RESEARCH ARTICLE

Quantitative methods for group bibliotherapy research: a pilot study [version 1; peer review: 1 approved with reservations]

Emily T. Troscianko¹, Emily Holman², James Carney³

¹The Oxford Research Centre for the Humanities, University of Oxford, Oxford, UK

²Margaret Beaufort Institute of Theology, University of Cambridge, Cambridge, UK

³The London Interdisciplinary School, London, UK

v1 First published: 07 Mar 2022, 7:79
<https://doi.org/10.12688/wellcomeopenres.17469.1>
Latest published: 07 Mar 2022, 7:79
<https://doi.org/10.12688/wellcomeopenres.17469.1>

Abstract

Background: Bibliotherapy is under-theorized and under-tested: its purposes and implementations vary widely, and the idea that 'reading is good for you' is often more assumed than demonstrated. One obstacle to developing robust empirical and theoretical foundations for bibliotherapy is the continued absence of analytical methods capable of providing sensitive yet replicable insights into complex textual material. This pilot study offers a proof-of-concept for new quantitative methods including VAD (valence-arousal-dominance) modelling of emotional variance and doc2vec modelling of linguistic similarity.

Methods: VAD and doc2vec modelling were used to analyse transcripts of reading-group discussions plus the literary texts being discussed, from two reading groups each meeting weekly for six weeks (including 9 participants [5 researchers (3 authors, 2 collaborators), 4 others] in Group 1, and 8 participants [2 authors, 6 others] in Group 2).

Results: We found that text-discussion similarity was inversely correlated with emotional volatility in the group discussions (arousal: $r = -0.25$; $p = \text{ns}$; dominance: $r = 0.21$; $p = \text{ns}$; valence: $r = -0.28$; $p = \text{ns}$), and that enjoyment or otherwise of the texts and the discussion was less significant than other factors in shaping the perceived significance and potential benefits of participation. That is, texts with unpleasant or disturbing content that strongly shaped subsequent discussions of these texts were still able to sponsor 'healthy' discussions of this content, as evidenced by the combination of low arousal plus high dominance despite low valence in the emotional qualities of the discussion.

Conclusions: Our methods and findings offer for the field of bibliotherapy research both new possibilities for hypotheses to test, and viable ways of testing them. In particular, the use of natural

Open Peer Review

Approval Status ?

1

version 1

07 Mar 2022

1. **Thor Magnus Tangerås**, Kristiania University College, Oslo, Norway

Any reports and responses or comments on the article can be found at the end of the article.

language processing methods and word norm data offer valuable complements to intuitive human judgement and self-report when assessing the impact of literary materials.

Keywords

bibliotherapy, evaluation, group reading, narrative, literature, linguistic analysis

Corresponding author: Emily T. Troscianko (emily.troscianko@humanities.ox.ac.uk)

Author roles: **Troscianko ET:** Conceptualization, Data Curation, Investigation, Methodology, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Holman E:** Conceptualization, Funding Acquisition, Investigation, Methodology, Project Administration, Writing – Review & Editing; **Carney J:** Formal Analysis, Methodology, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome [205493; a fellowship awarded to James Carney]; and a grant from the Balliol Interdisciplinary Institute awarded to Emily Holman.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2022 Troscianko ET *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Troscianko ET, Holman E and Carney J. **Quantitative methods for group bibliotherapy research: a pilot study [version 1; peer review: 1 approved with reservations]** Wellcome Open Research 2022, 7:79 <https://doi.org/10.12688/wellcomeopenres.17469.1>

First published: 07 Mar 2022, 7:79 <https://doi.org/10.12688/wellcomeopenres.17469.1>

Introduction

It is intuitively plausible that reading ‘literature’ might have effects relevant to mental health and wellbeing, but is it also true? If so, what effects, and by what mechanisms do they arise? Is it possible to generalize, given the vast scope for variation in texts, readers, and contexts of reading?

Research on ‘creative bibliotherapy’ has begun to address these questions. Creative bibliotherapy, the reading of literary texts (which may include prose fiction, poetry, and/or drama) for health benefits, is distinct from ‘poetry therapy’, which tends to use poetry rather than narrative or dramatic forms of literature, and which often includes writing as well as reading poetry. When assessing bibliotherapy research conducted so far, an interesting divergence arises: The majority of existing bibliotherapy theory concerns individual reading, while most empirical work has involved group reading. The theory is based on minimal empirical evidence, and the empirical work has not yet been used to derive a more evidence-based theoretical account, although it has generated many hypotheses for efficacy and mechanisms of change.

Drawing on existing theoretical and empirical research on bibliotherapy, and using relevant tools from other areas (experimental psychology, natural language processing, cognitive literary studies), this study aimed to contribute to the project of mapping out testable hypotheses of bibliotherapeutic change. We adopted a group-reading methodology to connect the theoretically and empirically driven traditions via investigation of both text-centred and broader social aspects of how reading exerts change. In the rest of this introduction, we consider 1) the purported and observed therapeutically relevant effects of creative bibliotherapy in group settings and 2) the hypothesised mechanisms of therapeutically relevant change (in group and individual settings).

Therapeutically relevant effects

Many wide-ranging claims are made for the therapeutic value of reading, taking in purported benefits to self-understanding, self-expression, and self-esteem; interpersonal and communication skills; and creativity, change, and coping and adaptive functions, amongst others.

When it comes to putting such claims to the test, the most extensive empirical studies of group bibliotherapy have been carried out by Josie Billington and colleagues in the Centre for Research into Reading, Literature and Society at the University of Liverpool, in collaboration with The Reader. Their interventions typically involve reading a mixture of fiction and poetry, and have included:

- Reading groups at a GP practice, run by a trained facilitator for patients and local residents¹.
- Groups in healthcare settings led by a project worker, for adults with depression².
- Reading groups in a range of community and healthcare settings led by English undergraduates³.
- Groups in prisons led by a trained Reader in Residence^{4,5}.

- Groups led by a project member in healthcare environments, for people with dementia and the staff who care for them, using mostly poetry⁶.
- Groups for Mersey Care NHS Trust service users, led by a trained Reader in Residence, and also training NHS staff to grow a lasting reading-group culture⁷.

Billington and colleagues hypothesised that group reading should bring improvement in the areas of social, mental/educational, and emotional/psychological wellbeing, and found qualitative evidence of improvements across all areas, including: enhanced concentration, interest in learning, self-awareness, and capacity for self-expression; increased confidence; reduced sense of isolation². Similarly, the Mersey Care initiative, assessed in a Merseyside Service User Evaluation, documented ‘improvements in confidence, self-esteem, self-expression, memory, concentration, creativity, social engagement, listening skills and overall health and well-being’⁷. Robinson¹ reported positive effects on mood, loss of self (being ‘taken out of oneself’), concentration, confidence and self-esteem, pride and achievement, and communication skills, as well as appreciation of the opportunity to reflect on experiences in a supportive environment, and of a common purpose and shared ‘journey’.

Where quantitative measures have been used, reduction in dementia symptom severity⁶ and improvement on depression markers on the PHQ-9^{2,8} have been observed, though numbers were small and causality cannot be established because neither study included a control group. Other quantitative measures of change are rare, but a systematic review⁹ found a small to moderate effect on internalizing, externalizing, and prosocial behaviours amongst children from studies with a range of methods involving stories, poems, or films, plus various forms of interpretive support. A later review of creative bibliotherapy for post-traumatic stress disorder (PTSD)¹⁰ found no high-quality studies, but some suggestions that understanding and communication may be enhanced by group interventions involving reading. Meanwhile, a study using Persian poetry found evidence of quantitative improvement in mood, specifically a reduction in depression and an increase in hope amongst women with breast cancer receiving chemotherapy¹¹.

In other existing qualitative work, mood improvement is a common focus of inquiry. Mood was treated as the central dimension of change in a qualitative study using poetry about disability, which draws distinctions between *nervous* (i.e. emotional) *arousal*, *energetic arousal* (action readiness), and *hedonistic tone* (valence)¹². Pettersson’s user-focused study using poetry and short stories found six main categories of reading function reported by the four reading group participants who completed subsequent interviews: informational, escapist, social, perspective-creational, aesthetic, and therapeutic¹³. Pettersson points out that the first three of these align with Brewster’s outline of four user-centred models of bibliotherapeutic outcome for mental health problems: informational, escapist, social, and emotional (including empathic and cathartic)¹⁴. The absence of the emotional function raises the possibility that, for her participants, the emotional is subsumed in the therapeutic.

That said, however, the primary observed benefits were interpersonal and pragmatic, including improved self-confidence and ability to perform simple daily activities (including reading, willingness to engage in social activities, and capacity to complete daily chores).

In sum, then, changes on a wide range of social, cognitive-emotional, and behavioural dimensions are sought and observed in existing literature. Beyond the possibilities that in the absence of controls, randomization, and blinding, researchers are observing what they want to observe and participants are telling researchers what they know they want to hear or what they themselves want to believe, other questions arise. In particular, the breadth of documented effects raises the question of the extent to which they can be attributed to the group meetings or the reading of the text(s). Would a regular group meeting with no literary object of focus, or reading the same texts on one's own, have similar effects? Is engaging with both text and group contributing something specific? If so, what does each element offer; by what means? Here questions arise regarding mechanisms of change, and there is less evidence to draw on.

Elicitors and mechanisms of change

What 'active ingredients' are responsible for bibliotherapeutic change?

