What were you Thinking? A Comparison of Rater Coding and word Counts for Content Analysis of Thought Samples in Depression

By: Logan Stiles, Aaron Frazier, Kari M. Eddington

Stiles, L., Frazier, A. & Eddington, K.M. What were you Thinking? A Comparison of Rater Coding and word Counts for Content Analysis of Thought Samples in Depression Journal of Rational-Emotive & Cognitive-Behavior Therapy (2023). DOI: 10.1007/s10942-023-00507-0

Abstract:

This study examined the convergence between two methods of thought content analysis, manual coding by trained raters and computer-generated word counts, in a sample of clinically depressed participants assessed before and after treatment. Automated word count programs have traditionally used longer narrative texts so their utility for shorter thought samples is uncertain. Aims were to evaluate their direct correspondence and to determine whether the two methods yield similar results in assessing change from pre- to post-treatment. Thirty participants recorded in-the-moment thoughts during random phone-based signaling. Thought samples were analyzed for presence of negative emotion (NE), positive emotion (PE), and self-focus (SF), using hand coded ratings and automated word counts. Correlations between ratings and word counts for each of the three content categories were significant for all but post-treatment NE. Thought samples rated as showing the presence of NE, PE, or SF showed significantly higher NE, PE, and SF (respectively) word counts than those without. Comparisons of pre/post data showed significant decreases in NE and no differences in PE across both methods; increases in SF emerged only for ratings. While limited by a small sample size, these findings suggest that word count analyses may be a reasonable replacement for more laborious hand coding in thought sampling data, but there may be important differences across content categories. These results contribute to knowledge about the methodology of thought sampling analysis in clinical samples.

Keywords: depression | word count | thought sampling | experience sampling | cognitive change

Article:

Researchers have had a longstanding interest in assessing thought content in clinical conditions and in examining how content changes over time, but the methods for generating and evaluating thought content vary. Depression, for example, is characterized in part by negative emotional and self-focused thinking (Mor & Winquist, 2002), and treatment can lead to decreases in negative thinking (Christopher et al., 2009; Hofheinz et al., 2020; Parrish et al., 2009) and rumination (van der Velden et al., 2015). However, evidence of changes in cognitive content has come primarily

from self-report questionnaires that contain standard thought ratings, checklists, or vignettes. While these measures have good psychometric properties, concerns have been raised about retrospective reporting as well as the limited number of situations and occasions sampled (Chamberlain & Haaga, 1999), prompting recommendations for the use of real-time measures (Ben-Zeev et al., 2009).

Thought sampling procedures allow researchers to examine in-the-moment thought content and patterns using open-ended thought probes that prompt idiographic verbal or written responses, an approach frequently used in studies of attention and mind-wandering (McVay & Kane, 2009). Studies using thought sampling collected in lab settings have shown greater negative thinking associated with depression (Josephson et al., 1999). A less common but a more ecologically valid approach involves sampling in-the-moment thought content in the context of daily life, such as in experience sampling or daily diary studies. Experience sampling methods allow researchers to gather thought samples at random times throughout the day and in a variety of personally relevant daily life situations; aggregation across those data points can increase reliability of thought content measurement (Chamberlain & Haaga, 1999).

While idiographic thought sampling can overcome some of the concerns about the validity of nomothetic approaches, the downside is the time-consuming coding process required. Linguistic and speech analysis tools, such as the Linguistic Inquiry and Word Count program (LIWC; Tausczik & Pennebaker, 2010), have the potential to provide a less labor-intensive approach. Programs like LIWC include a variety of word dictionaries that can be used to categorize (and count) word usage. However, there are challenges in using LIWC for analyzing thought probes. First, it was originally designed for longer written passages (Tausczik & Pennebaker, 2010), and its accuracy for characterizing shorter language samples, such as those used in experience sampling or diary designs, is not well established. In addition, simple word counts do not account for context when analyzing data, so a sentence like, "I do not feel happy" would be counted as indicating positive emotion, despite the fact that the overall statement suggests the opposite.

