Validating judgments of perspective embedding

Further explorations of a new tool for literary analysis

D. H. Whalen,[12,3,] Lisa Zunshine,[4] Evelyne Ender,[5] Eugenia Kelbert,[6] Jason Tougaw,[7] Robert F. Barsky,[8] Peter Steiner[9] and Michael Holquist[310,]

[1]City University of New York, USA / [2]Haskins Laboratories, New Haven, CT USA / [3]Yale University, USA / [4]University of Kentucky, USA / [5]Johns Hopkins University, USA / [6]Universität Passau, Germany / [7]Queens College, USA / [8]Vanderbilt University, USA / [9]University of Pennsylvania, USA / [10]Deceased

Previous work (Whalen, Zunshine, & Holquist, 2012) has shown that perspective embedding ("she thought I left" embedding her perspective on "I left") affects reading times for short vignettes. With increasing levels of embedment 1–5, reading times rose almost linearly. Level 0 was as slow as 3–4. Embedment level was determined by the authors, but validation by others is desirable. In Experiment 1, we trained 12 literature students to make embedment judgments. Their judgments correlated highly with ours (.94 on average) and agreed exactly in the majority of cases (74.5%); almost all were within one (94.2%). In Experiment 2, judgments of the first three paragraphs of "To Kill a Mockingbird" (Lee, 1960) yielded a lower level of agreement; literature uses subtle means for introducing perspective embedment, and individuals differ about including them. Assessment of perspective embedding, and exploration of sources of disagreements, provide new tools for analyzing literature.

Keywords: reading, fiction, embedment, perspective, Theory of Mind, nested mental states

How literature engages its readers has long been a subject of avid discussion among scholars from a wide variety of theoretical perspectives (Brooks, 1992; Hakemulder, 2000; Miall, 2006; Scholes, Phelan, & Kellogg, 2006). One recent proposal, inspired by research in cognitive science, is that literature sustains readers' interest by activating the same cognitive mechanisms that make possible complex

Scientific Study of Literature 6:2 (2016), **278–297**. DOI 10.1075/ssol.6.2.05wha
ISSN 2210–4372 / E-ISSN 2210–4380

social interactions in real life (Kidd & Castano, 2013; Oatley, 2012; Palmer, 2010; Schneider, 2001, 2013; Zunshine, 2006). In real life, we attribute mental states to ourselves and others to explain behavior, a phenomenon known as Theory of Mind: "She didn't look up when I entered the room. Is she angry with me?" Some situations require a more complex embedment of perspectives – "Is she *angry* because she *thinks* that I didn't *want* her to *know* about the party?" Most of the time, we are barely aware of the process of attribution because it happens too fast. Moreover, we are often wrong about other people's mental states. (For instance, the "she" in question may have been preoccupied with her own feelings and thus had not noticed the speaker entering the room.) Still, fast and rough (Sperber & Wilson, 2002), ridden with stereotypes prompted by group-based reasoning (Hirschfeld, 2013), and just plain wrong as they may be, attributions of mental states charged with emotions underwrite every step of our social functioning (Dunbar, 1996).

Fiction, it has been argued, builds on our theory of mind by presenting us with complex social situations that require constant embedment of mental states (those of characters, narrators, readers, and authors). Some works of fiction conceal their "sociocognitive" underpinnings by not describing their characters' thoughts, feelings, and intentions, and making readers infer those mental states from actions and circumstances. Stylistic devices, ranging from metaphors (that ostensibly have nothing to do with thoughts and intentions) to the portrayal of characters who seem to lack any psychological interiority (Bowes & Katz, 2015; Zunshine, 2011) further contribute to creating complex mental states in readers. Generally, it appears that to follow a fictional plot, readers have to embed, or nest, mental states on at least the third level, as in, Tom Sawyer *doesn't want* his friends to *know* that he *hates* whitewashing the fence; Romeo *doesn't know* that Juliet merely *wanted* some people to *think* that she is dead (Zunshine, 2014, 2015a).

Fiction comes in many forms, of course, from novels to drama to memoirs that focus on imagination and consciousness. Although readers are (typically) aware that what they are reading is primarily, if not entirely, untrue, fictional content can elicit a spectrum of strong responses (Hogan, 2003; Miall & Kuiken, 2001; van Peer, Hakemulder, & Zyngier, 2007). If perspective embedment (also known as nesting of mental states) is as fundamental to making sense of such content as it has been argued to be (Zunshine, 2014), then it deserves a sustained look by scholars from a broad range of disciplines who want to understand fiction's hold on readers' minds.