Some researchers draw on existing theoretical frameworks like Vygotsky's model of deep understanding^{3,15} or reader-response models of creative and participatory reading². Billington and colleagues² propose 'four significant components or "mechanisms of action": reading material, facilitator, group dynamics, and physical environment. They elaborate as follows:

1. 'A rich, varied, non-prescriptive diet of serious literature, including a mix of fiction and poetry (the former fostering "relaxation" and "calm", the latter encouraging focused concentration). Both literary forms allowed participants at once to discover new, and rediscover old and/or forgotten, modes of thought, feeling and experience.' (p. 6)
2. 'The role of the group facilitator in expert choice of literature, in making the literature "live" in the room and become accessible to participants through skilful reading aloud, and in sensitively eliciting and guiding discussion of the literature. The facilitator's social awareness and communicative skills were critical in creating individual confidence and group trust and in putting the group's needs above those of the individual where necessary. The facilitator's alert presence in relation to literature, the individual and the dynamics of the group is a complex and crucial element of the intervention.' (p. 6)
3. 'The role of the group in offering support and a sense of community.' (p. 6) (Evidenced by increased 'reflective mirroring' of others' 'thought and speech habits', and increased cooperation and personal confidence.)

4. 'The environment in contributing to atmosphere, group dynamic and expectation of the utility of the reading group.' (p. 7)

The researchers described the first three as 'essential in its success' and the last as 'influential'. At present, however, these are not falsifiable hypotheses. The observed effects may be due to all, none, or any subset of these factors. The relative importance of the four factors in different iterations may (or may not) also be highly variable between contexts.

In the dementia study cited earlier⁶, brevity and variety of texts, length of meetings (one hour), an informal setting (a lounge), and the presence of a staff member are highlighted as crucial. Robinson¹ stresses the importance of reading aloud (although she does not provide theoretical grounds for its importance), and of the expert facilitator's contributions, including in deciding how long to spend informally chatting before reading, judging when to stop reading for discussion, and helping people start with texts accessible and enjoyable enough to stay motivated for the longer term. Daboui and colleagues¹¹ suggest that the spiritual aspects of Persian poetry help in increasing hope, and that poetry as a form generally helps with communication about taboo subjects like death.

These observations of factors contributing to efficacy suggest ways of narrowing down the wide range of possible contributing factors, but they do not lead directly to accounts of the mechanisms by which efficacy is achieved. Several attempts have been made to set out a multistage cognitive process to account for observed changes. Gorelick, for example, sets out four phases of reading-stimulated therapeutic change: recognition, examination, juxtaposition, and application to self¹⁶. Billington, Longden, and Robinson⁵ single out 'memory and continuities' and 'mentalisation' (exercise of Theory of Mind) as mechanisms by which shared reading may have a protective or therapeutic function for problems such as depression, self-harm, and personality disorders. Montgomery and Maunders propose that the mechanisms of creative bibliotherapy might be roughly equivalent to those of cognitive behavioural therapy: They categorize these into 'cognitive reading processes' (recognition and reframing) and 'emotional reading processes' (empathy, emotional memories, identification), and suggest that parallel forms of identification, challenging, and replacing of negative thoughts occur in bibliotherapy, resulting in 'new attitudes and belief systems'⁹.

In their later paper, Glavin and Montgomery¹⁰ propose that the transporting effect of literary reading may permit a form of exposure therapy in which things that would be threatening in the real world can be safely engaged with in the fictional one. This application of the literary-theoretical concept of transportation brings their paper into contact with theories developed beyond the realm of bibliotherapy specifically, by cognitive literary scholars studying the psychological effects of reading in other contexts of psychological difficulty, including bereavement or post-traumatic distress. Kuiken and Sharma¹⁷, for instance, identify the complex phenomenon of 'sublime disquietude', resulting from a mixture of perceived emotional discord,

self-perceptual depth, and inexpressible realizations, all of which literary reading can induce. Sikora, Kuiken, and Miall¹⁸ explore the interplay of presence and absence that occurs during literary reading after bereavement in relation to a gradual acceptance of poignant memories of the dead person. Kuiken, Miall, and Sikora¹⁹ investigate how different forms of self-implication affect the emergence of this type of readerly response, especially via the blurring of boundaries between reader and narrator. This in turn relates to broader recent work on the many varieties of ‘personal relevance’ that a text may prompt, with a range of emotional and interpretive consequences²⁰.

Theoretical models from the individual reading paradigm tend to follow a common pattern: They emphasise similarity between the reader’s problematic experience and the arc of the protagonist’s story. This similarity is believed to prompt an identification-based connection between reader and protagonist which generates (possibly via a catharsis-like reaction) insight into the nature of a problem, in turn eliciting a problem-solving phase in which the reader learns from the protagonist and makes personal changes^{21–25}. This model faces both empirical and theoretical obstacles^{26–28}, not least a paucity of supporting evidence and a failure to pin down how and to what extent ‘similarity’ is therapeutically beneficial. The limit case here would be reading about one’s own experiences—for example, via diary-writing—and although therapeutic writing has a growing evidence base for many conditions and situations (on poetry therapy, see Ramsey-Wade and Devine’s²⁹ review), it seems unlikely that reading literature by others should exert its effects via the same mechanisms, merely diluted.

Building on survey data²⁸ suggesting that for at least one type of illness (eating disorders), heightened reader–protagonist similarity can generate reader perceptions of significantly harmful, rather than helpful, effects, in this study we were open to the possibility that reading might generate uncomfortable, distressing, and even apparently anti-therapeutic experiences for readers. We were also open to the possibility that short-term negative experiences, and individuals’ perceptions of them as unhelpful, might not be the whole story; such difficult experiences may contribute to positive longer-term effects. This hypothesis is compatible with both catharsis and exposure-therapy hypotheses of bibliotherapeutic efficacy, and with a more general view that the value of the experience of literary reading derives from complexity and multidimensionality rather than a simple feel-good effect.

Other perspectives are provided in research showing that readers who score high on a ‘search for meaning’ scale—a metric correlating with propensity for depression—are more receptive to literary (over non-literary) versions of a text³⁰. Similarly, other papers give theoretical and empirical grounds for thinking that fiction can reduce anxiety by offering predictive schemes for thinking about social interactions^{31–33}. Carney³⁴ also explores the role of predictability in culture more generally, suggesting how entropy (a measure of unpredictability) might differentially impact on anxiety and depression.

Objectives

The primary objectives of this study were as follows.

1. To record and transcribe a full set of reading-group discussions to generate detailed data on group–text interactions.
2. To trial new computational methods for sensitive analysis of the resulting complex textual material.
3. To generate new hypotheses on potential therapeutic efficacy and its mechanisms to guide future research in group bibliotherapy.

Methods

Ethical approval

Ethical approval for this study was provided by the University of Oxford Social Sciences and Humanities Interdivisional Research Ethics Committee, ref. MS-IDREC-C1-2015-155. All researchers involved in this project have completed the NIH online training course ‘Protecting Human Research Participants’ and have familiarized themselves with the BPS code of conduct and the University’s data protection and academic integrity guidelines.

Reading group procedures and participants

Our study took the form of two closed ‘Books, Minds, and Bodies’ reading groups in two consecutive university terms (October to December 2015 and January to March 2016), the first group meeting for seven weekly sessions, the second group for eight weekly sessions, as required to complete the selected texts. Two groups were run in order to generate a wider range of text–participant interactions than would have arisen in a single iteration. All 3 authors participated in the first group; EH and ET participated in the second group. EH and ET took responsibility for welcoming participants, covering house-keeping announcements, and establishing basic guidelines for conduct at the start of the first session, and thereafter participated in the reading and discussion in the same way as all other participants.

Other participants were recruited via an advert posted on noticeboards and local online events listings, offering a chance to “explore connections between reading and mental health and wellbeing” with a group of professional researchers interested in these topics. Group 1 included 9 participants (3 authors, 2 colleagues, 4 others); Group 2 included 8 participants (2 authors, 6 others, reducing to 5 others after session 3 after one participant dropped out due to other commitments). Group 1 included 6 females; Group 2 included 7 females, and the male participant was the one who dropped out after session 3, leaving an all-female group. Other demographic information (e.g. age, education, occupation, etc.) was not sought. We sought a minimum of 4 and a maximum of 6 recruited participants per group, to allow for a range of perspectives without compromising the intimacy and trust that can more easily be created in a smaller group. We required participants to be aged 18 years or over, to have an interest in narrative and/or mental health, and to be available for weekly two-hour sessions during the term. Other

than age, the only exclusion criterion was that participants be able to confirm that ‘I do not currently suffer from a mental health disorder’. This criterion was included to mitigate the risk of harmful effects being generated by participation. ET interviewed participants beforehand, providing the information sheet before the meeting, and asking the following questions in person:

1. What drew you to this project?
2. Are you able to commit to weekly sessions until [date]?
3. What are your expectations for the project?
4. Have you ever been part of a fiction reading group?
5. Are you comfortable with a rotating facilitator role?

After the roughly 20-minute meeting, potential participants were invited to take their time to decide whether they wished to take part, and after confirming their interest by email were sent further information on text selection (see below). Paper consent forms were provided for participants to check and sign at the first meeting of each group, with the opportunity to ask further questions. All but one of the interviewed candidates participated; the exception had to pull out due to scheduling difficulties.

Group 1 (henceforth *MT*, for ‘Michaelmas Term’, Oxford University’s autumn term) met at 6 pm on Mondays; Group 2 (henceforth *HT*, for ‘Hilary Term’, the spring term) met at 2:50 pm on Fridays. Meetings took place in one of two similar meeting rooms at Balliol College, Oxford. Texts were selected democratically: We circulated a list of five possible books suitable in length for a term’s meetings, providing either a short (one-page) excerpt or a link to an Amazon.co.uk preview. Each participant ranked the full selection in order of preference and vetoed any book they had read before. Fyodor Dostoevsky’s *Notes from the Underground* was selected for *MT* (in the Oxford World Classics translation by Jane Kentish), and Ted Chiang’s *Story of Your Life and Other Stories* for *HT*. We chose not to read the same text in both groups to ensure all participants shared in the experience of discovering a new text together, and also to capture as much variance as efficiently as possible (rather than attempting to control for variance as will be appropriate in a follow-up study).

