PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES OPEN ACCESS

For living well, behaviors and circumstances matter just as much as psychological traits

William R. Hobbs[a,b,1] and Anthony D. Ong[a]

Edited by Timothy Wilson, University of Virginia, Charlottesville, VA; received August 1, 2022; accepted January 15, 2023

In 2004 through 2016, three studies in the national Midlife in the United States (MIDUS) project asked participants the open-ended question "What do you do to make life go well?". We use verbatim responses to this question to evaluate the relative importance of psychological traits and circumstances for predicting self-reported, subjective well-being. The use of an open-ended question allows us to test the hypothesis that psychological traits are more strongly associated with self-reported well-being than objective circumstances because psychological traits and well-being are similarly self-rated—meaning that they both ask respondents to decide how to place themselves on provided and unfamiliar survey scales. For this, we use automated zero-shot classification to score statements about well-being without training on existing survey measures, and we evaluate this scoring through subsequent hand-labeling. We then assess associations of this measure and closed-ended measures for health behaviors, socioeconomic circumstances, biomarkers for inflammation and glycemic control, and mortality risk over follow-up. Although the closed-ended measures were far more strongly associated with other multiple-choice self-ratings, including Big 5 personality traits, the closed- and open-ended measures were similarly associated with relatively objective indicators of health, wealth, and social connectedness. The findings suggest that psychological traits, when collected through self-ratings, predict subjective reports of well-being so strongly because of a measurement advantage—and that circumstance matters just as much when assessed using a fairer comparison.

well-being | health | personality | machine learning | survey design

Significance

Perhaps the best way to see whether someone is doing well is to ask them. Happiness researchers do this using multiple-choice questions. Responses to these questions track life's ups and downs, but across individuals, they are far more strongly associated with subjective psychological traits, which are rated in similar ways as well-being, than more objective measures of behaviors or circumstances. Here, we assess whether well-being from an open-ended question paints a different picture. We show that it does. Compared to the open-ended measure, closed-ended measures appear to overstate associations with self-reported psychological traits–and, in relative terms, understate associations with health indicators, behaviors, and socioeconomic circumstances that might be more justifiably targeted and supported by public policies and behavioral interventions.

Over the last several decades, measuring happiness and life satisfaction has become increasingly valued in the assessment of the effects of policies, societal-level events, and behavioral interventions (1–6). As societies have become richer and increases in longevity have slowed, there has been a shift beyond efforts to promote economic well-being alone.

In measuring well-being,* simply asking someone how they are doing can be surprisingly effective. Across hedonic (e.g., positive minus negative affect) and eudaimonic (e.g., sense of purpose) indicators of well-being, multiple-choice questions produce measures associated with important nonsurvey outcomes, including mortality risk (9–13). Because many different questions tend to result in broadly similar responses (14–16)— although with some differences in their associations with other variables (4, 17)—and because of the importance of happiness and life satisfaction to a very wide range of research topics, research has now converged on using a small number of very short batteries that can be more readily added to data collection efforts, such as in the General Social Survey (18).

A methodological concern in the assessment of self-reported well-being is that life satisfaction and even day-to-day happiness appear to be strongly associated with many psychological traits, especially self-reported personality traits (16, 19, 20), and more strongly associated with these traits than circumstances (19, 20). An association with often (or perceived to be) enduring and difficult-to-change factors like personality traits, relative to circumstances, could suggest that individuals and policy-makers might have little influence on these forms of well-being in comparison or even that interventions should target personality traits themselves (21, 22). This might be especially the case if psychological traits are thought to primarily influence well-being directly—that is, without also depending on objective changes in a person's behaviors and circumstances. Personality trait associations are so large that they cannot be readily explained through traits' much weaker, or thought-to-be weaker, associations with objective behaviors and circumstances (23).

*"Subjective" and "psychological" well-being are associated with specific measures of well-being, which aim to distinguish hedonic and eudaimonic dimensions of well-being (7, 8). They are also now used to refer to concepts of well-being more broadly, especially across social science disciplines. Happiness is often used as an umbrella term for many forms of subjectively reported well-being and does not refer to any specific measure.

Author contributions: W.R.H. and A.D.O. designed research; W.R.H. performed research; W.R.H. analyzed data; and W.R.H. and A.D.O. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: hobbs@cornell.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2212867120/-/DCSupplemental.

Published March 13, 2023.

Of course, researchers have long recognized that the predominance of personality trait associations might be due to a quirk in how subjective reports of well-being are assessed (24). Personality traits and well-being are rated in similar ways, and some well-being and personality assessments even use the same questions (25). However, removing repeated questions across batteries still retains a very strong association between personality and subjective well-being (26), and many alternate forms of self-reports of well-being are strongly correlated with personality traits in comparison to more objective circumstances (27, 28). Indeed, some researchers in psychological science, based on careful statistical analysis, argue that well-being, along with many other scales, should be considered a facet of personality (22), given such strong correlations with existing personality measures.

If survey behaviors that affect both types of self-ratings drive the strong associations between personality traits and well-being—without also reflecting real-world circumstances or outcomes—then we should find less strong associations between these two constructs when we measure one of them very differently (27, 29) and perhaps especially if we do so using self-reports that do not rely on closed-ended questions at all. With this, we would anticipate that closed-ended questions and measures for circumstances, especially ones that rarely involve subjective self-reports, would be placed on a more equal footing. And recent advances in language modeling now allow directed and highly replicable scoring of well-being in open-ended survey responses.

We use open-ended survey responses and automated text scoring to conduct such a test here, assessing whether responses to an open-ended question can more fully break circularity from shared response tendencies to subjective closed-ended questions specifically—and that we would not expect to also affect closed-ended questions about circumstances. We also consider whether such an open-ended measure might provide a distinct and potentially uniquely valuable measurement of well-being, especially if it provides information not clearly captured in any multiple-choice responses.

To assess the value of an open-ended measure as an indicator of well-being and wellness across individuals, we use objective health indicators as bases of comparison, such as mortality risk over follow-up—based on the well-established finding that people who report "living well" tend to live healthier and longer too (4, 30–32). We also consider plausible and relatively objectively measured drivers and/or consequences of well-being (several of which are thought to be only weakly associated with subjectively reported well-being), focusing in particular on income (33–35), net worth (36), education level (37–39), marital status (40–42), and parenthood (42–44)—using family status as a minimal and more objective form of social connectedness, as survey respondents otherwise need to decide at what point social contacts no longer count as friendships in their closed-ended responses—as well as health behaviors (current smoking, self-reported exercise, and sleep quality, as measured by actigraphy) and health status (self-reported medical history and biomarkers for inflammation and long-term glycemic control) (4). These health variables are in the MIDUS data and have been a focus of past research on well-being and cardiovascular disease (30).[†]

To do so, we leverage an open-ended question ("What do you do to make life go well?") that appeared in three samples within the longitudinal and national Midlife in the United States study. To our knowledge, this is the only large such set of studies that systematically includes a fully open-ended response question specifically about well-being, and that is linked to many psychological constructs as well as objective indicators of wellness. Although we hope that similar questions will be more widely adopted in the future, other longitudinal studies currently include open-ended questions for other purposes, such as to list single-word responses that would be too varied for a closed-ended question or to provide an optional response to questions of the form "Is there anything else you would like to tell us?".

In scoring well-being from the open-ended survey responses, we use a recently developed text analysis technique (46), as implemented in ref. 47, to perform zero-shot classification, meaning classification without first training on example labels. Using a fully pretrained model (e.g., we do not train it on the MIDUS well-being measures or fine-tune the language model on the MIDUS open-ended texts), the method produces a probability that a text is "about" a specific label, as in a probability that "This text is about __.". To measure well-being, we use the method to score labels currently used by the Gallup organization in their widely used World Poll and Gallup-Sharecare Well-Being Index.[‡] In public-facing reports, Gallup summarizes levels of subjective well-being using the words "thriving," "struggling," and "suffering" (*Materials and Methods* for more information). This approach intentionally avoids, for our primary tests, a supervised model in which we train a model to reproduce the closed-ended responses. Such a model would have the capacity to reproduce biases in closed-ended responses, and, if a perfect fit, its predictions would be no different from the original self-ratings, whether or not predictive text features reflect underlying levels of well-being itself.

Although we emphasize the importance of an open-ended question over this (or any) specific method for evaluating responses to one, this scoring has several advantages for our research goals here, which we explain in more detail in *Materials and Methods*. Broadly, our goal is to evaluate an open-ended response format rather than a specific method for evaluating them. This scoring provides a highly replicable way of measuring well-being in written responses, and the use of a pretrained model minimizes researcher degrees of freedom. The method also makes full use of the unique and high-quality MIDUS studies, as we do not need to reserve any fraction of it for model training. In the last subsection of Results, we repeat our analyses using Linguistic Inquiry and Word Count (LIWC) (48), a widely used dictionary approach that also does not require in-sample training.

Finally, we use a supervised model to evaluate to what extent self-rating, closed-ended response styles might be at all reflected in open-ended text and, if personality traits are much more strongly associated with these predictions than circumstances and behaviors, what features in text responses tend to predict high levels of closed-ended well-being.

Results

Well-Being Score Descriptive Statistics and Comparison to Human Labels. The average well-being probabilities, the probability that "This text is about (thriving/struggling/suffering)," were 0.54 for thriving, 0.29 for struggling, and 0.10 for suffering. The index of these scores was calculated using the formula thriving minus struggling minus suffering, and the average score here was 0.16. This score, like the closed-ended well-being

[†] HbA1c and biomarkers for inflammation are less consistently associated with subjective or psychological well-being than more extensively studied incident cardiovascular disease (4, 30), although a number of studies have considered the possibility of associations. Salivary cortisol was not included in the biomarkers because of the complexity of its use as a biomarker for stress levels (45).