The feature that we are focusing on is not strictly a linguistic one. For example, perspective embedment is different from syntactic embedding, even though it often makes use of syntactic embedding. A syntactically embedded sentence like "The rat that the cat chased is grey" does not contain any embedment of perspective and its embedment level, from the point of view of the present study, is zero. However,

a sentence like "The rat knew that the cat was after it" treats (anthropomorphically) the rat's knowledge as a level of embedment. There are many ways in which embedment can be created, with the "state of mind" verbs like "know" and "believe" being the most obvious. Understanding the means of conveying perspective embedment may ultimately lead to a discovery of more directly linguistic mechanisms, but the initially plausible candidate of syntactic embedding is not the one.

Our earlier work (Whalen et al., 2012) found evidence of an effect of embedment on reading time and, by extension, complexity. A recent follow-up study (Whalen, Zunshine, & Holquist, 2015) indicates that the effect is not due to the somewhat unnatural way that sentences were presented, namely, the moving window paradigm (e.g., Just, Carpenter, & Woolley, 1982). In our presentations of those results to various audiences, a common complaint has been that the judgments of embedment were purely those of the authors of that study and that they may not have wider validity. Indeed, there is, as yet, no study examining consistency of judgments of embedment level, nor what the differences in such judgments might tell us about the response of individual readers to literary creations. To remedy that situation, we trained a set of participants (literature students) on making such judgments according to the criteria we ourselves used. In Experiment 1, we used the vignettes we had prepared for the other work for the main conclusions. This way, they gain additional validity as a potential tool for other researchers. Further, as a first foray into applying the same method to literary text, Experiment 2 elicited judgments on the first 12 sentences (three paragraphs) of *To Kill a Mockingbird* (Lee, 1960). We expected both agreement and disagreement, and viewed disagreements as a rich source of insight into the complex grammatical and semantic aspects of literary language "shot through with intentions and accents" (Bakhtin, 1981: 293).

Experiment 1

In the earlier experiment, embedment levels of the sentences had been agreed upon by the previous authors (Whalen et al., 2012). However, it was clear that differences of opinion were still possible, and that the degree of embedment might not be completely consistent for other readers. To explore how consistent such judgments might be, we recruited 12 interested students to be trained in making these judgments. Then, they judged each individual sentence within the 84 vignettes. This experiment was immediately followed by the twelve sentences from the beginning of the novel that constituted Experiment 2 (Lee, 1960).

Method

Participants

The participants were 12 students of literature. 11 were graduate students at the City University of New York, in either the Comparative Literature (9), English (1) or Master of Liberal Studies (1) programs. The twelfth was an undergraduate English major from Columbia University. There were 5 females and 7 males. They were paid for 2 hours of training and 5 hours of judgments. All signed a consent form approved by the City University of New York Institutional Review Board.

Stimuli

The 84 vignettes of Whalen et al. (2012) were used. These were constructed to ensure consistency in overall length, lexical difficulty, number of orthographic characters, number of actors in the vignette, and the embedment of each sentence. They are available at https://yale.box.com/s/qvk12d3vwrppimedrrdkkj5hgrdq5p76 for those who would like see the full content and/or use them in their own studies. These were constructed to have embedment levels ranging from 0–5. Each had three actors or one (the narrator, "I"); in some cases, the narrator was an actor in the three actor vignettes. Each of these 12 combinations of actors and embedments was instantiated in 7 different vignettes, for a total of 84. Every sentence in every vignette was designed to be at the target embedment level. There were 387 sentences in total. The number of text characters (including the alphabetic ones and punctuation, but excluding spaces) in each vignette was manipulated so that it was exactly 350. If we include spaces in the count, they averaged 426.1 characters (ranging from 412 to 436). There were, on average, 6.2 punctuation marks per vignette (ranging from 5 to 9). The number of sentences averaged 4.6 (ranging from 2 to 10) while the number of words per sentence averaged 20.2 (ranging by vignette from 9.4 to 42.0). The number of words per vignette averaged 84.4 (ranging from 69 to 95). Each vignettes was presented as a single paragraph.

The vignettes were given to the participants in a randomized order that shifted systematically between participants. We chose to introduce this strategy because of our concerns that the limited time set aside for this test (5 hours) would leave the same questions unanswered at the end should all participants receive the same sequence. The randomization process ensured that all vignettes received multiple judgments, even though some participants might not make it through the entire protocol. As it happened, all participants chose to give responses to all vignettes, even if they used uncompensated time to do so. Nonetheless, our randomization procedure was as follows: Our first list was simply a random shuffle of the 84 vignettes. For the second randomization, this order was shifted by 7 items to create the next list: the first vignette was number 8 in the original list, and the final 7

items in the second randomization were the first 7 from the original one. The second list was then shifted by 7 for the next participant, and so on. As mentioned, all participants did make it through the entire set, taking a mean of 4.3 hours to complete the task (a range of 2–8.5).