Participants were presented with their own copy of the selected book, provided with a pencil, and encouraged to make markings in the book if they wished. Copies were collected between sessions to ensure participants did not read ahead on their own, and given to participants to keep at the end of the term. In the final *MT* meeting, having finished reading the main text, we asked each participant to bring a poem of their choice to read and introduce it to the group. In the final two *HT* meetings, having read four Chiang stories, we read two short stories by Franz Kafka: ‘A Hunger Artist’ and ‘Jackals and Arabs’. These additional sessions are excluded from the linguistic analysis of the literary texts and the transcripts (six for each term) covered in this article.

Books were read aloud by participants over the course of a session, switching with each paragraph (or, where these ran for longer than a page, with each new page). Sessions lasted 2 hours, the first hour spent reading, followed by an hour for discussion with refreshments (wine / soft drinks and nibbles in *MT*, tea and biscuits in *HT*). Audio recording began at the start of the second hour, and no notes were taken in addition to the recording. In order to lessen any sense of hierarchy between researcher–participants and others, each session was facilitated by a different participant, usually self-nominated. Facilitation was described to participants in the facilitator information sheet as follows:

The facilitator role, which will rotate at each session between participants, is broadly to keep the conversation going. This involves making sure that all participants feel able to contribute and perhaps occasionally asking someone who has been speaking for a long time to open the conversation to others. No more than one person should be speaking at once—and no private comments or conversations. This is a group discussion.

We have put together a list of possible questions that might be useful when you are in this role.

The questions included 5–7 questions in each of the following 4 categories:

- Emotional response (e.g. ‘Do you care what happens to the main character(s)?’)
- Interpretive response (e.g. ‘Did you ever feel that what was being described wasn’t what the passage was really “about”?’)
- Mental imagery (e.g. ‘Do you find you’re imagining more or less or differently than you do when you read alone?’)
- Drawing connections between real life and the narrative (e.g. ‘Do you think anything in this passage could help you think through or deal with difficulties in your life?’)

In practice, the questions were barely used, since conversational prompts from facilitators were rarely needed, and facilitators in general preferred to generate them ad lib when required.

After both groups had concluded, EH and ET transcribed the audio recordings of the discussions, EH taking *MT* and ET transcribing *HT*. One *HT* session was transcribed by a group participant, who was paid for her contribution; her transcript was checked and edited. Participants’ contributions were pseudonymized using colour codes to introduce their discussion interventions, and the codes were stored separately from the transcripts in a password-protected file.

While aware of the potential for researcher bias in a participatory framework of this kind, we considered that it could not be directly minimized at the group participation stage given the naturalistic setting, but that variations in individual perspectives would be manifest by all participants (some of whom were

also academics in other fields, i.e. there was no simple dissociation between ‘researchers’ and ‘participants’). In contrast to many prior studies, all forms of linguistically manifested bias were made explicit in the discussion recordings and transcripts, and were thus treated as an integral part of the complex dynamics under investigation. At the stage of analysis as opposed to participation, meanwhile, our methods were expressly designed to reduce the problematic forms of bias evidenced in qualitative content analysis, as we go on to document below.

Participant feedback. After the final reading-group session, participants completed an online survey designed to complement our analysis of their discussion contributions with direct self-report on the experience of taking part. The survey included questions about the enjoyment and significance of different elements of the reading-group experience, and forms of learning and change participants perceived to have resulted from taking part. To test for differences between the two groups, feedback from participants was analysed using an independent samples t-test that compared means between the two groups.

Comparison with other studies’ methods

The methods of our study both resemble and differ from other studies along several dimensions. [Table 1](#) summarises the comparisons.

Analytical methods

All analysis as outlined below was conducted on the full dataset resulting from 17 total individual participation instances (this included 2 repeat participations by researchers EH and ET) minus 1 participant who dropped out for non-study-related reasons after session 3 of HT.

Rationale. The guiding principle of our data analysis was that the discussion transcripts, including their linguistic relationship to the texts under discussion, stand at the centre. This analytical principle derives from the hypothesis that if shared literary reading is having wellbeing-relevant effects, they will be visible in the linguistic patterns of the text-prompted discussion. Although some previous studies have recorded and transcribed discussions¹, transcripts have been relatively under-used as sources of analytical insight. Our methods focused on two dimensions of participant response: emotional reaction and cognitive elaboration. Reading is a complex amalgam of emotional and cognitive responsiveness, and any thorough appraisal of its processes and impacts needs to attend to both³⁵. As our ambitions included quantifying the reactions of our reader groups, this meant finding ways to measure the cognitive and emotional variation implicit in our data. We resolved this problem by making use of word norm data and unsupervised machine learning in the following ways.

Table 1. Comparison of methods across cognate studies.

	Clinical participants	Poetry/mixed genres	Participants co-choose text	Topic-led reading	Trained facilitator	Refreshments	Reading aloud by participants
Our study (BMB)	N	N	Y	N	N	Y	Y
Robinson (2008)	N	Y	N	N	Y	N	Y
Billington <i>et al.</i> (2010)	Y	Y	Y	N	Y	N	Y
Robinson & Billington (2013) / Billington, Longden, & Robinson (2016)	Y	Y	N	N	Y	N	Y
Tukhareli ³⁶	Y	Y	N	Y	Y	N	N
Pettersson (2018)	N	Y	N	N	Y	N	N
Dubrasky <i>et al.</i> (2019) ³⁷	Y	Y	N	N	Y	N	Y
Daboui <i>et al.</i> (2018)	Y	Y	N	N	Y	N	N
COMMENTS		Last session of MT involved poetry				Wine & nibbles in MT; tea & biscuits in HT	

Emotional response. Traditionally, an impediment to supplementing qualitative assessments of the impact of reading has been the difficulty of measuring subjective responses. This is especially so with respect to emotional response, given that there is no universally attested taxonomy of emotion. When this emotional response is expressed verbally, the facility of language for expressing the same emotions in different ways compounds the problem. It follows that quantitative analysis of our transcript data has to try to solve the problem of extracting nuanced emotional information from text.

Our response to this challenge was to make use of word norm data. Essentially, word norms are corpora of language that have been rated by human participants along a specific dimension or set of dimensions^{38–40}. Warriner and colleagues³⁹ present 13,914 common English lemmas (words stripped of morphological variation) that have been rated by human participants for *valence*, *arousal*, and *dominance*. According to dimensional models of emotion, each discrete emotion can be represented in terms of an underlying set of finite components⁴¹. The VAD model of affect identifies valence, arousal, and dominance as the underlying factors responsible for emotional variation^{42,43}. That is, each discrete emotion can be thought of as a specific combination of valence (how pleasant or unpleasant it is), arousal (how stimulating or sedating it is), and dominance (how in-control or controlled it makes someone feel). Thus, anger is a low-valence, high-arousal, low-dominance emotion, while contentment is high-valence, low-arousal, and high-dominance. These extensive word-norm data provide an empirically validated way of assessing the overall emotional impact of a word, making mean VAD easy to calculate. The great value of the VAD norms is that they provide a low-dimensional proxy for emotional variation that is not restricted to words that are ostensibly emotional in their character (mood words like ‘happy’, ‘sad’, ‘angry’, etc.). They therefore provide a versatile means of establishing emotional variation on the basis of word use in linguistic documents. To this extent, they have an obvious value when it comes to capturing how our participants responded to texts and to each other on a per-session basis. Necessarily, they are also limited: emotion is conveyed not just by lexical choice, but also by phrasing, tone, and body language; equally, as averages of rater responses, VAD word norms are not sensitive to homonymy and other subtleties of usage. Moreover, competing dimensional models exist, with different dimensions as well as different numbers of dimensions⁴¹. Nevertheless, the VAD model has the best available data for measuring emotional variation in language, which makes it an appropriate choice in a study attentive to current possibilities for quantitative analysis.

The emotion of each text (both the discussion transcripts and the literary texts under discussion) was calculated by taking the mean of valence, arousal, and dominance across all the words of that text. Although this has the advantage of computational simplicity, within-text random variation means that the longer the text, the more they will tend towards the background mean for these values for English. As our texts were relatively short and of roughly equal length, we felt that any effects of

reversion to the mean would be small and spread equally across all texts. This analysis was performed by JC using bespoke scripts written in python 3; the script ascertained the value for each word on the transcript for each session for valence, arousal, and dominance using the Warriner³⁹ data and returned an average for the entire session.

Cognitive elaboration. Perhaps a greater challenge than measuring emotional response is measuring cognitive response. Allowing that the difference between ‘emotional’ and ‘cognitive’ is somewhat artificial, the fact remains that the space of possible concepts is wider than the set of possible emotions—it is, in fact, at least the size of a language’s vocabulary. Given that concepts can also be recursively combined to produce new concepts, this means that the set of possible concepts is combinatorially large. Until recently, this has meant that the task of extracting topics and themes from linguistic documents has been the preserve of the best pattern-matcher we know: the human brain. However, advances in unsupervised machine learning mean that it is now possible to identify automatically the specific ways in which words are combined with others to produce recurring items of content. In particular, word-embedding algorithms like word2vec, doc2vec, GloVe, and BERT provide empirically robust methods for capturing semantic variation at the level of word, document, and context that can be deployed at scale^{44–48}. Like all machine-learning methods, these algorithms are sensitive to initial parameter selection, so there is no sense in which they provide objective measures of semantic variation. Nevertheless, they inject a useful amount of statistical rigour into a practice that is often a hostage of ideological agendas.

For our purposes, the algorithm of most value was doc2vec, a document-level analogue to word2vec. Where word2vec represents the behaviour of a word across a corpus of documents by training a shallow neural network to predict its association patterns, doc2vec is trained by associating document tags with the word vectors comprising the document. (These word vectors are mathematical descriptions of how a word behaves in a corpus; the tags are the document name used to group the word vectors associated with a specific document.) Doc2vec thus represents high-level semantic variation across documents in a precise way, thereby making the discrete documents of a corpus comparable. In our case, the relevant corpora consisted of the transcripts of each session and the relevant portions of each text read in that session. The specific implementation of the doc2vec algorithm we used was that associated with the [Gensim](#) natural language processing library for python⁴⁹.

