

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Shearer, NJ;Gillespie, AN;Olds, TS;Mensah, FK;Edwards, B;Fernando, JW;Wang, Y;Wake, M;Lycett, K

Title:

Insights into adolescent well-being from computerised analysis of written language

Date:

2021-06-01

Citation:

Shearer, N. J., Gillespie, A. N., Olds, T. S., Mensah, F. K., Edwards, B., Fernando, J. W., Wang, Y., Wake, M. & Lycett, K. (2021). Insights into adolescent well-being from computerised analysis of written language. *Acta Paediatrica International Journal of Paediatrics*, 110 (6), pp.1880-1889. <https://doi.org/10.1111/apa.15813>.

Persistent Link:

<https://hdl.handle.net/11343/298334>

KATE LYCETT (Orcid ID : 0000-0002-8988-4038)

Article type : Regular Article

Insights into adolescent wellbeing from computerized-analysis of written language

Natalie J. Shearer (MBBS),^{1,2,3} Alanna N. Gillespie (BSc Hons),^{3,4} Tim S. Olds (PhD),^{3,5} Fiona K. Mensah (PhD),^{3,4,6} Ben Edwards (PhD),⁷ Julian W. Fernando (PhD),² Yichao Wang (PhD),^{3,4} Melissa Wake (MD),^{3,4,8} Kate Lycett (PhD).^{3,6}

Affiliations: ¹Royal Melbourne Hospital, Parkville, VIC, Australia; ²Deakin University, Geelong, VIC, Australia; ³Murdoch Children's Research Institute, Parkville, VIC, Australia, ⁴University of Melbourne, Parkville, VIC, Australia; ⁵Alliance for Research in Exercise, Nutrition and Activity, University of South Australia, Adelaide, South Australia, Australia; ⁶The Royal Children's Hospital, Parkville, VIC, Australia; ⁷Australian National University Centre for Social Research and Methods, Canberra, Australian Capital Territory, Australia; and ⁸The Liggins Institute, The University of Auckland, Auckland, New Zealand.

Correspondence: Kate Lycett, Murdoch Children's Research Institute, The Royal Children's Hospital, Flemington Road, Parkville VIC 3052, Australia. E-mail: kate.lycett@mcri.edu.au

Short title: Language-use patterns and wellbeing in adolescents

Word count: 3,534

Abstract

Aim: To examine associations between patterns of language use and early adolescent wellbeing.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1111/APA.15813](https://doi.org/10.1111/APA.15813)

This article is protected by copyright. All rights reserved

Methods: Participants were 1,763 Australian 11-12 year-olds in the Child Health CheckPoint. Six patterns of language use were identified from a writing activity using Linguistic Inquiry and Word Count and factor analysis: *Acting in the present and future*, *Positive emotion*, *Gender and relationships*, *Self-aware*, *Inquisitive and time focused*, and *Confident*. Wellbeing measures represented a spectrum from negatively to positively-framed psychosocial health. Associations between language use and wellbeing were estimated using linear regression adjusted for age, sex and social disadvantage.

Results: *Positive emotion* (high emotional tone, positive emotion) was associated with better general wellbeing (standardized regression coefficient (SRC) 0.05; 95% confidence interval 0.00 to 0.11; $p=0.04$), life satisfaction (0.06; 0.01 to 0.11; $p=0.03$), psychosocial health (0.07; 0.02 to 0.12; $p=0.01$), and QoL (0.06; 0.01 to 0.11; $p=0.02$). Similarly, *Self-aware* (high first person singular pronouns, authentic, low clout) was associated with better general wellbeing, life satisfaction and psychosocial health (SRC 0.05, 0.09, 0.08), but *Confident* (high clout, first person plural pronouns, affiliation) was associated with worse life satisfaction, psychosocial health and QoL (SRC -0.06, -0.09, -0.06).

Conclusion: If replicated in 'real-world' settings (e.g. social media), language patterns could provide naturalistic insights into early adolescents' wellbeing.

Keywords: Automated computerized analysis; language patterns; linguistic analysis; mental health; Wellbeing.

Key notes:

- This Australian cross-sectional population-based study of 11-12 year-olds examined patterns of written language and their associations with wellbeing.
- Written language patterns, identified through automated computerized analysis, were consistently associated with the spectrum of adolescent wellbeing.
- Our findings pave the way to explore the utility of automated language-based analyses in 'real world' settings to offer insights into adolescent's wellbeing.

1 INTRODUCTION

2 Automated analysis of natural language can offer novel insights into adults' inner psyche (e.g.
3 social relationships, emotions, and cognitive processes).¹⁻³ Such a tool could be particularly
4 useful during adolescence, when mental health problems often emerge but go undetected, and
5 when there is a reluctance to engage in prevention and treatment interventions.^{4,5} Equally,
6 being able to identify thriving adolescents could assist in identifying protective factors for a 'good
7 life'.⁶ At the population level, health professionals and researchers often rely on self-reported
8 screening measures to assess wellbeing.⁷ However, such measures often ask overt questions
9 about symptoms of anxiety and depression and issues such as bullying that some children may
10 not openly share.⁷ Analysis of natural language from various sources (eg social media posts,
11 diary entries, free-writing activities and school work), offer an alternative way to gather insights
12 into adolescents' inner psyches, but free-text responses require time-consuming analysis.
13 Technological advances, such as computer-based analysis of natural written language, could
14 provide a novel platform to gain insights into adolescents' underlying wellbeing without specific
15 enquiry. This could lead to the development of automated detection methods to assist in
16 identifying not only individuals who require additional assessment and treatment, but also those
17 who could benefit from targeted preventative strategies.^{8,9}