Procedure

A two-hour training session was conducted by the second author, in which the concept of embedment was introduced and means of recognizing the levels were discussed. The participants were first presented with several individual sentences featuring different levels of embedment. For instance, the sentence "My last name begins with a Z, while Jim's last name begins with a P," was presented as containing no mental states, hence zero embedment. "I can't believe how warm LA is in January!" was presented as containing one mental state, that of being *surprised* by the LA weather, hence one embedment. "I used to think, when living in Santa Barbara back in the 90s, that I would enjoy some snow now and then" was presented as containing two embedded mental states: *thinking* about *enjoying* snow. Finally, "I can't believe that I used to think, when living in Santa Barbara back in the 90s, that I would enjoy some snow now and then," was presented as containing three embedded mental states: the speaker's incredulity about her former perception of her preferences.

The participants then discussed together several vignettes similar in structure to those featured in the upcoming experiment. None of these vignettes appeared in the experiment itself. They also considered examples of sentences featuring subtler ways of creating mental states, for instance, sentences containing no explicit descriptions of thoughts and feelings. As those may be more common in literary works taught in college and high school (as opposed to popular genre fiction, such as romance, in which mental states tend to be explicitly described), they were of direct interest to our participants. The participants were then asked to work individually, marking up levels of embedment in several vignettes, which was followed by general discussion. Two days later, they received an email, which contained a close analysis of two more vignettes, clarifying some issues that came up during the general discussion.

The randomized vignettes, followed by Experiment 2's 12 sentences from Lee (1960), were distributed as a computer text file, along with a file for the answers. There were ten columns for answers, matching the largest number of sentences in the vignettes. (There was a second line for the final two sentences from Lee (1960).) A final column was available for comments. Each sentence of each vignette received a rating.

Participants were then able to perform the experiment on their own computers at their own pace. They were informed that we could only pay for five hours

of time, and that they would be paid for five hours regardless of how long it took them. They reported how long it did take; two participants volunteered extra time to the task (an extra hour and an extra 3.5 hours).

Results

Numerical assessment of the embedment levels was performed across the 12 participants, resulting in 4,644 judgments.

There were four cases (two each for two participants) in which participants interpreted a colon (":") as indicating a sentence break, when we had treated the two clauses as one sentence. In these cases, the higher of the two judgments was taken as the overall embedment level. There were, in addition, 4 vignettes for which an extra judgment was entered, and 20 vignettes (1.9%) for which a judgment (or, in two cases, 2) was missing. The responses for all sentences in those vignettes were deleted, given that we could not be certain which sentence was skipped or added. This total was 116 judgments, or 2.5% of the total possible. Three participants had no such errors, while others had between 1 and 6.

Participant judgments agreed with our target levels most of the time (see Table 1). Very few ratings differed by more than 1 level. We assessed the degree of agreement using Cohen's kappa (Cohen, 1960). This measure takes account of the fact that raters will appear to agree, even if only chance is involved, if they select answers with the same, nonrandom proportions. Thus, for example, if most raters choose "average" in an experiment where most of the answers were indeed "average," they could seem to be accurate even though they were fairly random. Cohen's kappa takes values from 0 or even below 0 (no agreement) to 1 (perfect agreement). Our values ranged from .55 to .81, with an average of .71. Values from .61 to

Table 1. Percentage of instances in which the participants' judgments differed from the experimenters' by a particular amount (0 = agreement).

Difference from target	% responses
−4	0.04
−3	0.21
−2	0.78
−1	9.28
0	74.46
1	10.50
2	2.05
3	0.17

.80 are considered "Substantial" (Landis & Koch, 1977). All but two of the current values were in that range, with one being in the "Moderate" range and one in the "Almost Perfect" range. Another measure, Fliess' kappa, which assesses a group of raters at once, was .63, indicating "fair to good agreement beyond chance" (Fleiss & Chilton, 1983: 603). In sum, our raters were quite consistent with our judgments even with the limited training they could receive in a single an afternoon's session.

There were 107 of the 387 (27.6%) sentences that received complete agreement both among the participants and with the designated level. There were an additional 13 in which the mean was identical to the designated level, although some higher and lower judgments occurred (cancelling each other out). There were 343 sentences (88.6%) for which the mode (the most frequent response) was the same as the designated level. This left 44 sentences for which there was relatively wide disagreement.