When using the doc2vec algorithm, the key parameters involve 1) choosing the size of the moving window of words amongst which semantic relations are assumed to obtain and 2) specifying the number of training epochs used by the algorithm. The first parameter is essentially a discourse specifier: In genres like poetry, this window should be large, given the dense interdependence of semantic elements; in technical writing it should be small, in view of the emphasis on strict denotation. In our case, we resolved on using the average

sentence length per group (Figure 1), given that sentence lengths in discursive conversation can often be quite long. The number of training epochs was 30k, which trial and error show is the point at which results stabilized.

Data cleaning. Both emotional and cognitive measures required texts to be cleaned in a way that made them amenable to automated analysis. In practice, this meant tokenizing the text into words and phrases and eliminating redundant variation across these words and phrases. Tokens were extracted by splitting character strings on whitespace; these were regularized using parsers from the spaCy natural language processing library for python⁵⁰. In practice, this involved lemmatizing each token, removing case variation, eliminating punctuation markers, and dropping stopwords ('it', 'a', 'an', 'on', 'the', etc.) that conveyed no semantic information. This reduced each text to a list of words that captured its semantic content in a consistent way.

Qualitative coding. EH and ET read the full set of transcripts and conducted a manual coding process for relevant features. Given our interest in the potential of group reading to have impacts on life beyond reading itself, and given existing theoretical work founded on notions of 'similarity', we adopted the categories of 'personal relevance' outlined by Kuzmičová and Balint²⁰ as our starting point. These included: *personal relevance*, *perceived similarity / simile-like identification*, *wishful identification*, and *metaphor-like identification*. Other categories identified in the first 2 transcripts from each term as necessary to the coding process included: *expression of emotional engagement*, *expression of no emotional engagement*, *engagement with character*, *liking* (of text), *dislike* (of text), *self-qualification*, and *human condition*. The resulting 11 categories were applied to coding the remaining transcripts.

Results

Emotion by session

VAD analysis of the transcripts showed clear differences in emotional profile both between individual sessions and between the two groups (Figure 2)⁵¹. On the whole, HT exhibited higher levels of valence and arousal and lower levels of dominance across sessions, though these differences were not statistically significant ($V: t = -1.52, p = .16$; $A: t = -1.3, p = .21$; $D: t = -1.38,$

$p = .19$). This may point to greater emotional febrility in HT, but it may also be an artefact of discussing very different texts. Although the lack of significance indicates that the differences may be the result of random chance, linguistic data exhibit high variation, meaning that significance is unlikely to be found in a sample size this small whether an effect is present or not. We therefore proceeded to analyse the emotional profile of the texts under discussion.

We found that text–discussion similarity was inversely correlated with emotional volatility in the group discussions (arousal: $r = -0.25$; $p = \text{ns}$; dominance: $r = 0.21$; $p = \text{ns}$; valence: $r = -0.28$; $p = \text{ns}$). (See additional data in the online repository: text_sim.)

It should be noted that MT-S6 had no text as such; instead, participants were invited to reflect on their experiences of participation and their relevance for further investigation of bibliotherapy. This meta-reflective structure presumably accounts for the outlier low arousal value for this session relative to the others. In Table 2 we offer a brief indication of some possible reasons for the outlier status of the HT-S6 VAD values.

Emotion by literary text

Since one possible driver of emotional responsiveness in participants is the emotional profile of the texts they are discussing, we took VAD measures of the discussed texts. Considered in aggregate, these measures positioned Chiang as on average lower in arousal and higher in valence and dominance than Dostoevsky (Figure 3). These differences were not statistically significant; nor did we expect them to be, given likely reversion to the mean for both authors over longer stretches of text. Within each author, there were relatively small differences between specific sections or stories (Figure 4).

Emotion by word norm data

An important element of establishing emotional variation involves identifying the actual words used. We did this by concatenating all transcripts for each group so as to create two corpora. After establishing VAD values for each word, we performed a decile split for each of valence, arousal, and dominance so as to split the data into ranked tiers. By comparing the top and the bottom deciles, this allowed us

Figure 1. Average sentence length per group for discussions in a) MT and b) HT. MT=Michaelmas Term; HT=Hilary Term.

Figure 2. Change in levels of arousal, dominance, and valence by group and session. MT=Michaelmas Term; HT=Hilary Term. The numbers on the y axis are scaled between 0 (min.) and 1 (max.).

to gain insight into the kinds of word driving the emotional profile of each group (see an example in [Figure 5](#)).

Cognitive elaboration

Our expectation was that measuring doc2vec document similarity between transcripts and texts for each group would show up a pair-wise relationship, such that the transcript of a session would be semantically closest to the text read in that session. Surprisingly, this was not the case: there was no obvious pattern linking a text to a session, and the method was, in any event, highly sensitive to parameter selection. What did emerge were striking between-group differences with respect to whether the group reproduced literary content in general, relative to non-literary content (i.e. the degree to which transcripts resembled the texts being read or resembled other transcripts) ([Figure 6](#)). As can be seen, the MT discussions reproduced far more of the semantic content of the Dostoevsky selections considered in aggregate than the HT sessions with the Chiang stories. In other words, the MT group was more ‘on topic’ than the HT group, if being ‘on topic’ is taken to mean discussing

the texts. An additional finding was that in sessions where the doc2vec similarity was low, VAD ratings varied more, suggesting that these sessions were more emotionally febrile.

Participant feedback

In both groups, most responses to the post-participation questionnaire followed a similar pattern ([Figure 7](#)). In assessing the significance of different elements of their reading and reflection during the sessions, the participants rated highest their emotional responses to the text and discussion, plus the perceived relevance of the texts to their lives. Engagement with the language of the text was rated lowest; engagement with textual meaning was rated of intermediate significance. Participants rated enjoyment of all elements of participation (the group, the additional short texts, reading aloud, listening to others read) significantly higher than enjoyment of the main text. Learning (about the text, about reading, and about oneself) was rated relatively low, and change (in reading habits, self-esteem, interactions with others, and in general) was rated as low (with general change higher than more specific types).

Table 2. HT, Session 6 versus other HT sessions.

HT, Session 6		
VAD	Value (values scaled from 0 to 1)	Commentary
Valence	Term average: $M = 0.616$; $SD = 0.153$ Session average: $M = 0.609$; $SD = 0.153$	<ul style="list-style-type: none"> • Discussion of skin problems, disfigurement, and anxiety about looks. • Topics related to negative aspects of religion, e.g. extremism, sociocultural tension, gender roles, shame, guilt, and hell/damnation. • Discussion of mental health issues, e.g. suicide and addiction in family contexts.
Arousal	Term average: $M = 0.381$; $SD = 0.101$ Session average: $M = 0.390$; $SD = 0.101$	<ul style="list-style-type: none"> • Inflammatory topics (religion and female attractiveness in contemporary culture). • Frequent lack of direct textual connection.
Dominance	Term average: $M = 0.594$; $SD = 0.109$ Session average: $M = 0.577$; $SD = 0.106$	<ul style="list-style-type: none"> • Frequent interruptions. • Discussion of female oppression by men and/or religion.

Figure 3. Average **a)** valence, **b)** arousal, and **c)** dominance values for Ted Chiang's *Story of Your Life and Other Stories* and Fyodor Dostoevsky's *Notes from the Underground*.

Figure 4. VAD (valence, arousal, dominance) values in **a)** Dostoevsky and **b)** Chiang by session selection.

Figure 5. Sample word clouds for valence in **a)** MT and **b)** HT. **a)** The words most responsible for driving positive valence in HT. These were collated by removing the words that the most positively valent session had in common with the least positively valent, leaving the words most responsible for the high positive valence. **b)** The words most responsible for driving positive dominance in HT. These were collated by removing the words that the most dominant session had in common with the least dominant session, leaving the words most responsible for the high positive dominance. MT=Michaelmas Term; HT=Hilary Term.

Intergroup analysis of participant feedback revealed three statistically significant differences amongst the 32 questions posed (Figure 8):

- How much did you enjoy *Notes from the Underground* / *Story of Your Life and Others*? ($p = .02$)
- How much did you enjoy the short texts in the final session? ($p = .03$)
- Did you enjoy listening to other people read sections of the text aloud to the group? ($p = .04$)

Figure 6. Semantic similarity between transcripts and texts in MT and HT. Note that a value of 0 means purely random association and a value of 1 means identity. Thus, lighter hue means more similar semantic content. Here, this is visible in the bottom half of the left-hand figure, which shows that transcripts were more similar to Dostoevsky selections than to each other. Note that, unlike in MT, HT transcripts in aggregate semantically resemble each other more than they resemble the Chiang texts. MT=Michaelmas Term; HT=Hilary Term.

Figure 7. Differences in post-participation feedback, by group and theme.

Qualitative coding

For the full coding results, please see the online repository (qualitative_coding_results). Five categories (expression of emotional engagement with the text; self-qualification [no. of instances]; self-qualification [no. of words]; and liking and dislike of the text) were considered relevant and coded by both EH and ET; 3 categories (engagement with a character, expression of complexity or difficulty, and personal revelation) were coded by EH only; and 6 (explicit absence of emotional engagement, personal relevance, perceived similarity, metaphor-like identification, wishful identification, and human condition) by ET only. Differences were manifest between our coding outputs for the 5 common categories; we felt these were in theory resolvable, but not relevant to the purposes of

the present analysis. ET's use of the categories derived from Kuzmičová and Balint resulted in near-zero outputs for 'metaphor-like identification' and 'wishful identification' (1 instance across the 2 categories in all 12 transcripts), low levels of perceived similarity (21 total instances), and medium levels of personal relevance (68 instances). ET's self-generated category 'human condition' (expressions of commonalities in human experience) yielded a far higher total (120 instances), with significantly more in MT than HT (82 v 38).