Given concerns about the "context-blind" nature of automated word count programs, how do results compare with hand-coding when examined head-to-head? One study (Ziemer & Korkmaz, 2017) compared hand coding with LIWC for expressive writing narratives in chronic pain patients and found a modest correspondence for positive emotion but not for negative emotion or first-person pronoun use. Another study (Alpers et al., 2005) compared hand- versus LIWC-coded online support group messages and reported Spearman r's ranging from 0.23 to 0.52, with lower values associated with specific emotion categories (e.g., anger). Zheng and Schweickert (Zheng & Schweickert, 2021) compared the LIWC to hand coding of dream reports and found good correspondence between the two analysis methods. In these studies, text passages were multiple sentences or paragraphs in length, and direct comparisons of shorter text samples is important in establishing the validity of word count approaches.

Only one study of depression to our knowledge has used open-ended thought sampling procedures in the context of daily life. This study of a non-clinical sample of adolescents (Mor et al., 2010) probed thoughts in a daily diary design and hand coded participants' written responses for self-focus; they found a stronger relationship between self-focus and negative mood among a community-based sample of adolescents with (versus without) a diagnosis of depression. A more recent study used the LIWC program to analyze daily diary event descriptions in a sample of depressed and nondepressed individuals (Krejtz et al., 2020). They found a greater use of negative emotion words and first-person pronouns, and a lower use of positive emotion words, among the participants with major depressive disorder. Although this study used event descriptions rather than

thoughts, it is notable that the descriptions were short (a mean of around 9 words per description), potentially supporting the validity of LIWC with shorter text samples. Other studies have also examined aspects of language (e.g., negative affect, self-focus) in written responses to prompts about daily events rather than thoughts (Wood et al., 1990) in general samples. Tov and colleagues looked at the correspondence between LIWC positive and negative emotion words extracted from descriptions of daily events in a diary study; they reported good correspondence with self-reported negative emotion in both studies and with positive emotion in one of the two studies (Tov et al., 2013).

The primary aim of the current study was to examine the convergent validity of two content analysis methods – manual hand coding and automated word count – as applied to thought samples obtained from participants with depression. We consider hand coding to be more accurate, as it allows for consideration of contextual factors. However, if there is good convergence between the two methods, it would suggest that automated word count approaches may be a reasonable substitute for the more labor-intensive coding procedure. This study has several novel features. First, it involved collecting in-the-moment open-ended thought samples in the form of phone-based audio samples from a clinical sample of treatment-seeking depressed participants in the context of their everyday lives. Second, thought sampling was collected both before and after completion of short-term, structured individual therapy in order to examine changes in thought content longitudinally. Finally, the audio samples were transcribed and analyzed both by hand and with automated word counts (LIWC; we used the 2015 version, LIWC2015), specifically assessing the thought samples for presence of positive and negative emotion words and for self-focus.

We used two methods for examining convergence between the two content analysis approaches. First, we evaluated their direct correspondence both at pre-treatment (Time 1) and at post-treatment (Time 2). We based this analysis on the most commonly used content categories in the depression literature: negative emotion, positive emotion, and self-focus. Second, we evaluated whether the two methods are similar in their detection of cognitive changes from pre- to post-treatment and consistent with previous findings. Based on the limited previous work, we expected good correspondence between the two methods of content assessment. We also predicted that participants would show a decrease from Time 1 to Time 2 in their use of first-person singular pronouns and negative emotional words, as well as an increase in positive emotional words, as indicated by both automated word counts and by hand coding.

Method

Participants

This study used archival data from a randomized controlled trial in which participants with a diagnosis of major depression or dysthymia received one of two forms of brief structured therapy (Eddington et al., 2015). These treatments, both of which significantly and substantially reduced depressive symptoms, were not compared against each other as there were no major differences found between the two treatments in terms of effectiveness (Eddington et al., 2015), and they are similar in terms of their structure and focus (e.g., building coping skills, consistent use of home practice). Exclusion criteria were antidepressant use in the past 4 months, history of mania, diagnosis of antisocial or borderline personality disorder, history of psychotic symptoms, and active suicidal intent or self-harm. Thirty of the 56 eligible participants completed phone-based experience sampling surveys at both pre- and post-treatment (excluded participants were treatment

non-completers). Excluded participants had slightly and nonsignificantly higher BDI-II scores at baseline than the included participants [M = 36.6 vs. 32.9; $t(54) = 1.70$; $p > .05$]. Participants were majority female (80%) with a mean age of 38. Participants were recruited using advertisements in mental health magazines, flyers in clinics, and websites, and all participants provided informed consent per the approved IRB protocol.