[‡] https://news.gallup.com/poll/122453/understanding-gallup-uses-cantril-scale.aspx.

measures, was then centered and standardized to SD units prior to all analyses. In *SI Appendix*, section S6, we replicate the paper's findings using each of these labels independently and show that an alternate index, thriving minus the average of struggling and suffering, does not alter the results. We chose thriving minus struggling minus suffering primarily because this index was simple; however, negative statements in response to a positively framed prompt could also be especially informative.

We use hand labels to evaluate the open-ended measure in a particular sense—not that it is necessarily measuring well-being itself (which is an ongoing debate even for long-standing well-being measures, ref. 49) but rather to create a set of instructions by which humans, together with simple document statistics, are able to reproduce the machine labels or vice versa. If the rater and machine labels are not equivalent, we then need to assess a) whether the labels still replicate similar associations with the personality and circumstances' auxiliary variables and b) whether the two might represent different forms of well-being that, in future work, might be used to create a more refined and all-encompassing well-being measure. We do not average the human and machine labels because the results of separate tests of associations with auxiliary variables (especially without strongly related variables) already reflect statistical reliability—and allow us to additionally assess convergent validity. The machine labels are useful because they are highly replicable and ones which we did not design ourselves, but there is little reason to expect that they will be superior to hand labels in general.

Research assistants were instructed not to try to rank how the respondent was doing or to label merely using the sentiment or general "feel" of a response but to instead focus on whether the respondent writes about: doing well or, as a direct response to the prompt, making life go well (thriving), having some form of difficulty when trying to make life go well (struggling), or life not going well (suffering)—whether or not they are referring to their own life (e.g., not distinguishing among own suffering, spouse's suffering, or suffering in the world), as the prompt for machine coding does not make this distinction. Research assistants did not discuss the labeling with each other as attempts to resolve coding differences would bias estimates of interrater reliability. All coding instructions were provided in the written codebook, which we show in *SI Appendix*, Fig. S1. All texts in the main analysis set (1,044 responses) were labeled, with 25% of texts labeled by both coders to assess interrater reliability.

In this labeling, the single-measure intraclass correlation was 0.58 (95% CI: 0.49 to 0.65) across hand raters. The single-measure correlation was lower, 0.39 (95% CI: 0.33 to 0.44), across the hand labels and machine scores. These intraclass correlations are comparable to those seen within closed-ended measures of well-being, such as general affect balance (0.50, 95% CI 0.47 to 0.52) and purpose in life (0.29, 95% CI 0.26 to 0.32; *SI Appendix*, Table S5), which achieve higher reliability only for averages across multiple and sometimes many closed-ended questions (e.g., the commonly used metric Cronbach's alpha, which tends to increase as more related or repetitious items are averaged).[§] This said, we cannot provide a simple comparison between multiple measures from a single open-ended question (one that could be perhaps viewed as a multipart question or battery) and multiple closed-ended questions (ones that respondents could interpret as repetitions of the same question). Instead, associations with auxiliary variables not measured through open-ended text or closed-ended self-ratings provide more comparable tests. Further discussion of the interpretation of reliability can be found in *SI Appendix*, section 1.3, and we display example texts[¶] with zero-shot labels and MIDUS percentiles in *SI Appendix*, section 3.

Some of the discrepancies between human and machine labels could be explained by differences in associations with document length and lower agreement for shorter documents (*SI Appendix*, Table S1). Document length was associated with slightly higher well-being in both the closed-ended well-being measures and the machine measure (*SI Appendix*, Table S2), while it was associated with substantially lower well-being in raters' labels. Raters were not asked to attempt to adjust for document length by, for example, scoring the fraction of texts referencing "thriving," "struggling," and "suffering."

Conversely, both the closed-ended well-being measures and the hand labels were weakly associated with a higher closed-ended measure of religious identification, while the machine scores were not associated (*SI Appendix*, Table S3), and the machine scores also appeared to associate mentions of religion in the text with lower well-being (*SI Appendix*, Table S4). This is perhaps because human raters were better able to identify the context of these responses, as they were aware of the question prompt. For example, although prayer might be associated with phrases used during a crisis like "Please keep us in your thoughts and prayers" in general text, which forms the basis of the zero-shot classification model, it could more often reference routine, daily prayer in response to the open-ended prompt about living well. In future work, the machine-coding approach here might be adjusted for mentions of religion that would on average be associated with struggling and suffering in general contexts but more often with thriving in response to this specific open-ended question.

In *SI Appendix*, we reproduce findings in the main paper to assess to what extent hand labels provide convergent validity. These analyses help assess whether discrepancies between hand and machine labels might reflect random differences, leading to attenuated correlations and larger confidence intervals across all analyses, or systematic differences, which could alter associations for personality versus circumstance. *SI Appendix*, section 1 shows that findings for the comparison of personality traits versus circumstances, as well as mortality risk over follow-up, are reproduced after we adjusted the research assistant ratings for document length, with slightly different associations on a correlate-by-correlate basis (e.g., the hand labels were more positively associated with having any living children and less positively associated with few symptoms/conditions in medical history) and with somewhat attenuated associations for the hand labels compared to the machine labels. In *SI Appendix*, section 2, we further display keywords associated with each measure of well-being. The keywords suggest that the two measures do place varying emphasis on different aspects of well-being, with words related to love, family, and religion used more often in the high-scoring hand labels; however, these varying foci do not influence the relative associations with personality and circumstance.

General Well-Being Correlates. In our first analysis, we present associations between well-being measures and prominent hypothesized correlates with well-being that have been documented in previous research, as listed in the introduction (*Materials and Methods* for descriptions of all variables).

[§] If we are willing to convert the intraclass correlations for the open-ended measures to Cronbach's alpha, these would be 0.72 between hand raters and 0.56 between the hand raters and the machine scores.

[¶] Real MIDUS responses cannot be publicly shared and can be requested by contacting midus_help@aging.wisc.edu.

For the closed-ended well-being measures, we focus on two measures that we expect to be maximally distinct from each other and from the personality trait ratings. We use one measure that is phrased entirely differently from closed-ended measures of personality traits (a single question asking about life satisfaction) and another measure that is temporally distinct from the personality trait questions, in that it asks a series of questions over eight days about whether the respondent felt good or bad the previous day (27 items, each repeated 8 times). In addition to being maximally distinct from the personality trait ratings, life satisfaction and day-to-day affect are also expected to be influenced by circumstances and behaviors in different ways. For life satisfaction, as argued in ref. 50, survey respondents might attempt to make a global assessment about their well-being by tallying aspects of their lives that they think should make their lives go well or that are viewed positively by society. This effect (what should make one happy, rather than what does) could be partly counteracted in more momentary assessments like the day-to-day affect balance measure, as respondents will be better able to assess their happiness while experiencing everyday events.

To expand the analyses further, we also include abbreviated analyses of the "psychological well-being" scale (42-item scale) and how the respondent has generally felt over the last 30 d (12-item scale) in the main text and complete analyses for these measures of well-being in *SI Appendix*. Question texts for these closed-ended questions can be found in *Materials and Methods*.

Fig. 1 displays the correlations between each predictor and the three primary measures of well-being. For personality traits, this figure displays adjusted multiple R (square root of adjusted R squared), which is the adjusted correlation between the predictions from ordinary least squares and the observed data. The adjustment accounts for increases in R-squared merely due to the inclusion of more variables in a regression—its influence is small given the sample size and the relatively small number of included variables. The figure displays each predictor in order of its association with life satisfaction. *SI Appendix*, Fig. S11 displays trait-by-trait associations for each measure of well-being.

The first finding to note is that personality traits are the strongest predictor of both life satisfaction and day-to-day positive minus negative affect ("day-to-day affect balance"). For this everyday happiness, other predictors are noticeably less strongly associated with well-being. For life satisfaction, wealth and marriage are comparably though less strongly associated with well-being than personality.

In contrast, personality traits are not significantly more strongly associated with the written well-being measure. This is not to say that personality does not matter—instead, because we measure personality and well-being using different survey formats, we are not providing an "unfair" advantage to personality relative to other factors that are not measured through multiple-choice self-ratings.

Fig. 2 displays combined models of personality traits versus behaviors and circumstances (i.e., all variables other than personality traits). This figure also displays associations for positive minus negative affect over the last 30 d ("general affect balance") as well as psychological well-being (8). All closed-ended measures of well-being are far more strongly associated with personality traits than the open-ended measure, while all well-being measures are similarly associated with behaviors and circumstances. At the same time, the open-ended measure does appear to be slightly less strongly associated with circumstance than most of the closed-ended well-being measures, including general affect balance and psychological well-being which in comparison with life satisfaction are perhaps not thought to be very strongly influenced by a tallying of what people think should make them happy. This suggests that there might still be plenty of room to improve the open-ended measurement technique (we use a technique for zero-shot classification that relies on a large language model from 2019, and we intentionally do not specifically train the model for this task) or that the open-ended prompt could be refined to encourage longer written responses about well-being. Subsection "LIWC and Supervised Models" at the end of the main text *Results* speaks to both of these possibilities.

Readers may note that the writing-based measure is more consistently associated with health behaviors (smoking and leisure physical activity) and the biomarkers of inflammation (index) and long-term glycemic control (glycosylated hemoglobin, HbA1c). By a small margin, self-reported physical activity that is leisure (i.e., that is not work or chores) is the strongest predictors of written "thriving." This finding could be driven by the format of the open-ended question, which asks "What do you do to make life go well?". This question plausibly primes respondents to think more about their physical health. Nonetheless, as shown in *SI Appendix*, Fig. S12, self-rated physical health is more strongly associated with the multiple-choice measures of well-being than the writing-based measure, perhaps because self-rated health is subject to the same sorts of response tendencies as life satisfaction and happiness.