18 In small laboratory studies, the use of more self-related words (e.g. "I"), negative emotions and
19 lower rates of positive emotion words have been shown to distinguish adults with poor mental
20 health.^{3,10,11} Similarly, a review of 12 studies that took computerized language analysis into the
21 'real-world' showed that certain language patterns captured via social media predicted mental
22 illness with moderate-to-strong accuracy.⁸ Schwartz and colleagues⁶ recently extended these
23 findings using a 'big data' approach through public Facebook accounts. They identified four
24 language topics that predicted optimal life satisfaction: positive affect and psychological states;
25 relationships; high-value occupations; and communal engagement. In addition, swearing and
26 disengagement predicted poor life satisfaction. Such a tool could be particularly useful in
27 adolescence given the abundance of access to natural written language through their near-
28 universal adoption of social media and tendency to share more online than off.^{9,12}

29 However, there is a dearth of research in adolescence. The transition from small laboratory-
30 style studies to 'big data' studies in adults using social media has not yet occurred for
31 adolescents. This may reflect a lack of large well-phenotyped samples that have both the ability
32 to control for potential confounders and 'gold standard' measures of mental health and
33 wellbeing. Australia's Child Health CheckPoint (CheckPoint) study presented a unique
34 opportunity to fill this gap. Thus, we aimed to investigate how wellbeing in 11-12 year-olds
35 varies according to identified language patterns of written language captured through a pen-
36 and-paper free writing activity, which required adolescents to mentally travel into one's future
37 and imagine the life they would be living at age 25.

METHODS

Study design and participants

The CheckPoint study was a one-off biophysical assessment offered to 11–12-year-olds taking part in the Birth cohort of the population-based Longitudinal Study of Australian Children (LSAC). Details of the CheckPoint are outlined elsewhere.¹³ In brief, LSAC recruited a nationally representative sample of 5,107 0-1 year-olds in 2004, and has since followed them up in biennial data collection 'waves'. The initial proportion recruited was 57.2%, of whom 73.7% (n=3,764) were retained at wave 6 in 2014. At wave 6, 3,513 families consented to their contact details being shared with CheckPoint. These families were invited to take part in a CheckPoint assessment between February 2015 and March 2016, with 54% (n=1,874) of families participating.

Ethics and consent

The Australian Institute of Family Studies Ethics Committee (14-26) and The Royal Children's Hospital Human Research Ethics Committee (33225) approved the CheckPoint study. Informed written consent for the child's participation was provided by the attending parent/caregiver.

Procedure

CheckPoint families completed a single assessment rotating through a series of stations either at a 'pop-up' Main (3.5 hours) or Mini (2.5 hours) Assessment Centre. Families who were unable to attend an Assessment Center were offered a home assessment (1.5 hours) with a subset of portable measures. Adolescents completed the free writing activity and standardized wellbeing scales at these assessments.

Measures

Written language

Adolescents completed a free writing activity at the '*Life at 25*' station during a 15-minute period, which was based on the 1958 UK Birth Cohort's storytelling activity.¹⁴ This required children to mentally travel into one's future and imagine the life they would be living at age 25. After giving the child a single piece of lined A4 paper and a pen, the researcher read aloud the prompt at the top of the page, "*Imagine you are now 25 years old. Write about the life you are leading, your interests, study, your home life and your work at the age of 25.*"

As described elsewhere,¹⁵ stories were professionally transcribed in two versions: 'raw' and 'marked-up' version. We used the 'marked-up' version, where transcribers corrected spelling, punctuation and grammatical errors, which allowed for computer recognition of words and linguistic features.

Self-reported wellbeing

Adolescents self-reported the standardised wellbeing measures via a survey completed on iPad. Widely-used, validated measures were selected to represent the wellbeing spectrum and for their feasibility in a population-based study (full details Supplementary Table 2). Negatively-framed wellbeing was captured by *psychosocial health* (15-item Pediatric Quality of Life 4.0 Psychosocial Health Summary Score)¹⁶ and *health-related quality of life* (9-item Child Health Utility-9D Questionnaire; CHU9D).¹⁷ Positively-framed wellbeing was captured by *general wellbeing* (6-item Pediatric Quality of Life General Wellbeing Scale)¹⁸ and *life satisfaction* (5-item International Survey of Children's Wellbeing Brief Multidimensional Students' Life Satisfaction Scale).¹⁹ Higher scores on all measures indicate better wellbeing.

Potential confounders of the association between language use and wellbeing were identified *a priori*. Given that language use and wellbeing has been shown to differ by sex and age, these were considered as potential confounders.^{14,20} Neighbourhood disadvantage was also considered, as differences in socioeconomic status are strongly associated with variations in both vocabulary/language use²¹ and mental health.⁵ We used the Index of Relative Socioeconomic Disadvantage to measure neighborhood disadvantage (national mean 1000 and standard deviation (SD) of 100, where higher values represent less disadvantage).²²

Statistical analyses

Characterisation of patterns of written language (Aim 1)

We used the computer-based Linguistic Inquiry and Word Count (LIWC2015)² analysis software to categorize and quantify word use by checking each word in a text file against an internal dictionary of almost 6,400 words, word stems and emoticons. Each word recognized by the LIWC2015 dictionary is assigned into language variable categories. For instance, the dictionary for the 'positive emotion' language variable includes 620 words (e.g. "kind" and "proud"). Language variables are quantified as the number of times a word in the dictionary is used, as a percentage of the total number of words used.