Measures

Beck Depression Inventory-II (BDI-II; Beck et al., 1996): This 21-item self-report questionnaire is utilized to assess depressive severity within the past two weeks. Items are rated on a 0 to 3 expanded format scale with total scores ranging from zero to 63, with 63 reflecting more severe symptoms. Scores above 14 were required for inclusion in the study.

Semi-structured Clinical Interview for DSM-IV-TR: Research Version. The SCID-I (First et al., 2002) and SCID-II (First et al., 1997) are semistructured diagnostic interviews for DSM-IV-TR (American Psychiatric Association, 2000) Axis I and Axis II disorders, respectively, and they were used to determine study eligibility. From the SCID-I, trained study diagnosticians administered the overview and screening sections, mood modules, anxiety modules, and psychotic symptoms; from the SCID-II, the antisocial and borderline personality modules were used.

Thought Sampling: A phone-based Interactive Voice Response System (IVRS; Telesage https://telesage.com/) was used for the experience sampling (see procedures section below). The program automatically called the cell phone of each participant and asked them pre-recorded questions about topics such as activities, stressors, physical activities, as well as social activities specific to the time of the call. For the purposes of this study, the analyses specifically used an open-ended question, “Please describe what were you thinking at the time of the call.” Participants provided verbal responses which were recorded and output as audio files.

Procedures

After informed consent was obtained for this IRB-approved study, the BDI-II was administered at the initial screening. If the BDI-II score was above 14, relevant portions of the diagnostic interviews were conducted by clinical psychology doctoral students who received training on the administration of the SCID-I and SCID-II. Participants were given instructions for the experience sampling procedure and were then called on a mobile phone eight times a day (each within 90-minute segments) for seven consecutive days (56 times in total) prior to the first treatment session. Call segments were randomized throughout a 12 h call window and, if a call was missed during this time, participants had the opportunity to call back but only within a 10 min span. Short-term, structured, individual psychotherapy for depression (either cognitive-behavioral therapy or self-system therapy) was provided for up to 16 sessions. The experience sampling procedure was repeated during the week prior to the final therapy session.

Audio Transcription and Data Preparation

Audio recordings were transcribed by research assistants. After initial transcription of recordings, they were checked by a second transcriber and any spelling errors were corrected. After all responses were double checked for accuracy, inaudible markers and nonfluencies (e.g. um, uh, hm) were omitted. For the purposes of the current analyses, we were primarily interested in general

tendencies in thought content (e.g., positive or negative emotion), which calls for aggregation of experience sampling data (Zelenski & Larsen, 2000). In other words, daily probing produced a pool of thought sample statements from which to derive an overall index of the content categories (positive emotion, negative emotion, and self-focus) for each person, before and after treatment. This aggregation approach has been used in prior studies comparing coding methods for repeated measures text samples (Alpers et al., 2005; Mota et al., 2020).

Word Counts

The 2015 version of the Linguistic Inquiry and Word Count program (Pennebaker et al., 2015) was used to calculate word count percentages of each audio sample based on 3 different categories for the purposes of the current study. The dictionary categories of interest for this study include first-person singular Personal Pronouns (abbreviated as LIWCSF as an indicator of self-focus), Positive Emotion words (or LIWCPE, which contains 620 words like "good" or "happy"), and Negative Emotion words (or LIWCNE, which contains 744 words like "hate" or "enemy"). Total word counts for each audio sample were also calculated. It should be noted that nonparametric tests were used for word count totals and positive and negative emotion percentages, which had distributions that deviated from normal.