Another possibility is that the open-ended question was often completed on the same day on which the biomarker data were collected (although some respondents returned the supplemental survey later), while the life satisfaction, psychological well-being, and general affect balance measures were collected in the main survey (prior to follow-up biomarker collection), and the diary

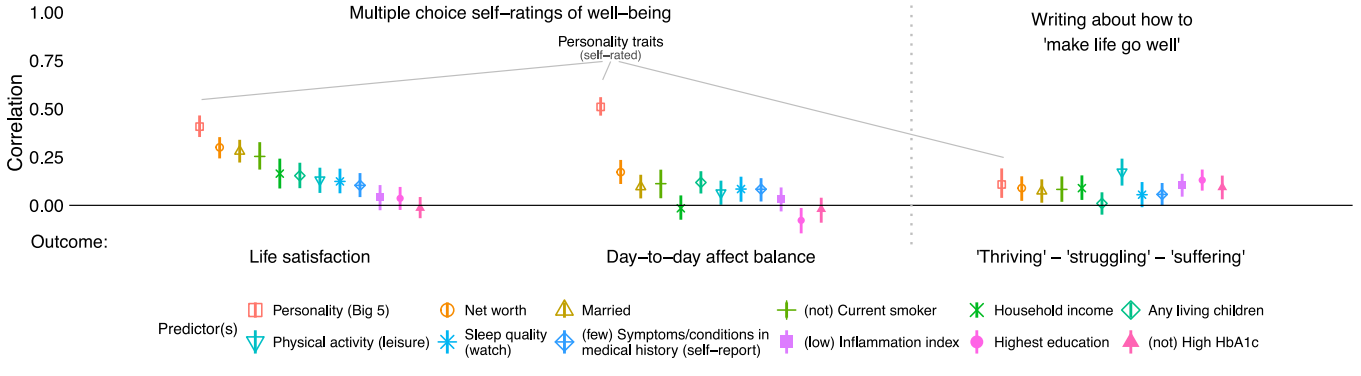

Fig. 1. Well-being associations for multiple-choice measures of life satisfaction and day-to-day affect versus a writing-based measure of "thriving." Correlation for personality traits represents adjusted multiple R (estimates are adjusted downward to account for the inclusion of more variables than comparisons). Error margins are bootstrapped 95% confidence intervals. N = 1,044.

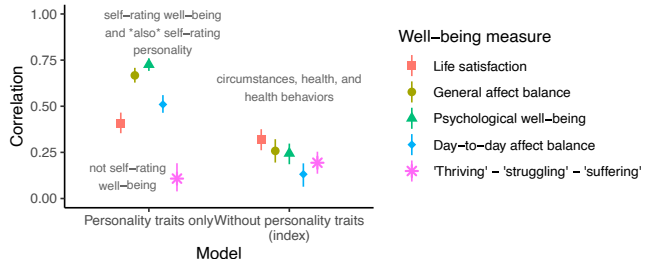

Fig. 2. Correlations for subjective personality traits versus circumstances and behaviors across measures of well-being. Closed-ended measures of well-being were strongly associated with other similarly rated closed-ended measures of personality traits, while each well-being measure, including the open-ended measure, was similarly associated with circumstances and behaviors. For the "without personality (index)" model, variables in Fig. 1, excluding personality traits, were summed after centering and standardization to SD units. We display the same finding for a non-indexed (multiple regression) version of this analysis in *SI Appendix*, Fig. S7; there, circumstances as associated with life satisfaction as personality (due to strong associations with marriage, having children, net worth, and not smoking specifically), but still much less strongly associated for other closed-ended measures of well-being. Correlation for personality traits represents adjusted multiple R (estimates are adjusted downward to account for the inclusion of more variables than comparisons). Error margins are bootstrapped 95% confidence intervals. N = 1,044.

data were also not collected concurrently with the biomarker data. If this drives the associations, then analyzing mortality risk over follow-up might be especially helpful in adjudicating to what extent the different measures are associated with health.

Well-Being and Mortality. Next, we consider associations between the well-being measures and mortality risk over follow-up. The findings support results from Fig. 1 suggesting that self-ratings of psychological traits overstate associations with (similarly rated) well-being.

Fig. 3 displays associations between the well-being measures and mortality risk over follow-up using Cox proportional hazard models (*Materials and Methods* for specification and control variable details). The first set of coefficients in each group displays all-cause mortality, and the second set displays death from cardiovascular disease. We analyze cardiovascular diseases separately because these are very common causes of death, for which we have sufficient power to analyze separately, and one which, based on substantial past research (30), we would expect to be consistently and bidirectionally related to well-being.

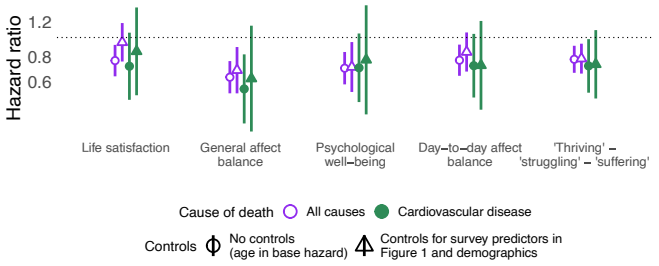

Fig. 3. Each well-being measure is similarly associated with mortality risk. Error margins are 95% confidence intervals. N = 1,044. The general affect balance appears to be somewhat more strongly associated with mortality risk over follow-up for all causes and with no controls for personality or circumstances; other coefficients for this measure are more imprecisely estimated. *SI Appendix*, Fig. S16 repeats this analysis for observations without diary data, with more precisely estimated associations with mortality risk in the larger sample.

These findings demonstrate that each well-being measure is associated with lower all-cause mortality over follow-up, though with somewhat varying levels of statistical significance. To assess whether the open-ended measure might contribute health information that is not reflected in other questions that might be easily collected by well-being researchers (i.e., outside of an extraordinary longitudinal study that was able to collect biomarker data), we also estimate mortality risk controlling for all the survey predictors in Fig. 1.

Overall, although the closed-ended measures are more strongly associated with personality and self-rated health, among other factors, they are not generally more strongly associated with mortality risk, even after controlling for measures of health status and behavior that can be collected through closed-ended questions.

Correlations Among Self-Ratings. As a final analysis comparing the zero-shot text labels to closed-ended responses, we consider to what extent well-being self-ratings (identified by searching for keywords in closed-ended response options; *SI Appendix*, section 11) contain duplicated information compared to the open-ended measure. For this, we test what fraction of the self-ratings in MIDUS were significantly associated with each of the well-being measures after corrections for multiple tests. These findings are displayed in Table 1.

This analysis suggests that a very large fraction of the survey self-ratings are strongly associated with the closed-ended well-being measures. The open-ended measure is significantly correlated (at $P < 0.05$—after the Bonferroni multiple testing correction, which here requires a raw P-value less than 0.05/1916) with only 1% of the self-ratings. In combination with the significant associations with measures that are not self-ratings and are instead more objectives measures of circumstances and health, this suggests that the open-ended measure provides unique well-being information compared to the self-ratings while the closed-ended measure contains often-repeated information across responses.

In *SI Appendix*, section 11, we display all self-rating associations for the writing-based well-being measure as well as the top 50 self-rating associations for life satisfaction and the day-to-day affect balance. Qualitatively, the associations suggest that the writing-based measure is relatively strongly related to the work situation, physical health self-ratings and activity

Table 1. Statistically significant associations ($P < 0.05$) with self-ratings after Bonferroni multiple testing corrections for 1,916 tests

Well-being measure	Statistically significant associations with **multiple-choice self-ratings and self-reported event frequencies** (q's duplicated across samples)
Life satisfaction	507 out of 1916 (26%)
General affect balance	768 out of 1916 (40%)
Psychological well-being	906 out of 1916 (47%)
Day-to-day affect balance	535 out of 1916 (28%)
Thriving - struggling - suffering	19 out of 1916 (1%)

Due to slight differences in question names and, in some cases, response options, each sample (MIDUS II, Milwaukee, and Refresher I) was analyzed separately.

reports, and survey responses about the participant's (in)ability to address problems they are facing in life than the multiple-choice measures.[#] Further, the open-ended measure does not appear to clearly represent any specific theoretical distinction across hedonic and eudaimonic conceptions of well-being. This should be expected. In such a broadly phrased open-ended response, respondents can choose to address whichever physical, emotional, financial, social, or spiritual aspects of their well-being feel most relevant and important to write about. This said, as we show in *SI Appendix*, Figs. S14 and S15, writing more or less about work and health in the open-ended responses does not drive the overall findings—across quartiles of work and health mentions, personality traits and circumstance are again similarly associated with "thriving"–"struggling"–"suffering."

LIWC and Supervised Models. A lingering question on these analyses might be that the results could be driven by a particular definition of living well (i.e., "thriving"–"struggling"–"suffering"), despite concordant results for hand labels (*SI Appendix*, section 4) and for each zero-shot label individually (*SI Appendix*, section 6). How much can the differences be explained by the open-ended responses themselves rather than the approaches to measuring well-being in text so far? For example, does a simpler dictionary-based approach, such as the widely used Linguistic Inquiry and Word Count (LIWC), reproduce these findings?

In *SI Appendix*, section 13, we show that these findings replicate with LIWC '22 positive minus negative tone in longer responses, of a corpus median length of 47 words or longer. Shorter responses have no association with either personality traits or circumstances. Consistent with past work (51), LIWC produces more reliable and predictive measures in texts containing more words. LIWC tone relies on a list of words that have been categorized as likely to be unambiguously positive or negative, but a single positive or negative word is often not a reliable indicator of the text's tone overall. We illustrate this level of noise in example texts (actual MIDUS responses cannot be publicly shared) in *SI Appendix*, section 3, where we show LIWC scores and their MIDUS percentiles, along with comparisons to the zero-shot thriving/struggling/suffering labels.