The initial aim was to characterize recognizable factors, or language patterns, within the free-text samples of written language. Exploratory factor analysis with varimax rotation was used to identify key factors explaining variability in the adolescents' use of the LIWC2015 language variables.² We included all written language samples with at least 17 words, which was the average length of one sentence in the free writing activity. Language variables were excluded before factor analysis if they met any of the following criteria:

- They were represented via another language variable (14 variables excluded). For example, large percentages of positive and negative emotion words were also captured in the language variable 'affect words'. Affect words, in this case, were not included.

- They did not refer to meanings or features of specific words, such as punctuation categories (16 variables excluded).
- They showed minimal variation in our sample (2 variables excluded).

Table 1 details the LIWC2015 language variables that were excluded and those that were considered for subsequent factor analysis; it also provides example words from the LIWC2015 dictionary for each variable. The exceptions are four summary LIWC2015 variables (analytical thinking, authentic, clout, emotional tone) that use proprietary dictionaries so example words are not provided. However, these are largely self-explanatory, with the exception of 'clout'.³ A high clout level suggests an author is confident and speaking from an area of high expertise, whereas low clout suggests a more anxious, humble and tentative style.³

As each language variable was ordinal, polychoric correlation matrices were used for the factor analysis.²³ Orthogonal and oblique rotations of factors yielded similar findings so the orthogonal rotation was used. Individual language variables with a Kaiser-Meyer-Olkin (KMO)²⁴ value <0.4 were removed as these variables cause diffused correlation patterns (3 variables excluded). The overall KMO statistic was >0.6, confirming it was suitable to perform factor analysis on the dataset as it produced distinct and reliable factors.²⁴

To determine how many factors (patterns) were retained, the following criteria were applied: (1) eigenvalue >1; (2) visual inspection of the scree-plot for the point of inflection; and (3) two (highly correlated variables, correlation >0.7) or more variables contributing to the one factor.^{23,24} Only language variables with a factor loading of ≥ 0.4 or ≤ -0.4 were included in each factor. In addition, to check collinearity within the regression models, we used the variance inflation factor (VIF), which quantifies the severity of multicollinearity in an ordinary least squares regression analysis.

Associations of patterns of written language with wellbeing (Aim 2)

In order to fit regression models, patterns of written language were examined as "predictors" (explanatory variables) and measures of wellbeing as "outcomes".

Unadjusted and adjusted linear regression models were used to assess associations, with sequential Bonferroni adjustments applied. Wellbeing measures were internally standardized to have a mean of 0 and standard deviation of 1. Standardized regression coefficients (SRC) represent the SD difference in wellbeing measures according to a one SD unit higher language pattern. To check assumptions of linearity, we visually inspected Lowess plots between each language pattern and wellbeing outcome, followed by formal tests to see whether quadratic associations were evident. There was no evidence of non-linear associations based on visual inspections and formal univariable tests (all $p > 0.1$). Univariable linear regression analysis was

141 performed and residuals were inspected. Given that residuals suggested a ceiling effect, a Tobit
142 regression was performed, which showed similar results to linear regression models.²⁵ Taking
143 this and our large sample size into account, we proceeded with linear regression models.²⁵
144 As LIWC recommends caution when interpreting samples with less than 50 words, a final
145 sensitivity analysis was conducted excluding such samples.²

Author Manuscript

146 RESULTS

147 Sample characteristics

148 Written language samples of at least 17 words (the average length of one sentence in the free
149 writing activity) were provided by 1,763 (94%) adolescents, of whom 1,718 (92%) had at least
150 one wellbeing measure available and were included in our analytic sample. The sample came
151 from slightly less disadvantaged and more homogeneous neighborhoods (mean 1023; SD 61)
152 than the Australian population as a whole (mean 1000; SD 100).²²

153 On average, adolescents were aged 11.5 years (SD 0.5; 50% boys) (Table 2). Girls written
154 language samples contained more words than boys (79 vs 62). Wellbeing scores were similar
155 for boys and girls and were compared to existing norms where available. Adolescents had
156 slightly lower (poorer) psychosocial health scores but slightly higher (better) health-related QoL
157 than the age-appropriate population norms.^{18,26}

158 Patterns of language used classified within written text (*Aim 1*)

159 Of the 92 LIWC2015 variables, 60 were retained for subsequent factor analyses. Our final six
160 factor solution (Table 3) contained 19 language variables. The six factors were named
161 according to language variables with the strongest factor loadings, as follows: (1) *Acting in the*
162 *present and future* – low analytical thinking and high first person singular pronouns, auxiliary
163 and common verbs, and words that focused on the present (eg now, say) and on the future (eg
164 soon, tonight); (2) *Positive emotion* – high emotional tone and positive emotion; (3) *Gender and*
165 *relationships* – high third person singular pronouns, family, female and male; (4) *Self-aware* –
166 high first person singular pronouns and authentic, and low clout; 5) *Inquisitive and time focused*
167 – high common adverbs, interrogatives and time; and 6) *Confident* – high clout, first person
168 plural pronouns and affiliation. Table 4 provides examples of stories that loaded highly in each
169 factor. Our final six-factor solution accounted for 14.0% of the total variance explained in
170 observed language use.

171 Patterns of language and wellbeing (*Aim 2*)

172 Since unadjusted and adjusted (age, sex and neighborhood disadvantage) results were similar,
173 we present adjusted models (Table 5).