Liking and disliking the text

We found that liking or disliking the texts read did not impact on how valuable participants felt the sessions to be. Statistical analysis of post-participation feedback indicated that

Figure 8. Significant intergroup differences in post-participation feedback: liking of the principal texts, liking of the short texts, and enjoyment of hearing others read aloud. Scales range from 0 to 5, where a higher score indicates greater agreement with the statement. MT=Michaelmas Term; HT=Hilary Term.

reported enjoyment of the main texts being read was low, and significantly lower for HT than MT. Even when texts were explicitly disliked, participants enjoyed the discussions they prompted, and all other aspects of their participation. One HT participant remarked in response to the question ‘How much did you enjoy the stories by Ted Chiang?’, ‘*It did not matter. The stories, even when not enjoyable, still triggered discussions, interactions and people expressed their opinions.*’ Other aspects of enjoyment may play an important role: for instance, reported enjoyment levels for listening to the texts being

read aloud by others were higher overall, and significantly higher for MT than for HT, supporting the relative insignificance of enjoyment of the text itself in overall experience and effects of participation.

Liking and disliking the discussion

Many discussions of emotional register focus on whether the register is positive or negative, i.e. on valence. We found that valence alone was insufficient to capture the range of emotional variation in how participants reacted to the sessions.

Valence levels were higher in HT discussions. However, HT also manifested higher levels of arousal (associated with stress) and generated reports of finding the space unsafe. Conversely, MT participants evinced lower valence and arousal but higher dominance. In both terms, however, participants reported low levels of change in self-esteem or social interaction.

Discussion

Our analysis used new quantitative methods in the attempt to provide a combination of richness and replicability needed to answer questions about human responses to complex aesthetic phenomena. Using VAD (valence–arousal–dominance) modelling of emotional variance and doc2vec modelling of linguistic similarity to analyse the discussion transcripts and texts under discussion from two reading groups, we found an inverse correlation between text–discussion similarity and emotional volatility in the group discussions. Specifically, doc2vec analysis demonstrated that in verbal similarity taken as a whole, MT manifested significantly higher levels than HT, but that high arousal plus low dominance were also present. We also found no link between discussion valence and therapeutically relevant outcomes: The higher-valence group discussion (in HT) involved higher arousal and lower dominance, and neither group reported appreciable change in self-esteem or social interaction attitudes/habits. This suggests that higher valence does not necessarily translate into outcomes that reflect autonomy and self-directed action—traits that are often absent in mental health conditions. Post-participation feedback also suggested that enjoyment or otherwise of both the texts and the discussion was less significant than other factors in shaping how participants perceived the significance and potential benefits of their participation. This suggests that any therapeutic use of fiction need not be restricted to straightforwardly enjoyable or accessible texts.

Our findings are limited by the absence of a direct control group in which either the same participants read a different book or new participants read the same book. We decided against this method in order to allow everyone to experience a new text together and to maximize rather than control for variance. Any follow-up research should involve a control condition to establish causal links between textual features and discussion variables. Further limitations emerge with respect to our not having captured factors of interpersonal variation to do with reading history, education levels, or personality type. As these data may have mediated the impact of the texts read, including them in future research may provide more material for hypothesis generation and testing.

Analytically, our ability to draw robust conclusions from the linguistic data was limited by the absence of both a strong signal and a large sample size (of discussion material plus texts under discussion). This limitation could be addressed by rolling out such group meetings on a larger scale and pooling the textual materials, although automated transcription methods would in this case ideally be trialled to reduce the time investment of manual transcription. Other expansions such as testing out effects of texts in non-narrative genres, or involving participants with specific demographic characteristics or (with

appropriate safeguards) with current mental health conditions, could serve to evaluate the generalizability of the current findings.

Our results are based on a small sample of individuals in an idiosyncratic study setup. As such, caution should be used when generalizing to larger samples of readers. However, almost all real-world reading occurs in idiosyncratic ways, and by running the two groups we sought to control for some of the relevant variation between individuals. Moreover, it is hard to see how any experimental setup concerning group reading can capture the large number of factors that attend group reading. For these reasons, we suggest that some aspects of our results may generalize, but further research will be needed to establish which.

In the remainder of this section, we offer some starting points for broader interpretation of our findings with respect to the types of text–discussion interactions that may be therapeutically positive, as well as timescales of potential change. We begin by linking our work with other researchers’ observations on group facilitation methods.

Reading group facilitation

The role of the reading group facilitator has been the subject of frequent discussion in exiting group bibliotherapy research. The experiences of ET and EH in this study suggest there may be benefits to rotating facilitation responsibilities amongst participants; the key, however, is that facilitation occur, and that it be active. Active facilitation is important in any scenario involving structured discussion of literature, for several reasons. In the first instance, it keeps discussion focused on the text. This aligns with Billington’s reflections on the facilitator’s role, which emphasise the specifically literary guidance facilitators give: one ‘essential’ component for ‘success’ is ‘The role of the group facilitator in expert choice of literature, in making the literature “live” in the room and become accessible to participants through skilful reading aloud, and in sensitively eliciting and guiding discussion of the literature’². In the second, the facilitator role regulates conversational dynamics between participants. This supports Robinson’s observation that facilitator duties include ‘bring[ing] people in as much as possible into the general discussion’, being ‘willing to give everyone space to talk and read and reflect, and being ‘able to make them feel that their contributions to any discussions was [sic] valuable and interesting to others in the group’¹. Finally, there is a value in having an individual present who can respond to unanticipated or problematic issues emerging over the course of the discussions, and minimize any potential damage—always a possibility when reading complex texts that elicit a wide range of experiences. All of this becomes more important when the group contains researcher–participants, because they may be inclined to take a more passive role in their capacity as observers. ET’s personal reports after HT sessions, for instance, highlight the difficulties of participating, facilitating, and researching all at once: *‘I tried to “facilitate” and thereby prevented myself participating.’* The researcher perspective is encapsulated in ET’s habit of *‘reminding myself that this is all part of the experiment, and that it doesn’t matter how it goes; it matters that we learn something from how it goes.’* Sometimes more active intervention is needed.

Open and closed interpretation

What types of text encourage constructive interpretive and discursive patterns? Our experience was that texts that were interpretively ‘open’ did a better job of sustaining discussion and promoting a sense of shared purpose than texts that were interpretively ‘closed’. The former are texts that allow readers freedom to speculate as to their meaning without having an obviously ‘right’ answer; the latter are more like puzzles that can be solved in a singular, unequivocal way. The distinction may be aligned with Barthes⁵² distinction between *lisible* (readable, or literally readable) and *scriptible* (writerly, or literally writable) texts: the former directing interpretation down well-worn paths, the latter drawing attention to themselves and inviting interpretive elaboration. Our result converges with the less specific suggestion made by Billington and colleagues² that one of the four mechanisms of bibliotherapeutic action is that the literature being read (a mixture of fiction and poetry) be ‘serious’ and ‘rich’, although it may contradict the suggestion that the fiction ought to foster ‘relaxation’ and ‘calm’. Responses to open (or closed) complexity may be beneficial without being relaxing or calming.

Another discovery was that interpretive openness in a text facilitated deeper social interactions between participants than are typically had between strangers. Participants reported that hearing interpretive contributions from others generated curiosity and that later this was rewarded by discovering the personal origins of the contribution:

they'd come up and have an opinion on some part of the chapter that we read, and then I'd think ooh why did they think that, you know, that's really mysterious [laughter] And then over the next couple of sessions I'd learn more about them, and that was really interesting. (MT-S6)

However, this positive effect is lost when discussion strays too far from the text. We should also note that change in social interactions was not widely reported in the post-participation feedback, perhaps because of the nature or duration of this intervention relative to the patterns of participants' everyday interactions.

Text–discussion proximity

Direct correspondences between text and discussion seem not to be manifested in the emotional sphere in either group. However, our direct probing of text–discussion similarity via the doc2vec analysis demonstrated a clear difference between the two groups: In verbal similarity taken as a whole, MT manifested significantly higher levels than HT. For HT, the highest level of text–discussion similarity occurred when ‘Liking what you see’ was under discussion, a story about physical appearance, social appraisal, and other topics that have a bearing on day-to-day social life. This was the only story that participants said (in discussion and in the post-participation feedback) they had thought and talked about outside the session, and one participant described it (near the start of the subsequent session) as having ‘*probably resonated with me more than the others,*

just I guess cos it's kind of more immediately applicable to everyday experience’ (HT-S6). Correspondingly, participants in both groups considered the act of drawing connections between the text and their own lives to be of relatively high significance to their experience of reading and reflecting during the sessions. One mechanism of its significance may simply be that by definition it encourages text–discussion proximity, and therefore a greater experience of control during the discussion.