In considering effects of document length, we also repeated our prior analyses for responses less than the median length and the median length or longer. To what extent should future analyses expect to require longer texts, such as 50 words or more? Would open-ended analyses be just as reliable with a handful of words? In these analyses, we still find an equal or stronger association with circumstance than personality traits across longer and shorter texts, but associations for both are notably weaker in the shorter texts. Additionally, for the mortality risk analyses, the shorter open-ended measures were not significantly related to lower mortality risk over follow-up. However, we also observe an equally small and not significant mortality risk association for the closed-ended day-to-day affect balance measure. Because of this, and the much lower statistical power in the mortality risk analyses compared to other tests, we cannot unequivocally attribute variation in associations to less reliable measurement in shorter texts alone. Nonetheless, an important question for future work is whether additional open-ended prompts would improve measurement, as well as whether online surveys can be used to reliably measure well-being through text. Texts in online samples could be shorter than ones collected through phone interviews or through the collection of hand-written text. There may also be other quality concerns as responses to open-ended surveys are sometimes used to identify "low-quality" or inattentive respondents (52), although these issues may be more transparent in analyses of open-ended responses than closed-ended ones.

Next, we also test to what extent common method variance from multiple-choice responses might be present at all in text. If the open-ended responses do contain variation related to the strong associations among self-ratings—to the point that we are able to reproduce much larger associations for personality traits relative to behaviors and circumstances—then we can assess to what extent text features reflect well-being itself versus, perhaps, indicators of response biases. Here, it is important to reiterate that we intentionally avoided a supervised model in our previous tests. Such a model has the capacity to reproduce biases in closed-ended responses—and perfect predictions merely return the original self-ratings.

For that purpose, and with the above caveats, we train supervised models on the closed-ended responses, using only supervised models that produce interpretable features—keywords and zero-shot labels. We use the widely used elastic net regression (53) (i.e., a linear regression with penalty terms to avoid overfitting) with, first, the words used by at least five respondents as independent variables and, second, zero-shot labels for the top 1,000 words used by respondents, which is again approximately the number of words in the corpus used by at least five respondents. This model specification is explained in more depth in *Materials and Methods*. Only two of the eight models identified predictive variables when trained on the 1,044 respondent dataset—1,000 zero-shot labels predicting psychological well-being and life satisfaction.

Across the models, and arguably similar to past work using social media data to predict personality traits (54), text-based predictions do not capture most of the variation seen in closed-ended responses. In the best fitting model, fitted predictions are correlated with psychological well-being and life satisfaction at 0.33. This suggests that patterns of closed-ended self-ratings are not clearly represented in the open-ended responses and perhaps that more training data would be required to assess whether more subtle associations exist in the texts. However, more importantly, these predictions do still reproduce an equal association for personality traits relative to circumstance. While the association with personality traits is markedly reduced, the association with circumstance variables declines only marginally. This finding is consistent with the argument that a large fraction of the variation in self-ratings could reflect idiosyncratic closed-ended response styles and ones that are not manifested in the same respondents' written narratives about their well-being. Because the large difference in associations with personality traits versus circumstance for psychological well-being can be more plausibly captured from a relatively small training set than the smaller gap for life satisfaction, we display associations in the 523 respondent test set for the original psychological well-being measure compared to the text-based well-being predictions in Fig. 4. This test set consists of respondents whose data were not used in model training and who were not included in prior analyses because they did not participate in the diary studies. Predictions for life satisfaction, and their associations with auxiliary variables, were largely indistinguishable from psychological well-being predictions (*SI Appendix*, Fig. S18),

[#]These most strongly associated external problems ratings are "Really no way I can solve problems I have" and "Little control over things happen to me." Closed-ended measures are more strongly associated with these variables than open-ended measures, but closed-ended measures are strongly associated with most self-ratings.

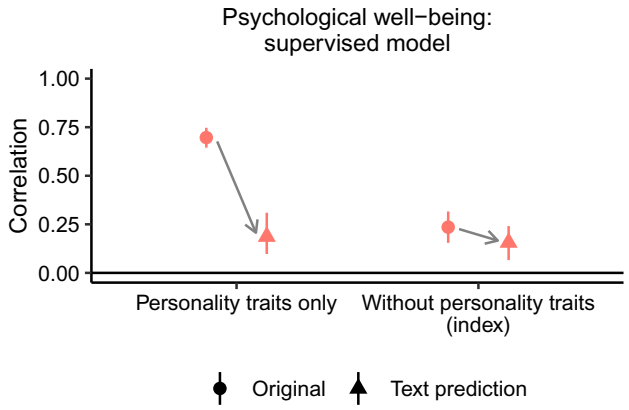

Fig. 4. Correlations, in the test set, for self-reported personality traits versus circumstances and behaviors for the original closed-ended measure of psychological well-being versus supervised text predictions based on 1,000 zero-shot labels. $N = 523$.

suggesting that these models captured similar source variation in the open-ended data, despite being trained on different dependent variables.

Further, although we largely fail to capture closed-ended variation related to self-ratings of personality traits, this analysis still provides some support for the claim that supervised models could capture patterns in text not necessarily reflecting well-being per se, such as strong associations with "Christian," "values," and "grateful" rather than "happiness" or "fulfillment" (for example, *SI Appendix*, Table S18). These predictors could mostly reflect the well-established importance of religiosity and gratitude for well-being (55, 56); however, a preponderance of these labels could also suggest that compliance with a social or religious norm against rating one's life too negatively might also have influenced self-ratings.

Discussion

Do self-reports of well-being matter outside of the survey context? Or do ratings in these self-reports reflect idiosyncratic survey behavior that might not reflect everyday experience? While comparisons across individuals will always be imperfect, additional measures of well-being (that do not only use a particular form of measurement) can help assess to what extent self-reports reflect experiences, behaviors, and circumstances in real life.

The inclusion of self-reports to measure well-being for use in public policy has been justified based on their associations with objective measures. Although associations for existing subjective measures are strong, they are nonetheless small relative to those collected within surveys, especially self-reported psychological traits that are rated in similar ways as measures of well-being. Existing measures of well-being, including life satisfaction and positive and negative affect (7), psychological well-being (8), and the more recently introduced psychological richness, are all strongly associated with self-reported personality traits (19, 57, 58)—and, in relative terms, typically much less strongly associated with circumstances and behaviors (19, 20).

Our findings here suggest that the difference in the strength of associations between self-reported traits and circumstances (as well as behaviors) may be due to similarities in survey response styles when respondents are asked to subjectively rate themselves and their lives on surveys, rather than a reflection of the strength of these associations outside of the survey context. Although personality traits are very strongly associated with well-being, they are not necessarily so much more strongly related than circumstances. To the extent that personality traits and well-being are interrelated, well-being and circumstances are just as tightly bound. From this, there is little reason to use the relative strength of observational associations between self-ratings of personality and well-being to argue that traits are likely to be a more promising intervention target than circumstances or behaviors. Organizations who have the specific capability to target individuals' circumstances to improve well-being should not necessarily consider their capability to be inferior to an organization with a greater ability or, perhaps more importantly, mandate (e.g., in an opt-in self-help intervention) to influence individuals' psychological traits.

We make this claim by measuring a form of subjectively reported well-being without using self-ratings—by using scores constructed from an open-ended survey question. This form of subjective well-being is similarly associated with more objective measures of wellness and much less strongly related to personality traits. While we do not argue that this open-ended approach reproduces existing conceptions of subjective well-being, it is important to point out that all existing measures of subjective well-being rely on closed-ended questions. Consistent with extensive past research (19, 57), all of the prominent measures evaluated here were very strongly associated with personality. Claims that measures are not strongly predicted by personality are made only relative to the extraordinarily strong relationships of other self-ratings.

Further, relatively little of the information about an individual's wellness was captured through closed-ended questions, even questions that asked about respondents' health behaviors. And controlling for health indicators and behaviors did not substantially alter associations with mortality risk. The open-ended approach adopted here appears to be a uniquely promising addition to the well-being and wellness studies repertoire—and, like existing subjective measures, it provides a view of well-being from the perspective of survey respondents. Including more, and more targeted, open-ended questions on future surveys could perhaps even provide more unique information about survey respondents than provided from tens or even hundreds of survey questions that ask respondents to rate themselves and their lives in a variety of ways.

Findings from the present study are of course relevant to classic and recurring debates in psychology about "person" versus "situation" (59, 60). However, perhaps counterintuitively, the findings do not strongly favor either side of this debate. The "weak" association between personality traits and behaviors was at one point used to argue that the situation was far more consequential. But here, in the same way that we underestimate the association between circumstances and well-being, we also underestimate associations between self-reported psychological traits and circumstances. Because of this, the findings more broadly suggest that behaviors and circumstances are likely to be more integral to both well-being and personality traits than might otherwise be appreciated.

With respect to the design of interventions, these findings on overstated associations between well-being and psychological traits do not speak against current trends toward personalized interventions in economics, public health, or in clinical settings. Experiments can of course customize interventions based on participants' personal circumstances (see, for example, ref. 61). Additionally, although the findings do suggest that using self-rated personality traits (or any single method measures) as

customization targets may lack diversity and risk conflating survey behaviors with the underlying traits themselves—to our knowledge—personalized interventions in psychology typically either allow substantial participant agency through partly self-tailored interventions (21, 62), which would not be constrained by closed-ended response styles, or leverage more complex within-person modeling approaches (63), which would also not tend to be sensitive to the cross-individual variation studied here. Instead, open-ended questions and preregistered measures in interventions, and measures that do not rely on training or codebook creation data from a pretest, could eventually be informative additions to experimental work and in largely the same ways that they would add diversity to longitudinal studies. These could, for example, be used to assess effect heterogeneity both within—and, perhaps especially usefully, across—studies or to check whether experimental effects might have been caused by shifts in closed-ended response styles alone.