174 Three of the six identified language factors were consistently associated with wellbeing across
175 the spectrum. *Positive emotion* was associated with better wellbeing across all four measures.
176 Each additional standard deviation in *positive emotion* was associated with slightly better
177 general wellbeing (Standardised Regression Coefficient (SRC) 0.05, 95% confidence interval
178 (CI) 0.003 to 0.11, p=0.04), life satisfaction (0.06, 0.01 to 0.11, p=0.03), psychosocial health
179 (0.07, 0.02 to 0.12, p=0.01), and general health-related QoL (0.06, 0.01 to 0.11, p=0.02).
180 Similarly, higher *self-aware* levels were associated with slightly better wellbeing across three of

the four measures (SRC ranging from 0.05 to 0.09). Associations between language use and health-related QoL were also consistently observed although these were weaker than for other outcomes (SRC 0.04, $p=0.12$).

Conversely, *Confident* language patterns were associated with worse wellbeing across three of the four mental health measures: life satisfaction (SRC -0.06, 95% CI -0.11 to -0.01, $p=0.03$), psychosocial health (-0.09, -0.14 to -0.04, $p=0.001$) and general health-related QoL (-0.06, -0.11 to -0.003, $p=0.04$). There is no evidence in our study showing that the remaining three language patterns (*Acting in the present and future*, *Gender and relationships* and *Inquisitive and time focused*) were associated with adolescent wellbeing.

Sequential Bonferroni adjustments suggested that the associations of *Self-aware* with life satisfaction, and of *Confident* with psychosocial health, were the most robust. Similar associations remained after excluding samples with <50 words (Supplementary Table 1), although some attenuated slightly.

VIF tests confirmed that collinearity between the six language patterns was not artificially inflating our estimated regression effects. For example, all VIFs were less than 1.5, well below the threshold of 10 used to indicate multicollinearity.

199 **DISCUSSION**

200 **Principal findings**

201 This is one of the first population-based studies to examine associations of patterns of written
202 language with psychological wellbeing in young adolescents. Using computer-based linguistic
203 analysis software, we identified six patterns of written language, of which three were
204 consistently associated with wellbeing spanning the negative-positive continuum. Language
205 patterns associated with greater wellbeing were characterized by the high use of positive
206 emotion and emotional tone words, authentic and first person singular pronouns and low clout.
207 Those associated with worse wellbeing were characterized by high clout, first person plural
208 pronoun and affiliation words.

209 While the reported effects were quite small, these findings suggest that computer-based
210 analysis of short excerpts of written language may provide useful insights into adolescent
211 wellbeing. A next step could be to identify the type of tasks (e.g. unprompted vs. prompted and
212 length of text) and settings (e.g. online, school and/or clinical settings) that produce the most
213 robust results (i.e. larger effects with reasonably high sensitivity and specificity) for
214 understanding young people's wellbeing. For example, if online settings (e.g. Facebook and
215 Twitter) showed robust results, these platforms could add background algorithms based on
216 natural language use to provide free evidence-based online support services. Similarly, clinical
217 settings could consider whether written tasks from young patients could be uploaded to a
218 computer-based screening model prior to, or during appointments, as an opportunistic way to
219 tap into wellbeing.

220 **Findings in light of previous research**

221 The magnitude of associations between language patterns and wellbeing were small (i.e. SRC:
222 0.05 to 0.09). However, associations were highly consistent across wellbeing measures
223 providing confidence in the results. A review by Guntuku and colleagues⁸ reported that
224 associations between mental *illness* and language use patterns were moderate to strong.
225 However, this could be explained by the fact that 10 of their 12 studies oversampled clinical
226 groups (often comprising half of the entire sample) for whom one would expect much larger
227 effects than in our community sample of mostly healthy children.

228 Findings from our study suggest better wellbeing is consistently associated with emotional tone
229 and positive emotion words in adolescents. This is in contrast to the adult literature, which has
230 shown mixed results between wellbeing and the use of positive emotion words.^{27,28} For
231 example, Kahn and colleagues²⁷ reported positive emotions words correlated with self-rated
232 amusement after watching a comedy clip, but not with positive mood. Tov and colleagues²⁸
233 reported positive emotion words predicted concurrent, retrospective, and global self-reported

wellbeing in daily (n=206) but not twice-weekly (n=139) open-ended diary responses. Our study is the first to report the *Self-aware* (i.e. high first person singular pronouns and authentic and low clout) language factor. However, aspects of this language factor are characterized in previous studies. For example, poorer mental health outcomes in adults have been characterized by high use of first person singular pronouns.^{1,3,29} Similarly, low levels of clout have also been linked to more humble, tentative or anxious traits in adults.³⁰

In contrast, our child study demonstrated slightly *better* wellbeing with the high use of first person singular pronouns and low clout. One surprising finding was the association between the high use of a *Confident* language factor (i.e. high clout, first person plural pronouns and affiliation) and slightly worse wellbeing. High use of clout, first person plural pronouns and affiliation words have been associated with confidence and social connectedness, characteristics normally considered protective patterns for mental health and wellbeing.^{4,5,30}

The language factor *Acting in the present and future, Gender and relationships and Inquisitive and time focused* are similar to patterns described in the adult literature.³¹ However, we extend these previous findings by also examining their relationship with positive wellbeing.

Strengths and limitations

Strengths of this study include its large, national, population-based sample enhancing the generalizability of the use of automated linguistic analysis as a marker of wellbeing in the Australian adolescent population. Our widely used and well-validated measures of wellbeing provide a comprehensive picture across the negative-positive spectrum of wellbeing, although not into the pathological end of the spectrum. Furthermore, the consistent associations of language patterns with both negatively-framed and positively-framed wellbeing measures enhance confidence in our findings, despite the different recall windows. Our well-phenotyped sample enabled adjustment for potential confounders known to impact on both language and wellbeing.