Hand Coding

Transcribed audio samples were hand-coded by trained coders based on whether each sample included positive emotion, negative emotion, and a focus on the self (each category was coded separately as either "1 = present" or "0 = absent"). Positive and negative emotion was coded as "1" when the thought sample included explicit emotion language or described positive or negative feelings, abbreviated as HCPE and HCNE (hand coded positive and negative emotion). Note that a single thought sample could include both positive and negative emotion (e.g., "I am worried about visiting my uncle but I'm glad my sister is going with me"). Self-focus was coded as "1" when a response referenced thought content (i.e., an event, activity, or emotion) in which the respondent identified themselves, abbreviated as HCSF (hand coded self-focus). This coding omitted the common start to the response "I was thinking…" which directly responded to the prompt.

Coders were initially trained on a sample of thoughts from participants who were excluded from the current study sample (due to dropping out) to refine coding instructions. Following the training phase, weekly coding meetings were held to prevent rater drift and to resolve uncertainties in the coding by consensus. Each sample was coded by 2 different coders who were blind to participant number, treatment condition, and time point (pre- versus post-treatment). Inter-rater reliability, calculated as Phi coefficients, was very good (HCPE, 0.95; HCNE, 0.92; HCSF, 0.91).

Results

Word Counts and Responsiveness

Regarding participant responsiveness to the 56 possible IVR surveys at each time point, there were more responses to the IVR signals at Time 1 compared to Time 2 (see Table 1), a difference that was significant [$t(29) = 3.14$, $p < .01$; 95% CI 1.84, 8.70]. Total number of responses ranged from

16 to 102. A Wilcoxin signed-rank test indicated that there was not a significant difference in average word counts for the thought sampling item between the two time points ($Z = -0.50$, $p > .05$). Correlations between the content metrics and overall word counts are displayed in Table 2.

Table 1 Descriptive Statistics and Mean Differences for Thought Sampling Variables (N = 30)

	Time 1 M (SD)	Time 2 M (SD)	Test statistic1	Effect Size
# IVR Responses	**35.0 (12.8)**	**29.8 (13.7)****	**Z = 3.04**	**d = 0.55**
Average Word Count	17.8 (17.2)	20.4 (21.3)	Z = 0.50	
LIWCPE	4.7 (7.3)	6.3 (11.7)	Z = 0.85	
LIWCNE	**4.9 (8.8)**	**2.0 (2.8)****	**Z = 2.61**	**d = 0.48**
LIWCSF	11.7 (5.6)	12.1 (4.2)	Z = 0.20	
HCPE	0.16 (0.19)	0.21 (0.21)	*t* (29) = -1.49	
HCNE	**0.30 (0.17)**	**0.22 (0.17)****	***t* (29) = 3.20**	**d = 0.58**
HCSF	**0.62 (0.30)**	**0.68 (0.26)***	***t* (29) = -2.17**	**d = − 0.40**

[1] Paired-samples t-tests or (for non-normal data) Wilcoxon signed rank
Note: Based on per-participant averages from Time 1 (pre-treatment) and Time 2 (post-treatment).
LIWC = word count percentage generated by the LIWC2015 program.
Difference is significant at * $p < .05$; ** $p < .01$.

Convergence of Content Analysis Methods

First, we examined correlations between within-person average word count percentages and hand-coded ratings for the Time 1 and Time 2 data. Two different participants, one at Time 1 only and one at Time 2 only, were identified as clear outliers, having word counts that were more than 3 standard deviations above the mean. We therefore excluded those two data points from the word count correlations, shown in Table 2. Spearman correlation coefficients were significant and positive for the self-focus category across time points and content analysis methods. For the emotion categories, correlations with word counts were nonsignificant for all but Time 2 HCNE.

Secondly, we examined correspondence by using the hand-coding as a content "benchmark" to test whether word count averages were higher in the samples coded as "1" (indicating the presence of negative emotion, for example) compared to those coded as "0" (negative emotion not present). Given the two major concerns with the validity of the LIWC approach for these data, the brevity of the thought samples and the lack of context inherent in simple word counts, we assumed that hand-coding provides a more accurate characterization of thought content. For each participant, an average for LIWCPE, LIWCNE and LIWCSF was calculated across both time points for the "1" and "0" thought samples, and those averages were then contrasted across all participants as paired samples.