Similarly, it is important to point out that we do not argue that open-ended survey questions should replace closed-ended ones, especially in aggregate (64), on tracking surveys (6), or in research that focuses on within-individual changes in well-being over time (65). This effort is also distinct from some prior work measuring well-being in text for tracking purposes. These studies have been intended to reproduce survey measures of well-being in large-N data and on more fine-grained temporal and geographic scales (66). The ultimate goal of these studies is to investigate changes in well-being which we would not expect to be strongly influenced (stable aspects of) by personality traits.

Also, these initial measures of well-being from text, which were assigned without any supervision or fine-tuning of an automated language model, appear to be somewhat less reliable than much more long-standing and refined measures of well-being from closed-ended survey questions, at least when those are scored from a long battery of closed-ended questions. Nonetheless, the text measure is particularly useful for evaluating the hypotheses here as it is a distinct measure of well-being and one that is not likely to be influenced by closed-ended response tendencies. All in all, the primary takeaway is that research output should not place excessive emphasis on a correlation between personality traits and well-being, especially in comparison to more objective measures that are not self-rated; closed-ended measures of well-being should still be used for many other research questions. Like for the refinement of closed-ended measures and the convergence on a small number of closed-ended measures, a longer period of research and community development/selection of text-based measures across contexts would be required to achieve comparable reliability and suitability for their general use.

It is possible that unknown, or at least unmeasured, psychological traits would better predict the open-ended well-being measure constructed here. And efforts to reproduce survey self-ratings of personality in text, which can to some extent reproduce closed-ended self-ratings (54, 67), could of course be expanded and primarily validated using theoretical predictions about nonsurvey personality associations, while allowing for more substantial divergence between closed- and open-ended measurement approaches. However, much like for subjective closed-ended survey responses, correlations between text-based response measures might also be inflated relative to correlations between dissimilar response formats.

Last, we anticipate that ongoing research in language modeling will enable continuous better measurement of well-being in text and without relying on training of algorithms to primarily reproduce closed-ended survey measures. But to truly capitalize on future methods, and to fully assess their strengths and weaknesses, we will need text data in high-quality, representative, longitudinal studies well before those technological advances. Because of this, it may be justified to begin to supplement closed-ended questions with more open-ended questions. Our analyses suggest that this will improve and diversify current research, while also greatly improving the value of well-being research in the future.

Materials and Methods

The following subsections describe the 3 MIDUS samples (II, Milwaukee, and Refresher I) and 3 MIDUS projects (main, biomarker, and diary) analyzed in this study, measures of well-being (closed-ended and open-ended), modeling specification (correlations, hazard models, and supervised models), and well-being predictor descriptions. This research project was reviewed and approved by the Cornell University Institutional Review Board (IRB0010653, exempt).

Midlife in the United States (MIDUS) Studies. The Midlife in the United States (MIDUS) study is a nationally representative longitudinal study begun in 1995. Participants in the study were recruited through random digit dialing. There have been three waves of the main study: 1995 to 1996 (MIDUS 1), 2004 to 2006 (MIDUS II), and 2013 to 2014 (MIDUS III). Data analyzed here were included in MIDUS II (68). Respondents were aged 25 to 74 during the first wave, and 35 to 86 in the second. The MIDUS study also included twin, metropolitan, and sibling samples, which are not included in this study.

In addition to the original study, two later samples were added to parallel MIDUS II and subsequent data collection. The MIDUS Milwaukee African American study began during the second wave of the original MIDUS in 2005 to 2006 (MIDUS Milwaukee). Participants were recruited through area-based stratified sampling. (See ref. 69 for more information.)

MIDUS also added the Refresher sample in 2011 through 2014 MIDUS Refresher 1 (70). This is a national probability sample to both replenish the original MIDUS study sample and to study respondents' perceived effects of the 2008-2009 economic recession. No diary data were collected from the corresponding Milwaukee African American Refresher sample, and this sample is not included in our analyses because of this.

MIDUS Biomarker Studies and Open-Ended Survey Samples. The open-ended survey question studied here was included in supplemental studies that collected biomarkers from the second wave of the MIDUS II, Milwaukee, and Refresher I samples. The open-ended survey question was included at the end of a supplemental survey.

National survey respondents who were living at the time of the biomarker study and "existing health information indicated an ability to travel to the clinic without excessive risk to the respondent or project staff" were eligible to participate in the biomarker data collection (71). (71) describes the full MIDUS II biomarker sample, including twin, city, and sibling samples.

In total, 637 MIDUS II (main RDD only), 195 Milwaukee (MIDUS II), and 735 Refresher I respondents participated in the biomarker studies (not all participated in the diary studies). These data were collected from 2004 to 2009 for MIDUS II and Milwaukee respondents (72) and 2012 to 2016 for Refresher I respondents (73). For more information on biomarker project data collection protocols, see refs. 71–73.

MIDUS Diary Studies. Finally, we use data from the MIDUS diary studies, called the National Study of Daily Experiences. These data were collected over eight consecutive days in 2004 to 2009, MIDUS II and Milwaukee (74) and 2012 to 2014, Refresher I (75). Our analyses use average values of responses across these 8 d (or the average across fewer days, for the 30% and 20% respectively of respondents who completed fewer than eight interviews).

In total, 569 MIDUS II (main RDD only), 132 Milwaukee (MIDUS II), and 343 Refresher I respondents participated in both the biomarker studies and the diary studies and also provided responses to both the open-ended question

and the closed-ended life satisfaction and day-to-day affect balance questions. These are the 1,044 respondents used in our primary analyses in the main text. The remaining 523 respondents were analyzed in a supplementary mortality risk analysis (*SI Appendix*, Fig. S16) and as a test set in the supervised model analysis (Fig. 4).

MIDUS Closed-Ended Well-Being Measures. MIDUS includes many self-ratings, including ratings used to construct various measures of well-being.

While many well-being measures produce similar associations, we focus on specific components of well-being to align with prior research that studies differences between global evaluations, such as life satisfaction ratings and happiness during experiences themselves (50).

For the life satisfaction rating, we use responses to the question "At present, how satisfied are you with your life? Very, somewhat, a little, or not at all?".

For the day-to-day affect balance, we use data from the diary study. Because the diary study did not ask whether respondents were "happy" (the closest question was "extremely happy," emphasis added), we use the MIDUS-provided positive and negative affect averages, which were in turn averaged across all 8 d of the diary period. Affect balance was positive minus negative affect. For these affect ratings, participants responded to the question "How much of the time today did you feel ... ?". The "day-to-day happiness index" comprises positive affect minus negative affect:

- positive affect–in good spirits, cheerful, extremely happy, calm and peaceful, satisfied, full of life, close to others, like you belong, enthusiastic, attentive, proud, active, confident
- negative affect–restless or fidgety, nervous, worthless, so sad nothing cheer you up, everything was an effort, hopeless, lonely, afraid, jittery, irritable, ashamed, upset, angry, frustrated

For psychological well-being, we use the average of the MIDUS II provided psychological well-being scales for autonomy (7 items), environmental mastery (7 items), personal growth (7 items), positive relations with others (7 items), purpose in life (7 items), and self-acceptance (7 items). These scales were summed and standardized to create the composite psychological well-being scale.

For general affect balance, we use the MIDUS II–provided positive and negative affect scales. These contained fewer adjectives than the provided index in the diary studies. Respondents responded to the question "During the past 30 d, how much of the time did you feel...." Affect balance was positive minus negative affect.

- positive affect–in good spirits, cheerful, extremely happy, calm and peaceful, satisfied, full of life
- negative affect–restless or fidgety, nervous, worthless, so sad nothing cheer you up, everything was an effort, hopeless

Open-Ended Well-Being Measure: Zero-Shot Classification. Because our goal is to evaluate an open-ended question rather than a particular technique for analyzing it, we choose an operationalization that can identify mentions of positive and negative well-being, but that has limited researcher degrees of freedom. In this, we attempt to target the "stance" of the responses rather than the sentiment of the responses (76); average expressed sentiments in the texts were almost entirely positive.

For this, we use the zero-shot classification approach proposed by Yin et al. (46) as implemented in the Python library (47), which uses language model BART (large-sized model) (77) after being trained on the MultiNLI dataset (78). Zero-shot classification is a task in natural language processing that involves labeling a dataset without any training examples–other than related tasks that have been "pretrained" using other, general language data. Roughly, this requires a machine to account for the grammar/ordering of words in a sentence and the approximate meanings of words, a form of which can be estimated and represented using a vector of numbers by using the shared contexts of words. More specifically, the pretrained BART language model used here was trained on texts in which spans of texts were masked and sentences rearranged, and the model's objective was to reconstruct the original text–meaning, to predict the hidden words and original sentence orderings based on the remaining context. This and related models succeed in this task by constructing complex representations of the relationships among words and their contexts. Next, the MultiNLI dataset training described in ref. 46 involves fine-tuning BART to answer questions about texts, specifically whether a piece of text (called the "premise") entails or does not entail a "hypothesis" about the text (and, for zero-shot classification, converting MultiNLI's third "neutral" label to nonentailment). Examples of these labeled texts can be found here: https://repeval2017.github.io/shared/. This process was the entirety of the model training. We did not further adjust the classification model ourselves.

Given a set of context-based word "meanings" and additional pretraining using a set of question and answer sets (i.e., external to our own data here as described above), these text models can answer basic questions about a text–for example, "Is this text about ___?". Using the meaning of ___, based on its context in general word use, as well as the meaning of a text being "about" it, the algorithm can assign a probability to whether the content of a text entails the hypothesis "This text is about ___".