Limitations include under-representation of very disadvantaged groups. It seems likely that similar findings would hold for older adolescents, but we cannot be sure of this. In addition, CheckPoint was not a natural setting, and may have over-elicited positive language by its positively-framed 'Life at 25' prompt and its fun and comfortable space, with adolescents reporting high levels of enjoyment in taking part.¹³ While it is possible that the knowledge that someone else would read their written story may have influenced the likelihood that participants would disclose their true feelings, it was free from, for example, the immediate peer feedback that may influence social media interactions. The short, context-specific story requirements could have constrained the content and style of language used.

In addition, the writing task required future thinking, a cognitive ability that in itself has been shown to be impaired with those with mental health conditions,³² and could bias results (eg those with poor wellbeing may have lacked the ability to write much). As the Life@25 data are progressively linked with the LSAC's rich longitudinal data, future studies can explore how the writing task varied by mental health conditions and academic ability.

Despite factor analysis being widely used and accepted to identify latent patterns, it does not involve any theoretical input, and this can result in complex factors where unexpected variables are grouped together – making interpretation and factor naming challenging. *Acting in the present and future* was such a factor, containing six variables, some of which do not logically go together. For example, the combination of 'present focus' and 'future focus' words is not intuitive, as these are often distinct characteristics of specific mental disorders (e.g. ADHD, characterised by the inability to think into the future and plan, while depression is often marked by the inability to be present). The sheer complexity of this factor may explain why it was not associated with wellbeing. It may also be that more detailed and longer tasks with different types of prompts could be needed to differentiate within complex constructs such as *Acting in the present and the future*.

Finally, our findings rely on the fixed group of words available in each conceptual LIWC dictionary. Whilst words in each dictionary are statistically related to each other they may encompass a number of different meanings (e.g. mad about you). Thus, pre-existing LIWC dictionaries should not be understood as perfect descriptors, but as reasonably reliable clustering of related words. Despite these limitations, it appears that taking a more tentative, humble, authentic and positive approach to one's projections for the future is indicative of better adolescent mental health than the use of more confident language and references to groups/others.

Conclusions

Patterns of written language can reveal insights into adolescents' wellbeing. Our findings pave the way to explore the utility of automated language-based analyses to identify and predict wellbeing in a 'real world' setting. Future directions include a 'big data' population-based study to identify associations between adolescent language use in a natural setting (e.g. social media) and wellbeing. Eventually, findings such as ours could potentially contribute to large scale, automated screening initiatives to detect at-risk individuals, who could then be provided with further support and targeted interventions.

Acknowledgements:

This article uses unit record data from Growing Up in Australia, the Longitudinal Study of Australia Children. The study is conducted in partnership between the Department of Social

Services (DSS), the Australian Institute of Family Studies (AIFS) and the Australian Bureau of Statistics (ABS). The authors thank the LSAC and CheckPoint study participants, staff and students for their contributions.

Abbreviations:

CheckPoint: Child Health CheckPoint; LSAC: Longitudinal Study of Australian Children; LIWC2015: Linguistic Inquiry and Word Count 2015; QoL: Quality of life; CHU9D: Child Health Utility 9D; Disadvantage Index: Index of Relative Socioeconomic Disadvantage; SD: Standard deviation; KMO: Kaiser-Meyer-Olkin; SRC: Standardized regression coefficients; CI: Confidence interval; p: p-value; n: number of participants.

Statements of conflict of interest and of funding:

The authors declare no potential conflicts of interest, including no specific financial interests relevant to the subject of this manuscript. The Murdoch Children's Research Institute (MCRI) administered the research grant for the study and provided infrastructural support to its staff but played no role in the conduct or analysis of the study. Research at the MCRI is supported by the Victorian Government's Operational Infrastructure Support Program. This work was supported by the National Health and Medical Research Council (NHMRC) of Australia [grant numbers 1041352, 1109355]; The Royal Children's Hospital Foundation [grant number 2014-241]; the MCRI, The University of Melbourne, the National Heart Foundation of Australia [grant number 100660]; Financial Markets Foundation for Children [grant numbers 2014-055, 2016-310]; and the National Centre for Longitudinal Data (NCLD). KL was supported by an NHMRC Early Career Fellowship [APP1091124]; and a National Heart Foundation Postdoctoral Fellowship [grant number 101239]. MW was supported by an NHMRC Principal Research Fellowship [grant number 1160906]. FM was supported by an NHMRC Career Development Fellowship [grant number 1111160]. The funding bodies had no role in relation to the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Financial Disclosure: There are no financial relationships relevant to this article to disclose.