Our results for both groups (an association between low doc2vec similarity and high VAD variance) suggest that discussion needs to be grounded in the interpretive possibilities of the text for it to be therapeutically positive. There is, however, no direct connection on VAD grounds between the emotional profile of the texts and the discussions. Our results therefore challenge the similarity hypothesis concerning bibliotherapeutic mechanisms, in which benefits are derived via a close pedagogical relationship between the protagonist's psychological situation and progression and the reader's. We found that value for personal insight and wellbeing was sometimes derived despite the lack of obvious markers of similarity between reader and protagonist. For instance, the Underground Man's lack of growth was perceived by this MT participant as a spur to personal growth:

Particularly after the first few sessions in terms of how people occupy such different headspaces, and also discussions about him wanting to control social situations and set up a moment where he is in a certain position of power/standing shoulder to shoulder. It made me think more about not controlling relationships around me, which was helpful! (post-participation feedback)

When asking how text–discussion similarity is maintained or lost, the difference between discussions centred on author intention versus character motivation seems instructive. The former tended to deflate the world of the text and have the effect of closing down interpretive activity by assuming that ‘definitive’ answers exist. The latter kept the text interpretively open and thereby stimulated ongoing discussion, perhaps because there is clearly no singular ‘fact of the matter’ when it comes to a fictional character's motivations. One linguistic manifestation of this difference across the two groups was that the pronoun ‘he’ referred more often to the author in HT (along with ‘they’, also for the author) and more often to the protagonist in MT; engagement with characters was also correspondingly higher in MT (see online data file *qualitative_coding_results*). This challenges the suggestion made by Robinson that author and character are equivalent as objects of interpretive engagement: that ‘participants appeared to become enmeshed in the plot, as they developed their own theories around characters' actions and motivations, and the author's intent in using particular words, or including descriptions of particular contexts’¹. In our experience, speculation about the author may be more likely to lead to closed discursive forms: statements of liking/dislike, or statements aimed at establishing biographical facts.

Human condition, self-qualification, and VAD-guided word frequency as indicators or mediators of positive group dynamics

One important mediator of positive discussion dynamics may be a category that emerged in the qualitative coding of the discussion transcripts: reflections on the human condition. This emerged in a bottom-up fashion out of the attempt to employ the four major categories of ‘personal relevance’ set out by Kuzmičová and Balint, and the realization that none of those categories accommodated an important and frequently recurring phenomenon: the act of seeing something in the text or character(s) as resonating with a general or universal human tendency, or of seeing a character as an ‘everyman’ figure, etc. For instance:

GOLD: *I really loved the um, the bit about memory. [pause] Again because I think it's true. There are things in everybody's memory that he doesn't divulge to everyone but only to friends and so on. I thought that was... [pause] And that writing is often the way of processing those things in a preparatory way, to revealing something. [pause] There's no self apart from an autobiography in some sense. (MT-S2)*

GREEN: *He's also got that typical outsider syndrome, of feeling that you're superior to everybody else, and looking down on them, while also feeling extremely insecure when he's actually in their presence, so that you're living constantly in this world of your own making. (MT-S3)*

Frequency of *human condition* mentionings was significantly higher in MT than in HT (though low in the outlier S6 for both groups), and sessions that included most mentions also tended to include higher numbers of expressions of emotional engagement. It is possible that this type of personal relevance-drawing serves as a happy medium between making and expressing a direct personal connection and keeping discussion at an unthreatening but perhaps also ineffectual level of generality. The group reading context in particular may encourage this form of less individualized relevance attribution, as a way of speaking for oneself but also for a collective. Such comments may have a beneficially inclusive effect, and may also involve cause-and/or-effect links, or ambiguous overlap, with more directly personal connection-making. In this exchange, for instance, **ORANGE**'s human condition mention generates **BLUE**'s personal relevance mention:

ORANGE: *Because um—let me have a look. [pause] Sorry, I just have to just check back to the things I underlined. [pause] Because he's obsessed with his social footing and the status, but he tries to achieve it by bumping into people or—the way people look at him, and there's this constant sense of shame and—he talks about things like physicality and physical size, and I just found it really sort of—a bit like the primate rank, fighting. In a way that I don't feel—or I don't—I haven't experienced or thought about as much in the sense of like the female gender. I don't know. I think that's why.*

BLUE: *I think I was probably just instantly translating from kind of jostling with shoulders to I don't know, idiots not getting out of the way of me when I'm on my bike or something, and then size of body to like shape of body, or... I think I was probably doing the switch very automatically and that was why it didn't feel alien. (MT-S3)*

The danger also exists that expressing a personal view on a universal phenomenon may make others feel excluded or misunderstood by unjustified generalization. Specifically, the capacious pronoun *you* may sometimes serve as a linguistic veil for a contribution arising directly out of personal experience not shared more widely. As a form of distinctly discursive relevance-seeking, we suggest that ‘human condition’ statements are worth further investigation.

A second, related phenomenon observed in the discussions was the conspicuous number of self-qualifying statements. These included statements like ‘I don't know whether anyone else felt that’ or ‘maybe that's just me’, and other indications of not knowing, not remembering, or otherwise emphasising one's subjectivity or bias. These statements, occurring significantly more in MT, were a frequent feature in both groups' discussions. Although self-qualification could be seen as a trivial indicator of default false modesty, it may also serve a helpful function for social cohesion by moderating the strength of claims being made (including softening human-condition observations). Pragmatically, it often also seemed to provide an easy entry point for the next speaker:

YELLOW: *Um, yeah, but again it's not in the same way, he's not hating himself anymore there. He's just indifferent to himself. It's a different kind of stance, you know. I mean he doesn't seem petty, whereas I think up to this point, nearly everything he did is petty, you know like the little breakdown he has, and you just you know, you wanna give him a slap and say snap out of it. But here, I felt that was – again, does anybody else have any feelings on that, cos I'm not sure...*

GOLD: *I might agree, if it wasn't for the fact that he carried on writing.*

YELLOW: *Yeah, fair point. (MT-S5)*

Self-qualification may also, rather than being a causal driver in its own right, be an effect or a correlate of other ways in which discussion is made more inclusive, for example a general awareness of the importance of maintaining conversational flow amongst participants. In this sense it may relate to the various form of semantic and syntactic echoing identified as a significant contributor to positive group dynamics by other researchers^{2,8}.

Finally, VAD-structured word clouds provide a different type of clue to textual manifestations of constructive conversational dynamics. For example, the word *solution* was the most frequently spoken in the high-valence and high-dominance selections for HT (Figure 5), having arisen 13 times in S4 (and almost never in any other session), which was the

highest-dominance and highest-valence of all 6 sessions. Instances of its use make clear that participants were grappling with the difficulties of trying to make sense of the plot and authorial intention (in this instance, examples of closed complexity), as well as its wider relevance to problems humankind are currently facing (e.g. overpopulation), with more open-ended scope. Throughout the discussion, the word *solution* operates as a fulcrum between exploration of closed and more open forms of complexity. Its usage patterns add detail to the suggestion that such transitional dynamics may be associated with positive valence and feelings of control over the discussion. Thus VAD-structured word frequency mapping may, alongside human-condition observations and self-qualifying statements, be a useful indicator as to where to embark on close reading as a source of more sensitive insights regarding the direction taken by a specific text-prompted discussion.

Taking the longer view

A general question underlying the analyses in this study is a question about timescales. In particular, is it possible (or indeed likely) that a reading and discussion experience which has negative qualities (feels unpleasant, uncomfortable, or even unsafe) may elicit positive change (increased understanding, constructive action, etc.) at some point following the reading and discussion? Conversely, are the most positive experiences (on whichever metrics we select) more or less likely to generate positive change (of whichever type is valued)? This question of short- versus longer-term good versus ill is one we addressed through in-depth analysis of the verbal dynamics of the discussion sessions and consideration of the participant feedback at the end of each group's series of meetings. The free-response feedback from both groups included observations on a) learning about oneself and others, b) mood/relaxation benefits, and c) changed habits or attitudes around reading that cannot necessarily be gleaned from analysis of what was said during the discussions, and particularly not from the levels of reported enjoyment of the texts being read. Here is a selection of the concrete changes reported:

I learned that it is hard for me to clearly articulate certain types of emotional or experiential responses and maybe that is something I need to work to improve!

I'm actively looking for a book to read in its entirety. How long this motivation will linger... I'm not sure.

I emerged with a sense that people are nicer than I thought they were, and more inclined to be charitable in my interpretation of motives more generally.

I became more relaxed with respect to other problems after the reading group.

(MT)

it has thankfully increased willingness to see others points of view.

I came back in a much better mood. I would listen to my daughter when we were reading together and discuss what was happening.

I built a momentum as it were, continuing with the state of free flowing self expression even after the session.

You realise who you may look and sound alike, have similarities with and it provided a new talking point(s), idea, concept to recommend to others (without discussing content of discussions).

It's been a busy and sometimes difficult term for me since recovering from surgery- it's one of several activities that I'm pleased I took part in and proud I completed.

Yes wanted to be able to have a new routine and read around topics not necessarily thought of or enjoyed before. Also to further number of folk we know.

(HT)

It may or may not ever be possible to predict on the basis of textual and/or discussion content which of these benefits is mostly likely to accrue, let alone to tailor text, select participants, or guide discussion to maximize its likelihood. We suggest the next steps for group bibliotherapy research and practice should be open to several possibilities:

- 1) that negative effects are possible, in the short and/or longer term

In two personal reports which ET made after HT-S5 and HT-S6, she reports feeling 'excluded', feeling that 'there was no space and no time for me in the conversation', feeling resentment of other participants for never allowing silence between contributions, being keen to get away after the end of the session, finding it hard to re-engage with people after the end of the session, and subsequently feeling 'distanced from everything and everyone, and angry, and very fragile' for a number of hours, including responding to mental illness-related online material in a 'more viscerally defensive/aggressive' mode than usual. She concludes: 'I'm left feeling that in the wrong hands, or with the 'wrong' text, or just through bad luck, this thing really could be dangerous for people. I'm generally tired at the moment, but I'm healthy and generally strong and fairly well-balanced. If I weren't, I imagine that getting past this might have taken much longer. And who knows, perhaps with more time still, I'll feel that I've learned something about myself that needed to be learned, and that the negative short-term reaction was a fine price to pay for insight that took longer to come. But right now, I feel dislike and resentment and a lingering sense of unsettlement that I can't quite see turning into something good.'