For negative emotion content, the LIWCNE percentage in samples hand-coded as "1" was significantly higher than in those coded as "0" [$M1 = 10.83$; $SD1 = 15.01$; $M0 = 0.92$, $SD0 = 2.22$; $t(28) = 3.79$; $p < .001$; $d = 0.70$]; the same results were found for LIWCPE [$M1 = 16.97$; $SD1 = 19.61$; $M0 = 2.32$, $SD0 = 3.96$; $t(28) = 3.5$; $p < .001$; $d = 0.78$]. Likewise, for LIWCSF, the mean percentage of first-person pronouns in samples hand-coded as "1" was significantly higher than in those coded as "0" [$M1 = 16.16$; $SD1 = 4.71$; $M0 = 6.32$, $SD0 = 4.89$; $t(28) = 5.98$; $p < .001$; $d = 1.3$].

Table 2 Within-Category and Word Count Correlations (N = 30)

	Time 1			**Time 2**		
	LIWC/HC	HC/WC	LIWC/WC	LIWC/HC	HC/WC	LIWC/WC
PE	**0.63****	0.30	− 0.27	**0.80****	0.15	− 0.18
NE	**0.62****	0.31	− 0.25	0.31	**0.59****	0.08
SF	**0.67****	**0.83****	**0.40***	**0.77****	**0.85****	**0.39***

Note: LIWC/HC = Spearman correlation between average LIWC2015 word count percentage and hand coded rating for the indicated category (PE = positive emotion; NE = negative emotion; SF = self-focus); HC/WC = Spearman correlation between average hand coded rating and total word count (averaged across all thought samples).
Significant at * $p < .05$; ** $p < .01$.

Pre- to Post-Treatment Differences

Table 1 summarizes the differences in pre- and post-treatment thought content for each of the content categories. Consistent with our expectations, across both the LIWC2015 and hand-coding results, there was a significant decrease in NE content at post-treatment, but (contrary to predictions) no differences emerged for PE content. For self-focus, results varied by content assessment method. First person singular pronoun use was similar at both time points based on the LIWC2015 data. However, the hand-coded data showed a significant increase in SF at post-treatment, in the opposite direction of our predictions.

Discussion

The current study is the first to our knowledge to examine open-ended thought samples before and after short-term therapy for depression using both hand coding and word count procedures. We aimed to examine the convergent validity of the LIWC2015-generated word count percentages in relation to the hand-coded data, which we considered more accurate because it allows for the content to be situated contextually within the broader text sample. If there is good agreement the two methods, it could support the use of less laborious word count programs for thought sampling data. Our evaluation included a direct comparison of the two methods via correlation coefficients and using mean comparisons between samples hand coded as positive (versus negative) for the presence of each content category. We also examined the relative correspondence between the methods in a more applied approach by examining pre- to post-treatment changes.

Given possible concerns with the utility of word count programs with shorter responses, we examined the extent to which the content metrics generated from the LIWC2015 program and from hand-coding were associated with overall word counts. There were notable individual differences in how "wordy" participants were in response to the thought probes. For PE and NE, the length of thought samples was significantly associated with hand-coded ratings indicating the presence of negative emotion only at Time 2. Given the significant reduction in NE content from pre- to post-treatment in this sample, raters may have been better able to detect NE in longer samples at Time 2, when there was less NE to detect. Overall word counts were consistently associated with presence of self-focus, with stronger correlations for hand-coded ratings than for LIWC2015 first person pronoun percentages. These findings may suggest that more self-focused responders tend to be wordier, which is consistent with our anecdotal observation that some

participants used the thought probes as an opportunity to "vent" about their personal experiences. In addition, proportionally, their thought samples also contained more first-person pronouns.

Regarding the correspondence between the two content analysis methods, correlations were moderate to strong and positive across all categories and time points with the exception of Time 2 NE, which was in the positive direction but nonsignificant. Across time points, the LIWC2015-generated percentages of PE, NE, and SF words were higher in the thought samples coded by raters as positive ("present") for the corresponding content compared to those coded as negative. In general, this pattern suggests good correspondence between the two content analysis methods for these categories.