References

1. Cohn MA, Mehl MR, Pennebaker JW. Linguistic markers of psychological change surrounding September 11, 2001. *Psychological science*. 2004;15(10):687-693.
2. *Linguistic Inquiry and Word Count: LIWC2015* [computer program]. Austin, Texas: Pennebaker Conglomerates; 2015.
3. Tausczik YR, Pennebaker JW. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Lang Soc Psychol*. 2010;29(1):24-54.
4. Das JK, Salam RA, Lassi ZS, et al. Interventions for Adolescent Mental Health: An Overview of Systematic Reviews. *Journal of Adolescent Health*. 2016;59(4):S49-S60.
5. Patton GC, Sawyer SM, Santelli JS, et al. Our future: a Lancet commission on adolescent health and wellbeing. *The Lancet*. 2016;387(10036):2423-2478.
6. Schwartz HA, Sap M, Kern ML, et al. Predicting individual well-being through the language of social media. Paper presented at: Biocomputing 2016: Proceedings of the Pacific Symposium2016.
7. Burns JM, Durkin LA, Luscombe GM, Hickie IB, Davenport TA. The internet as a setting for mental health service utilisation by young people. In: Australasian Medical Publishing Company Limited; 2010:22.
8. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*. 2017;18:43-49.
9. Park M, Cha C, Cha M. Depressive moods of users portrayed in Twitter. Paper presented at: Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD)2012.
10. Rude S, Gortner E-M, Pennebaker J. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*. 2004;18(8):1121-1133.
11. Pennebaker JW. *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury USA; 2013.
12. Lenhart A, Duggan M, Perrin A, Stepler R, Rainie H, Parker K. *Teens, social media & technology overview 2015*. Pew Research Center [Internet & American Life Project]; 2015.
13. Clifford SA, Davies S, Wake M. Child Health CheckPoint: cohort summary and methodology of a physical health and biospecimen module for the Longitudinal Study of Australian Children. *BMJ Open*. 2019;9(Suppl 3):3-22.

14. Elliott J, Morrow V. Imagining the Future: preliminary analysis of NCDS essays written by children at age 11. In: Centre for Longitudinal Studies, Institute of Education, University of London; 2007.
15. Clifford S, Davies A, Gillespie A, et al. Longitudinal Study of Australian Children's Child Health CheckPoint Data User Guide. Murdoch Children's Research Institute. Published 2020. Accessed.
16. Varni JW, Limbers CA, Burwinkle TM. How young can children reliably and validly self-report their health-related quality of life?: An analysis of 8,591 children across age subgroups with the PedsQL™ 4.0 Generic Core Scales. *Health and quality of life outcomes*. 2007;5(1):1.
17. Stevens K, Ratcliffe J. Measuring and valuing health benefits for economic evaluation in adolescence: an assessment of the practicality and validity of the Child Health Utility 9D in the Australian adolescent population. *Value in Health*. 2012;15(8):1092-1099.
18. Varni JW, Seid M, Kurtin PS. Pediatric health-related quality of life measurement technology: a guide for health care decision makers. *JCOM-WAYNE PA*-. 1999;6:33-44.
19. Strózik D, Strózik T, Szwarc K. The Subjective Well-Being of School Children. The First Findings from the Children's Worlds Study in Poland. *Child Indicators Research*. 2016;9(1):39-50.
20. Kern ML, Eichstaedt JC, Schwartz HA, et al. From 'Sooo excited!!!' to 'So proud': Using language to study development. *Developmental Psychology*. 2014(1):178.
21. Radesky JS, Carta J, Bair-Merritt M. The 30 million-word gap: Relevance for pediatrics. *JAMA Pediatrics*. 2016;170(9):825-826.
22. Australian Bureau of Statistics. *Census of population and housing: Socio-Economic Indexes for Areas (SEIFA) 2011*. 2011.
23. Costello AB, Osborne JW. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research & evaluation*. 2005;10(7):1-9.
24. Tabachnick BG, Fidell LS. *Using multivariate statistics*. Allyn & Bacon/Pearson Education; 2007.
25. Long JS. *Regression models for categorical and limited dependent variables*. Thousand Oaks : Sage Publications, c1997.; 1997.
26. Chen G, Ratcliffe J, Olds T, Magarey A, Jones M, Leslie E. BMI, health behaviors, and quality of life in children and adolescents: a school-based study. *Pediatrics*. 2014;peds. 2013-0622.

27. Kahn JH, Tobin RM, Massey AE, Anderson JA. Measuring emotional expression with the Linguistic Inquiry and Word Count. *American Journal of Psychology*. 2007(2):263.
28. Tov W, Ng KL, Lin H, Qiu L. Detecting well-being via computerized content analysis of brief diary entries. *Psychological assessment*. 2013;25(4):1069.
29. Settanni M, Marengo D. Sharing feelings online: studying emotional well-being via automated text analysis of Facebook posts. *Frontiers in psychology*. 2015;6:1045.
30. Kacewicz E, Pennebaker JW, Davis M, Jeon M, Graesser AC. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*. 2014;33(2):125-143.
31. Nagarajan M, Hearst MA. An Examination of Language Use in Online Dating Profiles. Paper presented at: ICWSM2009.
32. Sonuga-Barke EJS, Cortese S, Fairchild G, Stringaris A. Annual Research Review: Transdiagnostic neuroscience of child and adolescent mental disorders – differentiating decision making in attention-deficit/hyperactivity disorder, conduct disorder, depression, and anxiety. *Journal of Child Psychology and Psychiatry*. 2016;57(3):321-349.