In our scoring, we use an existing, content-free measure of well-being in a form that can be used with the machine learning technique outlined above. We require terms to fill in the blank "This text is about ___" and that reflect levels of well-being. Colloquial terms are important for this because they will be associated with well-being in general text on which the machine learning-based text scoring will be based. Without special tuning or training on academic samples (i.e., domain adaptation), the method is unlikely to accurately define terms that are strongly associated with well-being in academic language but used differently in everyday language.

For this scoring, we draw on terms used by Gallup[‖] in reports to summarize levels of well-being for the public. These summarize are used to explain locations on the Cantril self-anchoring striving scale (79), which Gallup uses in both its World Poll and its daily Gallup-Sharecare Well-Being Index. Gallup uses the words "thriving" to describe high well-being (high ladder-present and high ladder-future), "struggling" moderate well-being (mixed high-low for ladder-preset and ladder-future), and "suffering" and low well-being (low ladder-present and low ladder-future). However, thriving/struggling/suffering do not have these meanings in general, and our scoring will not reflect the specific meanings of these terms (i.e., placement on Cantril ladders) as used by Gallup.

Unlike dictionary approaches and some supervised algorithms, the process through which a score is assigned is not directly interpretable. Its validation instead relies on analysis of its output, such as the correlational analyses in the main text here, the keyword analysis described at the end of *Materials and Methods* (and shown in *SI Appendix*, section 2), and the scored example texts in *SI Appendix*, section 3.

Open-Ended Well-Being Measure: Hand Labeling and Validation. *SI Appendix*, Fig. S1 displays the labeling instructions and an example provided to research assistants when asked to code whether open-ended survey responses were about thriving, struggling, or suffering. Like the machine labels, the text responses could be coded in all categories–a text could be about thriving, struggling, and suffering. Although not used by Gallup (and so not analyzed in this study), we also included an additional category "striving," as research assistants initially perceived that there was a large gap between the thriving and struggling labels. This category is not scored or analyzed, and so it is no different from coding thriving, 1; striving, 0; and struggling: −1.

Models. Figs. 1 and 2 display the correlations for all variables other than personality traits and the square root of the adjusted R-squared from a linear regression for all big 5 personality traits combined. The square root of R-squared is the correlation between the model prediction and the data, and we use the adjusted R-squared to avoid assigning higher values to regressions that merely include more variables (as R-squared will generally increase with the inclusion of more predictors). Confidence intervals for both the adjusted R-squared values and (for consistency with R-squareds) the correlations were calculated using bootstrapping (1,000 replicates) and the percentile method.

[‖] https://news.gallup.com/poll/122453/understanding-gallup-uses-cantril-scale.aspx.

Mortality risk estimates were estimated using Cox proportional hazard models with age in the base hazard. Age was the respondent's age at the time of interview–the main interview date for life satisfaction, the first day of the diary study for the day-to-day happiness index, and the day of biomarker collection for the open-ended survey response. Following past work on mortality risk and well-being (80), respondents who died within 2 y of the interview date were excluded from these analyses. Complete mortality data were available through the end of 2020 and follow-up for living respondents ended in December 2020. Controls were included separately in this model in an identical format as included in Fig. 1 (Below for more details). Schoenfeld residuals for each model indicated that proportional hazard assumptions were met. "Demographic" predictors/controls were gender (indicator for "Female") and race (indicators for "Black and/or African American"). Note that, in these MIDUS samples, most respondents reporting race "Black and/or African American" were in the Milwaukee African-American sample; this sample (from Milwaukee) is not nationally representative, unlike the MIDUS II and Refresher I samples. Because of this, the race variable is similar to including an indicator for the Milwaukee African-American sample.

To ensure that each analysis was based on the same sample, all models are restricted to respondents participating in all studies containing the three well-being measures, "Life satisfaction," "Day-to-day affect balance," and "'Thriving'–'struggling'–'suffering"'–respondents who participated in the main interview, diary study, and the biomarker studies (1,044 respondents). This prevents the findings from being driven by sampling differences.

In analyses for Fig. 3, after subsetting to complete "Life satisfaction," "Day-to-day affect balance," and "'Thriving'–'struggling'–'suffering"' observations, we use multiple imputation by chained equations (20 imputations) and report pooled estimates and confidence intervals across models for each imputed dataset. In *SI Appendix*, Fig. S16, we repeat this procedure for observations with missing values of "Life satisfaction" or "Day-to-day affect balance." In Figs. 1 and 2, we conduct the same process with a single imputation. All variables used in analyses producing Figs. 1, 2, and *SI Appendix*, Fig. S16 were included in this imputation process. The most missing variables were sleep quality (watch) (62%) and positive log wealth (14%). All other variables in these analyses were less than 5% missing, and most were less than 1% missing.

SI Appendix, Table S8 displays all coefficients from analyses with controls in Fig. 3. There were 1,044 respondents, 132 deaths from all causes, and 34 deaths from cardiovascular disease in these models. There were 1,567 respondents, 166 deaths from all causes, and 42 deaths from cardiovascular disease for the models shown in *SI Appendix*, Fig. S16, except for "Day-to-day affect balance" (for which sample size did not appreciably increase).

All analyses other than the self-rating analysis were conducted on a pooled MIDUS sample, combining data from MIDUS II, MIDUS Milwaukee, and MIDUS Refresher I.

Fig. 1 Predictor Descriptions. *Personality* was measured by asking respondents to what extent adjectives described them. Big 5 personality comprises the following:

- agreeableness–helpful, warm, caring, softhearted, sympathetic,
- extraversion–outgoing, friendly, lively, active, talkative,
- neuroticism–moody, worrying, nervous, calm (reverse coded),
- conscientiousness–organized, responsible, hardworking, careless (reverse coded), and
- openness to experience–creative, imaginative, intelligent, curious, broad-minded, sophisticated, adventurous.

Married was modeled using a marital status of "married" rather than "divorced," "separated," "widowed," or "never married."

Any living children is any living children relative to no living children.

Net worth was constructed from the participant's responses to the questions: "Suppose you (and your spouse or partner) cashed in all of your checking and savings accounts, stocks and bonds, real estate, and sold your home, your vehicles, and all of your valuable possessions. Then suppose you put that money toward paying off your mortgage and all of your other loans, debts, and credit cards. Would you have any money left over after paying your debts or would you still owe money?". And "How much would that be (that you had left over, or would owe)?". Net worth was logged ($log(x+1)$ and, for simplicity, values less than 0 set to 0.

Household income ($log(x+1)$) was the total household income from wages, pensions, Social Security, and other government sources.

Counts of medical history symptoms and conditions were self-reported in the MIDUS Biomarker projects. These counts were logged ($log(x+1)$) and reverse-coded (i.e., few symptoms/conditions) in analyses. Respondents were asked about the following symptoms/conditions: heart disease, high blood pressure, blood clots, heart murmur, TIA or stroke, anemia or other blood disease, cholesterol problems, diabetes, asthma, emphysema/COPD, tuberculosis, positive TB skin test, thyroid disease, peptic ulcer disease, cancer, colon polyp, arthritis, glaucoma, cirrhosis/liver disease, alcoholism, depression, and blood transfusion before 1993. In addition, this count of symptoms/conditions also includes any listed under the two "Other, please specify" options. This does not include the number of chronic conditions reported in the main surveys, which, for many of the questions, asked more generally about "problems" or "troubles" with aspects of respondents' physical and emotional health.

Physical activity (leisure) is an index of 6 leisure activity self-reports, with separate answers for light, moderate, and vigorous physical activity, each of which was split by activity in the summer or winter (e.g., vigorous physical activity during winter is one question, and vigorous physical activity during summer is another). These self-reports were averaged to form the physical activity index. Leisure activity stands in contrast to job or chore physical activities, which were also asked in the survey.

Highest education was self-reported using 12 levels of highest education. These were converted to a numeric format for the analyses in Fig. 1.

The Sleep quality (watch) predictors use data recorded by an Actiwatch® activity monitor. Variables used for the sleep quality model were sleep onset latency (time required to fall asleep, logged), sleep efficiency (percentage of time in bed spent sleeping), and sleep time (amount of time scored as sleep by Actiware). These variables were summed after being centered and scaled to standard deviations. This variable was available for 38% of respondents in our 1,044 participant analysis sample.

MIDUS documentation lists CRP (C-reactive protein), ICAM (intracellular adhesion molecule), IL6 (interleukin 6), sIL6r (soluble IL-6 receptor), fibrinogen, and E-selectin as its primary set of inflammation markers. These markers–logged, centered and scaled to SD, and summed–comprise the inflammation index.

High HbA1c was zero for HbA1c less than 5.7, one for HbA1c greater than or equal to 5.7 and less than 6.5, and two for greater than or equal to 6.5.

Cardiovascular disease was defined as an ICD-10 cause of death code in the range C00 through C99.

Survey predictors are all predictors in Fig. 1 that can be recorded through standard survey questions. This excludes the biomarkers (low) inflammation index and (not) high A1C, as well as the actigraphy-based sleep quality (watch) variable.

MIDUS Self-Ratings. "Self-ratings" in the MIDUS studies were identified by searching the response options of questions in the surveys of the main interview, diary, and biomarker projects. The search terms for this analysis are included in *SI Appendix*, section 11, along with the response options identified from these search terms. In addition, we display in *SI Appendix*, Tables S12, S13, and S14 the questions most strongly associated (i.e., significantly associated at $P < 05$ after a Bonferroni multiple testing correction for 1,916 tests) with life satisfaction, the day-to-day happiness index, and the "thriving"–"struggling"–"suffering" open-ended measure.

Correlations were calculated for each sample separately because question names, and, at times, question response options varied across datasets. Questions with fewer than 100 complete responses in a sample were excluded from this analysis. The histogram in *SI Appendix*, Fig. S17 displays the sample sizes across these tests.