- 2) that enjoyment of the discussion may be minimally related to liking of the text

I've learnt that I can still enjoy reading a text I don't like when it's in this kind of context. (HT post-participation feedback)

- 3) that positive and negative effects of participation may be minimally related to liking of the text

My relationship with reading has definitely changed. I see the value that it has to create a new and sometimes uncomfortable experience. Previously, I only considered reading to be strictly for the purpose of gaining knowledge. I also feel confident in sharing my opinions about a book. (MT post-participation feedback)

The last two quotations also indicate the value of experiencing a text for the first time in the bibliotherapeutic group, as well as the significant felt impact of discovering it together. Reading aloud, as a process that heightens the experience of a shared and 'live' sensory-cognitive journey, is likely to be relevant here. In other words, the way a text is encountered matters, possibly more than the nature of the text itself. As for the group discussion, having begun with a sceptical attitude as to the value of actually talking about a book as opposed to simply convening for discussion about something else, we find ourselves concluding that the book really does make a huge difference. In particular, our results suggest that talking about the book is importantly different from not talking about the book, and that emotional unpredictability is greater when the book is left behind. *Literary scholars persuaded that literature has effects* is not the world's most attention-grabbing headline, but we hope this pilot study has demonstrated how novel quantitative methods for sensitive textual analysis can flesh out the claim and generate useful hypotheses for future testing.

Data availability

Underlying data

Oxford University Research Archive: Books, Minds, and Bodies dataset. <https://doi.org/10.5287/bodleian:gJZz9KDE0⁵¹>.

This project contains the following underlying data:

- Qualitative_coding_results.xlsx
- raw_text_data.xlsx
- text_sim.xlsx
- qualitative_coding_results.xlsx
- participant_feedback_results.xlsx

Data are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

Acknowledgements

We are grateful to Nela Brockington for her involvement in designing and organizing the reading groups, for taking part in the MT group, and for stimulating conversations on methods, analysis, and the bigger picture. We would also like to thank everyone who took part in our reading groups for giving their time and their enthusiasm to the project.

References

1. Robinson J: **Reading and talking: Exploring the experience of taking part in reading groups at the Vauxhall Health Care Centre.** *Liverpool*. 2008; (115). Report No.: 8. [Reference Source](#)
2. Billington J, Dowrick C, Hamer A, *et al.*: **An investigation into the therapeutic benefits of reading in relation to depression and well-being.** *Liverpool*. 2010. [Reference Source](#)
3. Billington J, Sperlinger T: **Where does literary study happen? Two case studies.** *Teach High Educ.* 2011; **16**(5): 505–16. [Publisher Full Text](#)
4. Billington J: **"Reading for Life": Prison Reading Groups in Practice and Theory.** *Crit Surv.* 2011; **23**(3): 67–85. [Reference Source](#)
5. Billington J, Longden E, Robinson J: **A literature-based intervention for women prisoners: preliminary findings.** *Int J Prison Health.* 2016; **12**(4): 230–43. [PubMed Abstract](#) | [Publisher Full Text](#)
6. Billington J, Carroll J, Davis P, *et al.*: **A literature-based intervention for older people living with dementia.** *Perspect Public Health.* 2013; **133**(3): 165–73. [PubMed Abstract](#) | [Publisher Full Text](#)
7. Billington J, Davis P, Farrington G: **Reading as participatory art: An alternative mental health therapy.** *Journal of Arts & Communities.* 2013; **5**(1): 25–40. [Publisher Full Text](#)
8. Dowrick C, Billington J, Robinson J, *et al.*: **Get into Reading as an intervention for common mental health problems: exploring catalysts for change.** *Med Humanit.* 2012; **38**(1): 15–20. [PubMed Abstract](#) | [Publisher Full Text](#)
9. Montgomery P, Maunders K: **The effectiveness of creative bibliotherapy for internalizing, externalizing, and prosocial behaviors in children: A systematic review.** *Child Youth Serv Rev.* 2015; **55**: 37–47. [Publisher Full Text](#)
10. Glavin CEY, Montgomery P: **Creative bibliotherapy for post-traumatic stress disorder (PTSD): a systematic review.** *J Poet Ther.* 2017; **30**(2): 95–107. [Publisher Full Text](#)
11. Daboui P, Janbabai G, Moradi S: **Hope and mood improvement in women with breast cancer using group poetry therapy: a questionnaire-based before-after study.** *J Poet Ther.* 2018; **31**(3): 165–72. [Publisher Full Text](#)
12. Czernianin W: **Poetry as a therapeutic medium in shaping mood.** *J Poet Ther.* 2016; **29**(3): 135–45. [Publisher Full Text](#)
13. Pettersson C: **Psychological well-being, improved self-confidence, and social capacity: bibliotherapy from a user perspective.** *J Poet Ther.* 2018; **31**(2): 124–34. [Publisher Full Text](#)
14. Brewster E: **An investigation of experiences of reading for mental health and well-being and their relation to models of bibliotherapy.** University of Sheffield. 2011. [Reference Source](#)
15. Soter AO: **Reading and writing poetically for well-being: language as a field of energy in practice.** *J Poet Ther.* 2016; **29**(3): 161–74. [Publisher Full Text](#)
16. Gorelick K: **Poetry therapy.** In: Malchiodi C editor. *Expressive Therapies.* New York: The Guilford Press. 2005; 128–9.
17. Kuiken D, Sharma R: **Effects of loss and trauma on sublime disquietude during literary reading.** *Sci Study Lit.* 2013; **3**(2): 240–65. [Publisher Full Text](#)
18. Sikora S, Kuiken D, Miall DS: **An Uncommon Resonance: The Influence of Loss on Expressive Reading.** *Empir Stud Arts.* 2010; **28**(2): 135–53. [Publisher Full Text](#)
19. Kuiken D, Miall D, Sikora S: **Forms of Self-Implication in Literary Reading.** *Poet Today.* 2004; **25**(2): 171–203. [Publisher Full Text](#)
20. Kuzmiceva A, Balint K: **Personal relevance in story reading: A research**

- review. *Poet Today*. 2019; **40**: 429–451.
[Publisher Full Text](#)
21. Shrodes C: **Bibliotherapy: a theoretical and clinical-experimental study**. University of California, Berkeley. 1950.
[Reference Source](#)
 22. Russell DH, Shrodes C: **Contributions of Research in Bibliotherapy to the Language-Arts Program. I.** *Sch Rev*. 1950; **58**(6): 335–42.
[Reference Source](#)
 23. Pardeck JT: **Using literature to help adolescents cope with problems.** *Adolescence*. 1994; **29**(114): 421–7.
[PubMed Abstract](#)
 24. Pardeck JT, Pardeck JA: **Treating abused children through bibliotherapy.** *Early Child Dev Care*. 1984; **16**(3-4): 195–203.
[Publisher Full Text](#)
 25. Shechtman Z: **The contribution of bibliotherapy to the counseling of aggressive boys.** *Psychother Res*. 2006; **16**(5): 645–51.
[Publisher Full Text](#)
 26. Detrixhe JJ: **Souls in jeopardy: Questions and innovations for bibliotherapy with fiction.** *J Humanist Couns Educ Dev*. 2010; **49**(1): 58–72.
[Publisher Full Text](#)
 27. Troscianko ET: **Fiction-reading for good or ill: Eating disorders, interpretation and the case for creative bibliotherapy research.** *Med Humanit*. 2018; **44**(3): 201–11.
[PubMed Abstract](#) | [Publisher Full Text](#)
 28. Troscianko ET: **Literary reading and eating disorders: Survey evidence of therapeutic help and harm.** *J Eat Disord*. 2018; **6**(1): 8.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 29. Ramsey-Wade CE, Devine E: **Is poetry therapy an appropriate intervention for clients recovering from anorexia? A critical review of the literature and client report.** *Br J Guid Counc*. 2018; **46**(3): 282–92.
[Publisher Full Text](#)
 30. Carney J, Robertson C: **People searching for meaning in their lives find literature more engaging.** *Rev Gen Psychol*. 2018; **22**(2): 199–209.
[Publisher Full Text](#)
 31. Carney J, Wlodarski R, Dunbar R: **Inference or Enaction? The Impact of Genre on the Narrative Processing of Other Minds.** *PLoS One*. 2014; **9**(12): e114172.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 32. Carney J, Robertson C, Dávid-Barrett T: **Fictional narrative as a variational Bayesian method for estimating social dispositions in large groups.** *J Math Psychol*. 2019; **93**: 102279.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 33. Carney J, MacCarron P: **Comic-Book Superheroes and Prosocial Agency: A Large-Scale Quantitative Analysis of the Effects of Cognitive Factors on Popular Representations.** *J Cogn Cult*. 2017; **17**(3–4): 306–30.
[Publisher Full Text](#)
 34. Carney J: **Culture and mood disorders: the effect of abstraction in image, narrative and film on depression and anxiety.** *Med Humanit*. 2020; **46**(4): 430–443.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 35. Mar RA, Oatley K, Djikic M, et al.: **Emotion and narrative fiction: Interactive influences before, during, and after reading.** *Cogn Emot*. 2011; **25**(5): 818–33.
[PubMed Abstract](#) | [Publisher Full Text](#)
 36. Tukhareli N: **Bibliotherapy in a Library Setting: Reaching out to Vulnerable Youth.** *Can J Libr Inf Pract Res*. 2011; **6**: 1–18.
[Publisher Full Text](#)
 37. Dubrasky D, Sorensen S, Donovan A, et al.: **“Discovering inner strengths”: a co-facilitative poetry therapy curriculum for groups.** *J Poet Ther*. 2019; **32**(1): 1–10.
[Publisher Full Text](#)
 38. Brysbaert M, Warriner AB, Kuperman V: **Concreteness ratings for 40 thousand generally known English word lemmas.** *Behav Res Methods*. 2014; **46**(3): 904–11.
[PubMed Abstract](#) | [Publisher Full Text](#)
 39. Warriner AB, Kuperman V, Brysbaert M: **Norms of valence, arousal, and dominance for 13,915 English lemmas.** *Behav Res Methods*. 2013; **45**(4): 1191–207.
[PubMed Abstract](#) | [Publisher Full Text](#)
 40. Lynott D, Connell L, Brysbaert M, et al.: **The Lancaster Sensorimotor Norms: Multidimensional measures of Perceptual and Action Strength for 40,000 English words.** *Behav Res Methods*. 2020; **52**(3): 1271–1291.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 41. Rubin DC, Talarico JM: **A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words.** *Memory*. 2009; **17**(8): 802–8.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 42. Mehrabian A: **Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies.** Cambridge MA: Oelgeschlager, Gunn & Hain, 1980; 381.
[Reference Source](#)
 43. Bakker I, van der Voordt T, Vink P, et al.: **Pleasure, Arousal, Dominance: Mehrabian and Russell revisited.** *Curr Psychol*. 2014; **33**(3): 405–21.
[Publisher Full Text](#)
 44. Rong X: **word2vec Parameter Learning Explained.** 2014; 1–21.
[Reference Source](#)
 45. Mikolov T, Chen K, Corrado G, et al.: **Efficient Estimation of Word Representations in Vector Space.** 2013; 1–12.
[Reference Source](#)
 46. Pennington J, Socher R, Manning C: **Glove: Global Vectors for Word Representation.** *Proc 2014 Conf Empir Methods Nat Lang Process*. 2014; 1532–43.
[Publisher Full Text](#)
 47. Campr M, Ježek K: **Comparing semantic models for evaluating automatic document summarization.** *Lect Notes Comput Sci*. 2015; **9302**: 252–60.
[Publisher Full Text](#)
 48. Devlin J, Chang MW, Lee K, et al.: **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.** Palo Alto, 2018.
[Reference Source](#)
 49. Řehůřek R, Sojka P: **Software Framework for Topic Modelling with Large Corpora.** *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, 2010; 45–50.
[Publisher Full Text](#)
 50. Honnibal M, Johnson M: **An Improved Non-monotonic Transition System for Dependency Parsing.** *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics; 2015; 1373–8.
[Publisher Full Text](#)
 51. Troscianko E, Carney J, Holman E: **Books, Minds, and Bodies dataset.** University of Oxford. 2022.
[Reference Source](#)
 52. Barthes R: **The Pleasure of the Text.** New York: Hill and Wang, 1975.
[Reference Source](#)