Table 1. Description of LIWC2015 language variables considered in factor analysis³

Language Variable	Example words	No. words	Included in factor analysis	Language Variable	Example words	No. words	Included in factor analysis*
Word count	-	-	No ^C	Perceptual processes	look, hear	436	No ^B
Summary variable				See	gaze, dark	126	Yes
Analytical thinking ^A	-	-	Yes	Hear	hear, clap	93	Yes
Authentic ^A	-	-	Yes	Feel	cold, grip	128	Yes
Clout ^A	-	-	Yes	Biological processes	eat, pain	748	No ^B
Emotional tone ^A	-	-	Yes	Body	teeth, spine	215	Yes
Language Metrics				Health	alive, pain	294	Yes
Words per sentence	-	-	No ^C	Sexual	lust, lover	131	Yes
Words > 6 letters	-	-	No ^C	Ingestion	cook, yum	184	Yes
Dictionary words			No ^C	Drives	bully, friend	1103	No ^B
Function words	No, I, very	491	No ^B	Affiliation	marry, met	248	Yes
Total pronouns	I, them	153	No ^B	Achievement	rank, succeed	213	Yes
Personal pronouns	I, her	93	No ^B	Power	boss, owner	518	Yes
1 st person singular pronoun	I, me,	24	Yes	Reward	add, benefit	120	Yes
1 st person plural pronoun	us, we	12	Yes	Risk	alarm, dread	103	Yes
2 nd person pronoun	you, your	30	No ^E	Time orientation			
3 rd person singular pronoun	him, her,	17	Yes	Past focus	past, worn	341	No ^E
3 rd person plural pronoun	them, they	11	Yes	Present focus	now, say	424	Yes
Impersonal pronouns	another, it	59	Yes	Future focus	soon, tonight	97	Yes
Articles	a, an, the	3	Yes	Relativity	bend, exit	974	No ^B
Prepositions	about, away	74	Yes	Motion	walk, zoom	325	Yes
Auxiliary verbs	am, are, be	141	Yes	Space	little, up	360	Yes
Common adverbs	about, fully	140	Yes	Time	back, begin	310	Yes
Conjunctions	also, but	43	No ^E	Personal concerns			
Negations	never, no	62	Yes	Work	retail, hire	444	Yes
Common verbs	do, hit, sell	1000	Yes	Leisure	bar, opera	296	Yes
Common adjectives	far, rich	764	Yes	Home	bed, garage	100	Yes
Comparisons	after, worse	317	Yes	Money	bill, sell	226	Yes
Interrogatives	how, what	48	Yes	Religion	soul, sin	174	Yes
Numbers	first, ten	36	Yes	Death	bury, die	74	Yes
Quantifiers	add, extra	77	Yes	Informal language	lol, hmm	380	No ^B
Affect words	happy, sad	1393	No ^B	Swear words	dumb, darn	131	No ^D
Positive emotion	proud, kind	620	Yes	Netspeak	brb, fav	209	Yes
Negative emotion	annoy, rude	744	Yes	Assent	ok, woo	36	Yes
Anxiety	worry, fear	116	No ^B	Nonfluencies	ah, er, um	19	Yes

Language Variable	Example words	No. words	Included in factor analysis	Language Variable	Example words	No. words	Included in factor analysis*
Anger	hate, kill	230	No ^B	Fillers	anyway, blah	14	No ^D
Sadness	cry, sad	136	No ^B	All punctuation	-	-	No ^C
Social words	mate, they	756	No ^B	Periods	-	-	No ^C
Family	cousin, dad	118	Yes	Commas	-	-	No ^C
Friends	ally, date	95	Yes	Colons	-	-	No ^C
Female references	girl, female	124	Yes	Semicolons	-	-	No ^C
Male references	boy, fiancé	116	Yes	Question marks	-	-	No ^C
Cognitive processes	know, cause	797	No ^B	Exclamation marks	-	-	No ^C
Insight	accept, clue	259	Yes	Dashes	-	-	No ^C
Causation	change, infer	135	Yes	Quotation marks	-	-	No ^C
Discrepancy	could, if	83	Yes	Apostrophes	-	-	No ^C
Tentative	bet, fairly	178	Yes	Parentheses (pairs)	-	-	No ^C
Certainty	clear, ever	113	Yes	Other punctuation	-	-	No ^C
Differentiation	differ, neither	81	Yes				

^A Summary language variables created from prior linguistic analysis, where dictionaries are unavailable. Higher analytic thinking is associated with confidence and expertise and lower analytic thinking with a humble, tentative or anxious style; higher authentic with honest and personal text and lower analytic with a guarded and distant discourse; higher clout with hierarchical, logical or formal thinking and lower with informal, narrative and personal thinking; and higher emotional tone with positive emotion and lower with anxiety, hostility or sadness.

^BExcluded from the factor analysis due to significant overlap with other language variables (14 variables excluded).

^CExcluded from factor analysis as the variable did not refer to meanings or features of specific words e.g. punctuation categories (16 variables excluded).

^DExcluded from the factor analysis due to inadequate variance (2 variables excluded).

^E Removed from factor analysis due to an individual Kaiser-Myer-Olkin value below 0.4.

Table 2. Characteristics of the analytic sample

Sample characteristics	Mean (SD)	
	Boys (N=882)	Girls (n=881)
Demographics		
Age (years)	11.5 (0.5)	11.5 (0.5)
Neighborhood disadvantage	1024 (62)	1022 (60)
Language measures		
Word count	61.8 (30.3)	78.5 (36.9)
Words per sentence	16.5 (7.9)	16.4 (8.5)
Wellbeing measures		
Psychosocial health	76.1 (13.4)	78.2 (14.1)
Health-related quality of life	0.90 (0.1)	0.90 (0.1)
General wellbeing	82.5 (13.1)	83.5 (13.3)
Life satisfaction	82.8 (13.2)	84.1 (13.8)

SD: Standard deviation.

All wellbeing measures, except health-related quality of life (scale from -0.11-1), are on a scale from 0-100, where higher scores represent better outcomes.