How well do these two methods correspond in a more applied context, examining pre- to post-treatment changes? We expected that, from pre- to post-treatment, there would be reductions in negative emotion and self-focused content as well as increases in positive emotion. Across both word count and hand-coded approaches, significant reductions in negative emotion were observed, with a large effect size in both cases. Contrary to predictions, however, positive emotion showed slight but nonsignificant increases from pre- to post-treatment across both methods. To the extent that the content of thought samples is reflective of a person's affective experience, these findings are partially consistent with the literature showing that intervention leads to an increase in positive affect and a decrease in negative affect (Boumparis et al., 2016). Quantitative item ratings of positive mood from the current sample, collected at the same time as the thought probes, in fact showed significant improvements over time (Eddington et al., 2017), suggesting that momentary thought samples may be less sensitive to changes in positive affective experiences.

For self-focus, only the hand-coded data showed a significant difference over time, but in the opposite direction – self-focus content was higher at post-treatment. Although studies have consistently reported higher rates of first person pronoun use in depressed compared to nondepressed samples (Edwards & Holtzman, 2017), it is a weak effect. Furthermore, the nature of thought sampling requires self-focus – to assess and report on their momentary thoughts, a person must "turn inward" and reflect on themselves and their inner world. As such, it may be difficult to discern thoughts that reflect the find of excessive, ruminative self-focus that is theorized to characterize depression. The extremely high correlations with overall word counts raise further questions about what exactly is being captured by these ratings. It is possible that thought sampling procedures like those used in this study do not provide an accurate snapshot of self-focus and may not be sensitive to individual differences in degree of self-focus (or to changes in self-focus over time). We should also note that a limitation of our sample is that it is predominantly female. Women are more prone to self-focus and self-focus plays a more prominent role in depression in females (Ingram et al., 1988), therefore differences may be obscured in our small sample.

Overall, these applied findings support the convergence of hand-coding and word count methods for the emotion categories but highlight a difference for the self-focus category. The lack of control group prohibits us from drawing causal conclusions about the impact of therapy on thought content, and the sample was small, a prevalent issue in the psycholinguistic literature. While this hindered our ability to conduct more fine-grained analyses (such as differences between therapy conditions), the intensive aspect of the methodology used permitted us to collect large quantities of data per participant. The naturalistic approach to thought sampling in this study is a clear strength -- a simple "what were you thinking?" prompt was given as participants went about their daily lives. However, this approach yielded audio responses varying from a maximum of 250 words and a minimum of 1. Providing more instruction to participants regarding response length or building in follow up prompts for shorter responses may help increase consistency.

One of the concerns with these brief thought samples is that word count approaches, which were initially developed for longer written narratives, may be less useful. We observed relatively good agreement between the word count results and the more laborious and context-sensitive hand coding, but not across the board. Mean word count proportions in thought samples hand-coded as showing the presence of emotional or self-focused content were consistently significantly higher than those without such content, and correlations were positive and significant for all but Time 2 NE. As noted above, when comparing pre- and post-treatment analyses, there was agreement in the emotion categories but a discrepancy in self-focus. A contributing factor in this discrepancy may be the fact that many of the thought samples started with something like, "I was thinking…". The hand-coders were instructed not to automatically code those as self-focused but instead to examine the rest of the thought sample, but these components were not excluded from the word counts, which may have obscured differences.

Because previous direct comparisons of LIWC and hand coding using longer text passages have produced mixed findings in terms of agreement, more research in this area is needed. Our results are somewhat optimistic when considering the correspondence between the two methods for brief thought samples, but caution is warranted given that results may differ across content categories. We also note that a newer version of the LIWC program was released after our data had been analyzed (Boyd et al., 2022). The newer version divides the positive emotional and negative emotional word categories into "emotion tone," which reflects general affective sentiments, and "emotion," denoting specific feelings reported or strongly implied. Our general NE and PE categories collapse across these two subcategories and therefore our results cannot directly address the question of their convergent validity.