Linguistic Inquiry and Word Count. To assess whether a dictionary-based sentiment analysis would replicate the zero-shot and hand label findings, we use the LIWC '22 positive and negative tone dictionaries (48). Text scores were assigned using LIWC software. These scores were the fraction of positive or

negative tone words in a text, and our well-being measure was positive minus negative tone.

Supervised Models. The goal of our supervised model analyses was to identify words and labels that predicted the closed-ended well-being measures and that could reflect either well-being or some other confounding variable, such as social desirability bias. Because the interpretability of these models was important–otherwise we would not be able to assess whether a supervised model was detecting underlying well-being or response bias in closed-ended self-ratings–we ran supervised models on the document-term matrices and also on 1,000 zero-shot labels. These supervised models returned coefficients for the predictive keywords or labels. We use the zero-shot labels in addition to the words themselves because the open-ended responses are relatively short and many of the words in the corpus were used by only a few respondents. Many respondents use different words with similar meanings, and the language model underlying the zero-shot labels is able to capture these similar meanings across different words. Note that a supervised model that only uses a language model, such as a supervised prediction using only Bidirectional Encoder Representations from Transformers (BERT) (81), does not provide interpretable predictions. To avoid surfacing words used by only a very small number of respondents, which could introduce some risk of reidentification, we considered only words used by at least five respondents.

To train on the closed-ended responses, we use the widely used elastic net regression (53) (a penalized regression), as implemented in ref. 82. The elastic net regression specification, an alpha of 0.01, we use is one very close to a ridge regression (a linear regression with an $l2$ penalty) but still incorporates a small lasso ($l1$) penalty for variable selection. This specification not only selects variables most strongly related to the outcome (if there are any strongly predictive variables across cross-validation folds) but also allows for highly correlated variables to predict the outcome reliably and in tandem. Highly correlated features will possess similar coefficients. A lasso specification (alpha of 1) identified no predictive variables for any closed-ended responses. The other elastic net hyperparameter λ was selected through cross-validation using the default settings of the "glmnet" R package (82). Note that a ridge regression can be viewed as a version of principal component regression (83) that leverages covariation in words in addition to words' association with the outcome. Although this regression could more loosely be seen as similar to running a topic model, e.g., latent Dirichlet allocation (84) or a structural topic model (85) and then a supervised model, the elastic net avoids that multiple stage modeling and also potential topic instability in a small training set.

Across these models, we used the main analysis set (1,044 respondents who took part in all 3 studies: main, biomarker, and diary) as the training set and the remaining analysis data as the test set (523 respondents who took part in 2 studies: main and biomarker but not diary). Only two of these models identified predictive variables when trained on the analysis dataset–1,000 zero-shot labels predicting psychological well-being and life satisfaction. We were able to produce fitted models for zero-shot labels when expanding the training set to all data for general affect balance, and we display predictive zero-shot labels in *SI Appendix*, section 12; however, this larger training set left us with no test data.

In addition to using supervised models to predict closed-ended survey responses, we also used them to identify keywords associated with each of the three other open-ended well-being measures–zero-shot classification, hand labels, and LIWC tone. These keywords are displayed in *SI Appendix*, section 2. It is important to recognize that these keywords are not themselves used to score documents–they are single words that are associated with the overall document scores. For example, the word "the" appears more often in documents with lower LIWC and lower hand label scores, but it is not a negative word and its inclusion in a text does not itself lead to lower scores.

Data, Materials, and Software Availability. All closed-ended data for MIDUS II and MIDUS Refresher I data are publicly available and accessible through the Inter-university Consortium for Political and Social Research (ICPSR) online portal. Some data for the MIDUS Milwaukee African American Sample require signing of a data use agreement – this data and the data use agreement are also made available through ICPSR. Researchers interested in studying the MIDUS open-ended survey responses should email the MIDUS Help Desk (midus_help@aging.wisc.edu) for guidelines on requesting access to that data (68–70, 72–75).

ACKNOWLEDGMENTS. We thank Jailyn Wilson, Fatima Al-Sammak, Chase Agheli, and Isabella Zhi for their work in the hand labeling of the open-ended survey responses. We are also grateful to Karl Pillemer and MIDUS conference participants for helpful feedback on this project. This research was supported by a Cornell Center for Social Sciences seed grant. Data used for this research were provided by the longitudinal study titled "Midlife in the United States," (MIDUS) managed by the Institute on Aging, University of Wisconsin. MIDUS was supported by a grant from the National Institute on Aging (P01-AG020166).

Author affiliations: [a]Department of Psychology, Cornell University, Ithaca, NY 14853; and [b]Department of Government, Cornell University, Ithaca, NY 14853

1. J. Ludwig *et al.*, Neighborhood effects on the long-term well-being of low-income adults. *Science* **337**, 1505–1510 (2012).
2. T. J. VanderWeele, On the promotion of human flourishing. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 8148–8156 (2017).
3. J. F. Helliwell, L. B. Aknin, Expanding the social science of happiness. *Nat. Hum. Behav.* **2**, 248–252 (2018).
4. A. Steptoe, Happiness and health. *Ann. Rev. Public Health* **40**, 339–359 (2019).
5. D. Buettner, T. Nelson, R. Veenhoven, Ways to greater happiness: A delphi study. *J. Happiness Stud.* **21**, 2789–2806 (2020).
6. J. C. Eichstaedt *et al.*, The emotional and mental health impact of the murder of George Floyd on the US population. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2109139118 (2021).
7. E. Diener, Subjective well-being. *Psychol. Bull.* **95**, 542–575 (1984).
8. C. D. Ryff, Beyond Ponce de Leon and life satisfaction: New directions in quest of successful ageing. *Int. J. Behav. Dev.* **12**, 35–55 (1989).
9. R. T. Howell, M. L. Kern, S. Lyubomirsky, Health benefits: Meta-analytically determining the impact of well-being on objective health outcomes. *Health Psychol. Rev.* **1**, 83–136 (2007).
10. P. L. Hill, N. A. Turiano, Purpose in life as a predictor of mortality across adulthood. *Psychol. Sci.* **25**, 1482–1486 (2014).
11. A. Steptoe, A. Deaton, A. A. Stone, Subjective wellbeing, health, and ageing. *Lancet* **385**, 640–648 (2015).
12. N. Martín-María *et al.*, The impact of subjective well-being on mortality: A meta-analysis of longitudinal studies in the general population. *Psychos. Med.* **79**, 565–575 (2017).
13. E. Puterman *et al.*, Predicting mortality from 57 economic, behavioral, social, and psychological factors. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 16273–16282 (2020).
14. T. B. Kashdan, R. Biswas-Diener, L. A. King, Reconsidering happiness: The costs of distinguishing between hedonics and eudaimonia. *J. Positive Psychol.* **3**, 219–233 (2008).
15. D. J. Disabato, F. R. Goodman, T. B. Kashdan, J. L. Short, A. Jarden, Different types of well-being? A cross-cultural examination of hedonic and eudaimonic well-being. *Psychol. Assess.* **28**, 471–482 (2016).
16. S. Margolis, E. Schwitzgebel, D. J. Ozer, S. Lyubomirsky, *Empirical Relationships Among Five Types of Well-Being* (Oxford University Press, 2021), pp. 377–407.
17. C. D. Ryff, Well-being with soul: Science in pursuit of human potential. *Perspect. Psychol. Sci.* **13**, 242–248 (2018).
18. T. W. Smith, M. Davern, J. Freese, S. L. Morgan, *General Social Surveys, 1972–2018: Cumulative Codebook* (NORC, Chicago, 2019).
19. J. Anglim, S. Horwood, L. D. Smillie, R. J. Marrero, J. K. Wood, Predicting psychological and subjective well-being from personality: A meta-analysis. *Psychol. Bull.* **146**, 279–323 (2020).
20. S. Margolis, J. Elder, B. Hughes, S. Lyubomirsky, What Are the Most Important Predictors of Subjective Well-Being? Insights From Machine Learning and Linear Regression Approaches on the MIDUS Datasets, (PsyArXiv), Preprint (2021).
21. M. Stieger *et al.*, Changing personality traits with the help of a digital personality change intervention. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2017548118 (2021).
22. T. F. Bainbridge, S. G. Ludeke, L. D. Smillie, Evaluating the Big Five as an organizing framework for commonly used psychological trait scales. *J. Pers. Soc. Psychol.* **122**, 749–777 (2022).
23. B. W. Roberts, N. R. Kuncel, R. Shiner, A. Caspi, L. R. Goldberg, The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspect. Psychol. Sci.* **2**, 313–345 (2007).
24. E. Diener, E. Sandvik, W. Pavot, D. Gallagher, Response artifacts in the measurement of subjective well-being. *Soc. Indic. Res.* **24**, 35–56 (1991).
25. H. S. Friedman, M. L. Kern, Personality, well-being, and health. *Ann. Rev. Psychol.* **65**, 719–742 (2014).
26. E. Diener, R. Lucas, *Personality and Subjective Well-Being in The Science of Well-Being, Social Indicators Research Series*, E. Diener, A. C. Michalos, Eds. (Springer Netherlands, Dordrecht, 2009), vol. 37.
27. R. E. Lucas, F. Fujita, Factors influencing the relation between extraversion and pleasant affect. *J. Pers. Soc. Psychol.* **79**, 1039–1056 (2000).
28. E. Diener, R. E. Lucas, S. Oishi, Advances and open questions in the science of subjective well-being. *Collabra: Psychol.* **4**, 15 (2018).