Table 3. Factor structure and rotated factor loading for exploratory factor analysis of LIWC2015 language variables^A

LIWC 2015 language variables	Identified language patterns					
	Factor 1 (Acting in the present and future)	Factor 2 (Positive emotion)	Factor 3 (Gender and relationships)	Factor 4 (Self-aware)	Factor 5 (Inquisitive and time focused)	Factor 6 (Confident)
Analytical thinking ^B	-0.72					
Auxiliary verb	0.85					
Verb	0.88					
Focus present	0.57					
Focus future	0.52					
1 st person singular pronoun	0.44			0.61		
Emotional tone ^B		0.82				
Positive emotion		0.86				
3 rd person singular pronoun			0.41			
Family			0.66			
Female			0.72			
Male			0.61			
Clout ^B				-0.41		0.70
Authentic ^B				0.80		
Common adverbs					0.54	

Interrogatives	0.59	
Time	0.48	
1 st person plural pronoun		0.52
Affiliation		0.75

^AOnly LIWC2015 language variables with factor loadings greater than 0.4 are shown. ^B Summary language variables created from prior linguistic analysis, where dictionaries are unavailable.

Table 4. Example stories with an extreme score on one identified language pattern and scores across other patterns (all patterns have a mean of 0 and SD of 1)

	Acting in the present & future (range -3.2, 3.1)	Positive emotion (range -2.5, 3.3)	Gender and relationships (range -1.6, 10.3)	Self-aware (range -3.9, 2.3)	Inquisitive & time focused (range -3.0, 4.2)	Confident (range -1.6, 6.0)
Story 1						
<i>"I would be live in <place>, and I'm be working at the train station, and I will be have two cats and one dog, and I'll have truck."</i>	3.1	-1.4	-0.2	-1.1	-0.8	-0.5
Story 2						
<i>"I would be doing a gaming career posting videos on YouTube because of entertainment and programing games to play for fun. I like listening to music so I would like to make music as well."</i>	-1.1	3.3	-1.3	-1.4	-0.1	0.8
Story 3						
<i>"Famous Singer. Live with Mum, Dad, <name>. Have 5 cats. Have a limo. Live in a mansion. Live in <place>."</i>	0.2	-1.9	4.5	-2.7	-0.3	0.8
Story 4						
<i>"I live in <place> and I play AFL for <team>. I live in my own house I have a wife and a kid. I am not very big and I have a lot of friends. I do a lot with my friends and I love my family."</i>	-0.9	0.5	1.0	2.3	-1.2	0.9
Story 5						
<i>"When I'm 25, I'm hopefully going to go to Indonesia to help the Orangutans, or say a speech to everyone who is listening and tell them what I think about what we are doing to nature and how we could help maybe?"</i>	1.7	-1.4	-0.7	0.3	3.5	-0.8
Story 6	1.0	-0.6	0.5	-0.5	0.1	6.0

“I would be traveling with my brother and cousin and we would go everywhere. We would be rich than anyone in the world. We are going to own 10 motorbikes, 15 dune buggies.”

Note: Stories are marked-up versions and the identifying details (e.g. names, places and sports teams) are removed from sample stories and replaced with <name>, <place> or <team>; The average score for each pattern is 0, positive values indicate adolescents who are high on this factors, whereas scores that are negative indicate adolescents that are lower on the pattern.

Table 5. Multivariable linear regression models examining associations between derived written language patterns and standardized wellbeing measures^A

Identified language patterns	Positively-framed wellbeing measures ^B				Negatively-framed wellbeing measures ^C			
	General wellbeing (n=1,718)		Life satisfaction (n=1,738)		Psychosocial health, PedsQL (n=1,740)		Health-related QoL, CHU9D (n=1,724)	
	SRC (95% CI)	p	SRC (95% CI)	P	SRC (95% CI)	p	SRC (95% CI)	p
Acting in the present & future	0.04 (-0.01, 0.09)	0.12	0.03 (-0.02, 0.08)	0.26	0.04 (-0.01, 0.08)	0.16	0.01 (-0.04, 0.06)	0.58
Positive emotion	0.05 (0.003, 0.11)	0.04	0.06 (0.01, 0.11)	0.03	0.07 (0.02, 0.12)	0.01	0.06 (0.01, 0.11)	0.02
Gender and relationships	-0.02 (-0.07, 0.04)	0.57	0.02 (-0.03, 0.08)	0.43	-0.003 (-0.06, 0.05)	0.93	-0.02 (-0.07, 0.04)	0.54
Self-aware	0.05 (0.001, 0.11)	0.05	0.09 (0.04, 0.14)	0.001*	0.08 (0.02, 0.13)	0.01	0.04 (-0.01, 0.09)	0.12
Inquisitive & time focused	-0.04 (-0.10, 0.02)	0.16	-0.005 (-0.06, 0.05)	0.87	-0.01 (-0.07, 0.04)	0.64	-0.01 (-0.07, 0.04)	0.69
Confident	-0.03 (-0.08, 0.02)	0.25	-0.06 (-0.11, -0.01)	0.03	-0.09 (-0.14, -0.04)	0.001*	-0.06 (-0.11, -0.003)	0.04

^A All models adjusted for sex, age and Disadvantage Index;
^B Questions are framed from a positive perspective (e.g. How happy are you today?). Higher scores indicate better wellbeing;
^C Questions are framed from a negative perspective (e.g. How upset are you today?). Higher scores indicate better wellbeing;
*p value remains significant after sequential Bonferroni adjustment;
n: number of participants; SRC: Standardized Regression Coefficient; CI: Confidence interval; p: p-values; QoL: Quality of life.

Author Manuscript

Author Manuscript