Open Peer Review

Current Peer Review Status: ?

Version 1

Reviewer Report 04 April 2022

<https://doi.org/10.21956/wellcomeopenres.19317.r49064>

© 2022 Tangerås T. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Thor Magnus Tangerås

Westerdals School of Communication, Kristiania University College, Oslo, Norway

The article is interesting and provides a valuable contribution to research on literary reading in groups and the development of quantitative methods for analysing emotional and cognitive aspects of text-discussion relationships and complex reader responses.

The article is clearly structured, well written and engaging. The introduction covers most, but not all (see further down), relevant research on the effects and mechanisms of various bibliotherapies.

The rationale and objective of the study is clearly laid out:

There is an absence of analytical methods capable of providing sensitive yet replicable insights into complex textual material.

This pilot study offers a proof-of-concept for new quantitative methods including VAD (valence-arousal-dominance) modelling of emotional variance and doc2vec modelling of linguistic similarity.

And the sections on method, results and discussion are thorough and transparent. The analytic method is replicable.

In sum, I recommend that the article be indexed, but there are issues that should be addressed in each section of the paper:

Introduction:

The article would profit from a sharper discussion of what is implied by “bibliotherapy”.

At least two things should be highlighted here:

- That “bibliotherapy” is a contentious term, and covers a highly diverse set of practices, contexts and rationales. For instance, some forms are centered around an interview with a prospective reader with a problem, the problem is analysed and the “bibliotherapist” selects and recommends apt reading material. What constitutes “apt” literature is problematic. E.g. should choice of text be based on “similarity of problems”? And the current study provides empirical findings of great value here.

- Shared Reading, the group reading practice that is now most widely disseminated (not just in the UK but across Scandinavia and Europe) does not call itself a “bibliotherapy”. The purpose of Shared Reading was to allow everyone regardless of background the access to and enjoyment of “great reading” (cf Jane Davis, *The Lancet*). Therapeutic effects were seen as secondary gains, and not the effect aimed for. I think the current article would do well to recognize that a central premise of SR is that the chosen literature may not be “likeable” – challenging, threatening, etc. As such, this study lends empirical support to this central premise.

Also, a little bit of historical context would be useful. The “theorization” of bibliotherapy consists largely of constructs borrowed from psychoanalysis (identification, catharsis, etc). It is only in recent years that empirical approaches to bibliotherapy and literary reading as such have gained currency. And the most relevant of this research – that carried out by Kuiken *et al.* on the one hand, and Phil Davis *et al.* – is primarily bottom-up and not theory driven.

Given that the article briefly discusses Billington’s account of four “mechanisms”, and the problems associated with establishing the impact of each, it would be useful to discuss how One could isolate variables, given that this study in fact revolves around discussing the role of two of these: choice of reading material and group discussion, but has altered one premise of SR – that the facilitator be trained and highly skilled.

Method:

What is referred to as “method” here (and occasionally also “methodology”) is problematic. On the one hand, it comprises reading group procedures and participants, on the other the methods of analysing data.

What should be stated clearly is that the authors have in fact invented their own group reading method. Whether that should be called Bibliotherapy or something else is another matter. Table 1, designed to show the differences in variables among various approaches, does not cover all relevant aspects here.

First of all, why have the authors decided to design their own unique blend? Given that Shared Reading forms the central touch point in terms of empirically-based group methods, perhaps it would be better to gather data from SR sessions? If not, why not?

Very little is said about the participants. Are they all academics/students at Oxford? Are they all avid readers? They are interviewed beforehand and primed as to their role, so they are highly motivated to participate. This is relevant. They get to choose the reading material, but from which options? And based on what? Perhaps the issue of identification or likeability or plot interest is already integral to that choice.

Why not discuss the relevant differences to Shared Reading? I think it is significant that the session is divided in two here: first read, then discuss. But the most problematic aspect is that facilitation is rotated among participants. The list of possible questions they are given beforehand are mostly evaluative, asking for a “head” response rather than immediate emotional reactions as they happen. This is a very important objection given that so much of the discussion of the results is about emotional responses.

Method of analysis:

This part is the central part of the method. How can one find and develop quantifiable and replicable ways of determining emotional and cognitive responses and interactions?

I find the explication of rationale for selecting VAD and doc2vec, and the procedures and documentation of their implementation, solid and convincing. This part of the article needs no alterations. What is worth discussing, however, given that this is a pilot study, is: what would the difference be if a different emotion model was used?

Results and Discussion:

Results are clearly laid out and analysed. Emotions by session, by literary text, by word norm data – and cognitive elaboration.

The discussion is relevant and thorough.

My main objective here is:

There is a significant problem in that the article does not clearly distinguish between reader responses, interactions and reported worth of participation on the one hand, and therapeutic effects on the other. Cognitive and emotional engagement with text and with group can be established, and also how the participants self-report the value of participating. However, there is no bridge to documentation of therapeutic benefits/improved psychological well-being over time.

As stated in the article: Our analysis used new quantitative methods in the attempt to provide a combination of richness and replicability needed to answer questions about human responses to complex aesthetic phenomena.

There is a big difference between the question of emotional-cognitive responses to aesthetic phenomena and the question of the therapeutic benefits of these responses. As such, the findings of the study may be more relevant to the discussion of reader response theory than bibliotherapy theory.

In the literature review, some highly relevant studies are omitted:

Longden, E., Davis, P., Billington, J., Lampropoulou, S., Farrington, G., Magee, F., ... Rhiannon, C. (2015). Shared reading: Assessing the intrinsic value of a literature-based health intervention. *Med Humanities*, 41, 113–120. doi:[10.1136/medhum-2015-010704](https://doi.org/10.1136/medhum-2015-010704).¹

Longden, E., Davis, P., Carroll, J., & Billington, J. (2016). An evaluation of shared reading groups for adults living with dementia: Preliminary findings. *Journal of Public Health*, 15(2), 75–82. doi:[10.1108/JPMH-06-2015-0023](https://doi.org/10.1108/JPMH-06-2015-0023).²

One of these studies does in fact have a control group where a different group activity is used. And in one of these studies, pre-/post testing is used. If objective/quantifiable measures of improved mental health/wellbeing are to be found, there must be a concept of wellbeing and a way of measuring it systematically. Thus, if it were shown that participants did improve mental health, then hypothesizing that this improvement would be reflected in the transcripts would be of great importance.

(Another article not cited, (Shared reading as an affordance-nest for developing kinesic engagement with poetry: A case study, Kjell Ivar Skjerdingsstad and Thor Magnus Tangerås) Discusses how participation can be of benefit even when one does not like the text.)³

In sum, the article would profit from:

- Clearer context for study.
- Mention Longden *et al.*s articles.
- Discussion of rationale and choice of procedure of group method.
- In the discussion section, draw a distinction between therapeutic effects and reader responses. And elaborate on how the method can be developed further, and which hypotheses can be tested and which variable must be controlled for

References

1. Longden E, Davis P, Billington J, Lampropoulou S, et al.: Shared Reading: assessing the intrinsic value of a literature-based health intervention. *Med Humanit.* 2015; **41** (2): 113-20 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Longden E, Davis P, Carroll J, Billington J, et al.: An evaluation of shared reading groups for adults living with dementia: preliminary findings. *Journal of Public Mental Health.* 2016; **15** (2): 75-82 [Publisher Full Text](#)
3. Skjerdingsstad K, Tangerås T: Shared reading as an affordance-nest for developing kinesic engagement with poetry: A case study. *Cogent Arts & Humanities.* 2019; **6** (1). [Publisher Full Text](#)

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Transformative reading experiences, shared reading, bibliotherapy

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.