Veterans Affairs. *Psychological Services*, *17*(3), 238–246. https://doi.org/10.1037/ser0000351

Sato, H., & Kawahara, J. (2011). Selective bias in retrospective self-reports of negative mood states. *Anxiety, Stress, and Coping*, *24*(4), 359–367. https://doi.org/10.1080/10615806.2010.543132

Sen, S., Kranzler, H. R., Krystal, J. H., Speller, H., Chan, G., Gelernter, J., & Guille, C. (2010). A prospective cohort study investigating factors associated with depression during medical internship. *Archives of General Psychiatry*, *67*(6), 557–565. https://doi.org/10.1001/archgenpsychiatry.2010.41

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*(1), 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Sun, Y., Fu, Z., Bo, Q., Mao, Z., Ma, X., & Wang, C. (2020). The reliability and validity of PHQ-9 in patients with major depressive disorder in psychiatric hospital. *BMC Psychiatry*, *20*(1), Article 474. https://doi.org/10.1186/s12888-020-02885-6

Wright, C. V., Goodheart, C., Bard, D., Bobbitt, B. L., Butt, Z., Lysell, K., McKay, D., & Stephens, K. (2020). Promoting measurement-based care and quality measure development: The APA mental and behavioral health registry initiative. *Psychological Services*, *17*(3), 262–270. https://doi.org/10.1037/ser0000347

Received September 2, 2022
Revision received December 14, 2022
Accepted January 3, 2023 ■