The vast majority of clinical psycholinguistic findings have come from experimental procedures which only include writing tasks. We suggest that even seemingly minor response latencies produced by writing tasks (e.g., the delay that occurs while a person carefully decides what to write and how to write it) compared to spoken ones may impact thought content, underscoring the importance of utilizing different methods of data collection. Efforts to interpret future studies like ours would probably benefit from clarifying whether differences exist between spoken language patterns and written ones, and if so, how natural language processing might account for such differences (e.g., modified word count dictionaries for written versus spoken language). Thought samples may provide a window into daily cognitive life, and as automated linguistic analysis tools evolve to make transcription and coding quicker and more efficient, these developments may facilitate the clinical study of cognitive mechanisms.

Data Availability

The dataset analyzed for the current study are not publicly available because they consist of audio samples that cannot be completely de-identified.

References

Alpers, G. W., Winzelberg, A. J., Classen, C., Roberts, H., Dev, P., Koopman, C., & Taylor, B., C (2005). Evaluation of computerized text analysis in an internet breast cancer support group. Computers in Human Behavior, 21(2), 361–376. https://doi.org/10.1016/j.chb.2004.02.008.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). Manual for the Beck Depression Inventory-II. Psychological Corporation.

Ben-Zeev, D., Young, M. A., & Madsen, J. W. (2009). Retrospective recall of affect in clinically depressed individuals and controls. Cognition & Emotion, 23(5), 1021–1040.

Boumparis, N., Karyotaki, E., Kleiboer, A., Hofmann, S. G., & Cuijpers, P. (2016). The effect of psychotherapeutic interventions on positive and negative affect in depression: A systematic review and metaanalysis. Journal of Affective Disorders, 202, 153–162. https://doi.org/10.1016/j.jad.2016.05.019.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22

Chamberlain, J., & Haaga, D. A. F. (1999). Convergent Validity of Cognitive Assessment Methods. Behavior Modification, 23(2), 294–315. https://doi.org/10.1177/0145445599232007.

Christopher, M. S., Jacob, K. L., Neuhaus, E. C., Neary, T. J., & Fiola, L. A. (2009). Cognitive and behavioral changes related to symptom improvement among patients with a mood disorder receiving intensive cognitive-behavioral therapy. Journal of Psychiatric Practice, 15(2), 95–102. https://doi.org/10.1097/01.pra.0000348362.11548.5f.psyh.

Eddington, K. M., Silvia, P. J., Foxworth, T. E., Hoet, A., & Kwapil, T. R. (2015). Motivational deficits differentially predict improvement in a randomized trial of self-system therapy for depression. Journal of Consulting and Clinical Psychology, 83(3), 602–616. https://doi.org/10.1037/a0039058.

Eddington, K. M., Burgin, C. J., Silvia, P. J., Fallah, N., Majestic, C., & Kwapil, T. R. (2017). The Effects of psychotherapy for major depressive disorder on Daily Mood and Functioning: A longitudinal experience sampling study. Cognitive Therapy and Research, 41(2), 266–277. https://doi.org/10.1007/ s10608-016-9816-7.

Edwards, T., & Holtzman, N. S. (2017). A meta-analysis of correlations between depression and first person singular pronoun use. Journal of Research in Personality, 68, 63–68. https://doi.org/10.1016/j.jrp.2017.02.005.

First, M. B., Gibbon, M., Spitzer, R. L., Williams, J. B. W., & Benjamin, L. S. (1997). Structured clinical interview for DSM–IV axis II personality disorders (SCID-II). Washington, DC: American Psychiatric Press.

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (2002). Structured clinical interview for DSM–IV–TR axis I disorders. New York: Biometrics Research, New York State Psychiatric Institute.

Hofheinz, C., Reder, M., & Michalak, J. (2020). How specific is cognitive change? A randomized controlled trial comparing brief cognitive and mindfulness interventions for depression. Psychotherapy Research, 30(5), 675–691. https://doi.org/10.1080/10503307.2019.1685138.

Ingram, R. E., Cruet, D., Johnson, B. R., & Wisnicki, K. S. (1988). Self-focused attention, gender, gender role, and vulnerability to negative affect. Journal of Personality and Social Psychology, 55, 967–978. https://doi.org/10.1037/0022-3514.55.6.967.

Josephson, B. R., Rose, R. D., & Singer, J. A. (1999). Thought sampling after mood induction in depressed vs non-depressed college students. Imagination Cognition and Personality, 19(1), 27–37. https://doi.org/10.2190/UWA1-HF64-B26N-8L8P. APA PsycInfo.