29. P. M. Podsakoff, S. B. MacKenzie, J. Y. Lee, N. P. Podsakoff, Common method biases in behavioral research: A critical review of the literature and recommended remedies. *J. Appl. Psychol.* **88**, 879–903 (2003).
30. J. K. Boehm, L. D. Kubzansky, The heart's content: The association between positive psychological well-being and cardiovascular health. *Psychol. Bull.* **138**, 655–691 (2012).
31. A. Steptoe, A. Deaton, A. A. Stone, Subjective wellbeing, health, and ageing. *Lancet* **385**, 640–648 (2015).
32. D. Gerstorf *et al*., Terminal decline in well-being: The role of social orientation. *Psychol. Aging* **31**, 149–165 (2016).
33. D. Kahneman, A. Deaton, High income improves evaluation of life but not emotional well-being. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 16489–16493 (2010).
34. B. Stevenson, J. Wolfers, Subjective well-being and income: Is there any evidence of satiation? *Am. Econ. Rev.* **103**, 598–604 (2013).
35. M. A. Killingsworth, Experienced well-being rises with income, even above $75,000 per year. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2016976118 (2021).
36. B. Headey, M. Wooden, The effects of wealth and income on subjective well-being and ill-being*. *Econ. Rec.* **80**, S24–S33 (2004).
37. P. Oreopoulos, K. G. Salvanes, Priceless: The nonpecuniary benefits of schooling. *J. Econ. Perspect.* **25**, 159–184 (2011).
38. J. Cuñado, F. P. de Gracia, Does education affect happiness? Evidence Spain*Soc. Indic. Res.* **108**, 185–196 (2012).
39. N. Powdthavee, W. N. Lekfuangfu, M. Wooden, What's the good of education on our overall quality of life? A simultaneous equation model of education and life satisfaction for Australia. *J. Behav. Exp. Econ.* **54**, 10–21 (2015).
40. H. K. Kim, P. C. McKenry, The relationship between marriage and psychological well-being: A longitudinal analysis. *J. Family Issues* **23**, 885–911 (2002).
41. A. Stutzer, B. S. Frey, Does marriage make people happy, or do happy people get married? *J. Soc.-Econ.* **35**, 326–347 (2006).
42. M. Luhmann, R. E. Lucas, M. Eid, E. Diener, The prospective effect of life satisfaction on life events. *Soc. Psychol. Pers. Sci.* **4**, 39–45 (2013).
43. S. K. Nelson, K. Kushlev, T. English, E. W. Dunn, S. Lyubomirsky, In defense of parenthood: Children are associated with More Joy Than Misery. *Psychol. Sci.* **24**, 3–10 (2013).
44. S. K. Nelson, K. Kushlev, S. Lyubomirsky, The pains and pleasures of parenting: When, why, and how is parenthood associated with more or less well-being? *Psychol. Bull.* **140**, 846–895 (2014).
45. R. Ryan, S. Booth, A. Spathis, S. Mollart, A. Clow, Use of salivary diurnal cortisol as an outcome measure in randomised controlled trials: A systematic review. *Ann. Behav. Med.* **50**, 210–236 (2016).
46. W. Yin, J. Hay, D. Roth, Benchmarking Zero-shot Text classification: Datasets evaluation and entailment approach. *EMNLP* (2019).
47. A. S. Maiya, Ktrain: A low-code library for augmented machine learning. *J. Mach. Learn. Res.* **23**, 1–6 (2022).
48. R. L. Boyd, A. Ashokkumar, S. Seraj, J. W. Pennebaker, *The Development and Psychometric Properties of LIWC-22* (University of Texas at Austin, Austin, TX, 2022).
49. M. T. Lee, L. D. Kubzansky, T. J. VanderWeele, *Measuring Well-Being: Interdisciplinary Perspectives from the Social Sciences and the Humanities* (Oxford University Press, 2021).
50. D. Kahneman, A. B. Krueger, Developments in the measurement of subjective well-being. *J. Econ. Perspect.* **20**, 3–24 (2006).
51. J. C. Eichstaedt *et al*., Closed and open vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychol. Methods* **26**, 398–427 (2021).
52. J. Ziegler, A text-as-data approach for using open-ended responses as manipulation checks. *Political Anal.* **30**, 289–297 (2022).
53. H. Zou, T. Hastie, Regularization and variable selection via the elastic net. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **67**, 301–320 (2005).
54. G. Park *et al*., Automatic personality assessment through social media language. *J. Pers. Soc. Psychol.* **108**, 934–952 (2015).
55. T. B. Smith, M. E. McCullough, J. Poll, Religiousness and depression: Evidence for a main effect and the moderating influence of stressful life events. *Psychol. Bull.* **129**, 614–636 (2003).
56. A. M. Wood, J. J. Froh, A. W. Geraghty, Gratitude and well-being: A review and theoretical integration. *Clin. Psychol. Rev.* **30**, 890–905 (2010).
57. K. M. DeNeve, H. Copper, The happy personality: A meta-analysis of 137 personality traits and subjective well-being. *Psychol. Bull.* **124**, 197–229 (1998).
58. S. Oishi, E. C. Westgate, A psychologically rich life: Beyond happiness and meaning. *Psychol. Revi.* **129**, 790–811 (2021).
59. W. Mischel, Y. Shoda, A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychol. Rev.* **102**, 246–268 (1995).
60. W. Mischel, Y. Shoda, Reconciling processing dynamics and personality dispositions. *Ann. Rev. Psychol.* **49**, 229–258 (1998).
61. P. Bergman *et al*., "Creating moves to opportunity: Experimental evidence on barriers to neighborhood choice" (Tech. rep., National Bureau of Economic Research, 2019).
62. N. W. Hudson, R. C. Fraley, Volitional personality trait change: Can people choose to change their personality traits? *J. Pers. Soc. Psychol.* **109**, 490–507 (2015).
63. J. Burger *et al*., A clinical PREMISE for personalized models: Toward a formal integration of case formulations and statistical networks. *J. Psychopathol. Clin. Sci.* **131**, 906 (2022).
64. A. J. Oswald, S. Wu, Objective confirmation of subjective measures of human well-being: Evidence from the U.S.A. *Science* **327**, 576–579 (2010).
65. M. Luhmann, W. Hofmann, M. Eid, R. E. Lucas, Subjective well-being and adaptation to life events: A meta-analysis. *J. Pers. Soc. Psychol.* **102**, 592–615 (2012).
66. K. Jaidka *et al*., Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 10165–10171 (2020).
67. A. Cutler, D. M. Condon, *Deep Lexical Hypothesis: Identifying personality structure in natural language* (PsyArXiv, Preprint, 2022). https://arxiv.org/abs/2203.02092. Accessed 1 April 2022.
68. C. D. Ryff *et al*., Midlife in the United States (MIDUS 2), 2004-2006. Inter-university Consortium for Political and Social Research [distributor], 2021-09-15. 10.3886/ICPSR04652.v8. Accessed 22 June 2021.
69. C. D. Ryff *et al*., Midlife in the United States (MIDUS 2): Milwaukee African American Sample, 2005-2006. Inter-university Consortium for Political and Social Research [distributor], 2022-10-12. 10.3886/ICPSR22840.v6. Accessed 27 June 2022.
70. C. D. Ryff *et al*., Midlife in the United States (MIDUS Refresher 1), 2011-2014. Inter-university Consortium for Political and Social Research [distributor], 2017-11-20. 10.3886/ICPSR36532.v3. Accessed 10 February 2022.
71. G. Dienberg Love, T. E. Seeman, M. Weinstein, C. D. Ryff, Bioindicators in the MIDUS national study: Protocol, measures, sample, and comparative context. *J. Aging Health* **22**, 1059–1080 (2010).
72. C. D. Ryff, T. Seeman, M. Weinstein, "Midlife in the United States (MIDUS 2): Biomarker project, 2004–2009" in *Inter-university Consortium for Political and Social Research* (National Archive of Computerized Data on Aging, 2022), https://doi.org/10.3886/ICPSR29282.v10.
73. M. Weinstein, C. D. Ryff, T. E. Seeman, "Midlife in the United States (MIDUS Refresher 1): Biomarker project, 2012–2016" in *Inter-university Consortium for Political and Social Research* (National Archive of Computerized Data on Aging, 2019), https://doi.org/10.3886/ICPSR36901.v6.
74. C. D. Ryff, D. M. Almeida, "Midlife in the United States (MIDUS 2): Daily stress project, 2004–2009" in *Inter-university Consortium for Political and Social Research* (National Archive of Computerized Data on Aging, 2017), https://doi.org/10.3886/ICPSR26841.v2.
75. C. D. Ryff, D. M. Almeida, "Midlife in the United States (MIDUS Refresher 1): Daily diary project, 2012–2014" in Inter-university Consortium for Political and Social Research (National Archive of Computerized Data on Aging, 2020), https://doi.org/10.3886/ICPSR37083.v2.
76. S. E. Bestvater, B. L. Monroe, Sentiment is not stance: Target-aware opinion classification for political text analysis. *Political Anal*. 1–22 (2022). 10.1017/pan.2022.10.
77. M. Lewis *et al*., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2020), pp. 7871–7880.
78. A. Williams, N. Nangia, S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference" in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (Association for Computational Linguistics, New Orleans, Louisiana, 2018), pp. 1112–1122.
79. H. Cantril, *The Pattern of Human Concerns* (Rutgers University Press, New Brunswick, NJ, 1965).
80. A. D. Ong, A. Steptoe, Association of positive affect instability with all-cause mortality in older adults in England. *JAMA Netw. Open* **3**, e207725 (2020).
81. J. Devlin, M. W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding" in *Proceedings of NAACL-HLT* (2018), pp. 4171–4186.
82. J. H. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
83. T. Hastie, R. Tibshirani, J. J. H. Friedman, *The Elements of Statistical Learning* (Springer, New York, 2001).
84. D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
85. M. Roberts, B. Stewart, D. Tingley, Structural topic models for open-ended survey responses. *Am. J. Polit. Sci.* **58**, 1064–1082 (2014).