Krejtz, I., Rohnka, N., Holas, P., Rusanowska, M., & Nezlek, J. B. (2020). Manifestations of clinical depression in daily life: A daily diary study of descriptions of naturally occurring events. Cognition & Emotion, 34(8), 1664–1675. https://doi.org/10.1080/02699931.2020.1795627.

McVay, J. C., & Kane, M. J. (2009). Conducting the train of thought: Working memory capacity, goal neglect, and mind wandering in an executive-control task. Journal of Experimental Psychology: Learning Memory and Cognition, 35(1), 196–204. https://doi.org/10.1037/a0014104.

Mor, N., & Winquist, J. (2002). Self-focused attention and negative affect: A meta-analysis. Psychological Bulletin, 128(4), 638–662.

Mor, N., Doane, L. D., Adam, E. K., Mineka, S., Zinbarg, R. E., Griffith, J. W., Craske, M. G., Waters, A., & Nazarian, M. (2010). Within-person variations in self-focused attention and negative affect in depression and anxiety: A diary study. Cognition and Emotion, 24(1), 48–62. https://doi.org/10.1080/02699930802499715.

Mota, N. B., Weissheimer, J., Ribeiro, M., de Paiva, M., Avilla-Souza, J., Simabucuru, G., Chaves, M. F., Cecchi, L., Cirne, J., Cecchi, G., Rodrigues, C., Copelli, M., & Ribeiro, S. (2020). Dreaming during the Covid-19 pandemic: Computational assessment of dream reports reveals mental suffering related to fear of contagion. Plos One, 15(11), https://doi.org/10.1371/journal.pone.0242903.

Parrish, B. P., Cohen, L. H., Gunthert, K. C., Butler, A. C., Laurenceau, J. P., & Beck, J. S. (2009). Effects of cognitive therapy for depression on daily stress-related variables. Behaviour Research and Therapy, 47(5), 444–448. https://doi.org/10.1016/j.brat.2009.02.005.psyh.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Austin, TX: University of Texas at Austin.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and Computerized text analysis methods. Journal of Language and Social Psychology, 29(1), 24–54. https://doi.org/10.1177/0261927X09351676.

Tov, W., Ng, K. L., Lin, H., & Qiu, L. (2013). Detecting well-being via computerized content analysis of brief diary entries. Psychological Assessment, 25(4), 1069–1078. https://doi.org/10.1037/a0033007.

van der Velden, A. M., Kuyken, W., Wattar, U., Crane, C., Pallesen, K. J., Dahlgaard, J., Fjorback, L. O., & Piet, J. (2015). A systematic review of mechanisms of change in mindfulness-based cognitive therapy in the treatment of recurrent major depressive disorder. Clinical Psychology Review, 37, 26–39. https://doi.org/10.1016/j.cpr.2015.02.001.

Wood, J. V., Saltzberg, J. A., Neale, J. M., Stone, A. A., & Rachmiel, T. B. (1990). Self-focused attention, coping responses, and distressed mood in everyday life. Journal of Personality and Social Psychology, 58(6), 1027–1036. https://doi.org/10.1037//0022-3514.58.6.1027.

Zelenski, J. M., & Larsen, R. J. (2000). The distribution of Basic Emotions in Everyday Life: A state and trait perspective from experience Sampling Data. Journal of Research in Personality, 34(2), 178–197. https://doi.org/10.1006/jrpe.1999.2275.

Zheng, X., & Schweickert, R. (2021). Comparing hall Van de Castle coding and linguistic Inquiry and Word Count using canonical correlation analysis. Dreaming, 31(3), 207–224. https://doi.org/10.1037/drm0000173.

Ziemer, K. S., & Korkmaz, G. (2017). Using text to predict psychological and physical health: A comparison of human raters and computerized text analysis. Computers in Human Behavior, 76, 122–127. https://doi.org/10.1016/j.chb.2017.06.038.

American Psychiatric Association (2000). Diagnostic and statistical manual of mental disorders (4th ed - Text revision). American Psychiatric Publishing, Inc. http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=1994-97698-000&site=ehost-live