As could be expected, there was a wider spread of values for the higher levels of embedment (see Figure 1). Only at levels 4 and 5 do the quartile ranges even

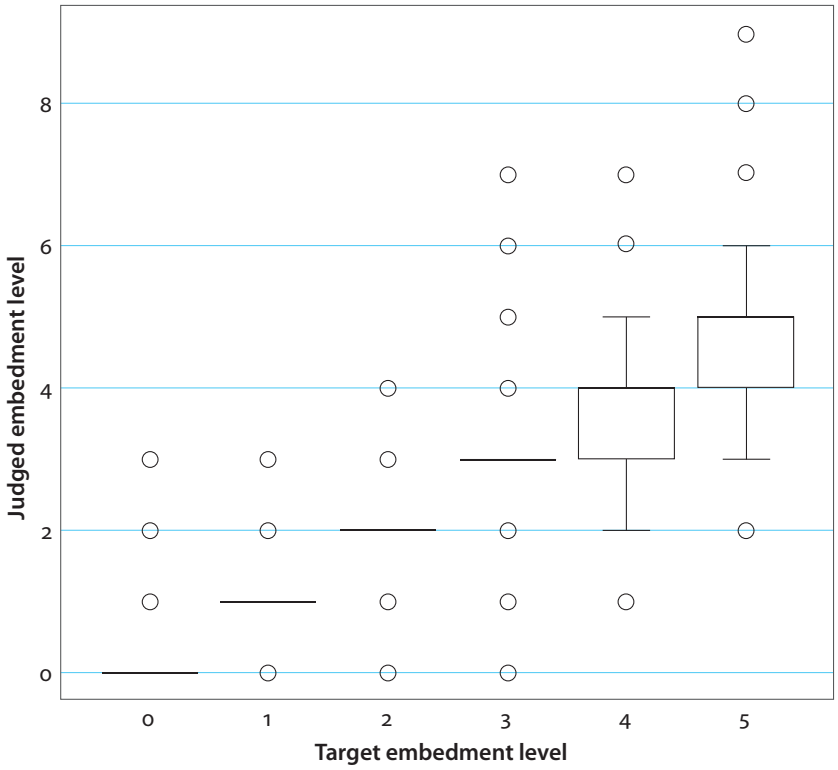

Figure 1. Box plot showing target embedment level vs. judged level. Means are shown with the solid horizontal bar. Quartiles (when visible) are shown as boxes. Outliers (individual judgments) are shown as circles.

appear graphically. The greater the number of cognitive states there are in a sentence, the greater the possibility for differences of interpretation. There are thus two obvious sources for the increased variability. The larger numbers allow for a greater range intrinsically (since embedments could not go below zero), and the more complex cognitive states will most likely be open to more interpretation (and disagreement) than the simpler ones. It remains to future research to examine the separate contributions of these factors and the possibility that the increase in complexity is nonlinear.

Comments were entered 44 times (out of the 1,008 responses (12 participants times 84 vignettes)), usually focusing on just one of the sentences of the vignette (there being only one comment field for each vignette). About half (45 of 84) vignettes received at least one comment; none received more than three. There was no tendency for the number of comments to increase with increasing embedment level ($r = .03$, n.s.). (This measure includes some comments that were not strictly about the sentences, such as "I'm starting to notice a pattern" or "took a break and espresso".) Some comments were made to sentences for which the participant's judgment agreed with ours. For example, one wrote for sentence 3 of vignette 74 (4 embedment level), "S3: "I guess I feel": uncertainty about one's feelings – she is embedding the possible feeling within the mental state of uncertainty." This is just what we had intended. Others commented on instances of disagreement. Particularly interesting are the cases in which a mental state was found in our level 0 (an example, perhaps, of what Miall and Kuiken call "stylistically induced feeling shifts" (2001)). For example, one participant found "or so" and "about" in vignette 7 ("Then I add bell pepper, coriander, turmeric, cinnamon, and cayenne pepper to the skillet and saute for another minute or so. I then stir in/tomatoes, chick peas, and raisins, and simmer, covered, for about 15 minutes. Finally, I add lemon juice and salt.") to be instances of mental states, giving those two sentences a 1 rather than our 0. Similarly, the "whizzed" of sentence 3 of vignette 2 ("My motorcycle whizzed through space at 80 miles an hour, taking me through different time zones, cuisines, and highway markings.") (level 0) was mentioned by one participant as implying an observer (i.e., level 1); two other participants also classified that sentence as level 1, though they did not comment on what formed the basis of their judgment. Indeed, the second author has since felt that this word can indicate a mental state (Zunshine, 2011). This may be the embedment version of the tree falling in the forest: Can a motorcycle whiz if no one hears it? The majority of the judges felt that it could, but a case could be (and has been) made that an observer is needed for something to whiz. More generally, participants occasionally wondered about the relationship between levels of embedment and the perceived "complexity" of a given sentence, speculating that sheer complexity may give rise to a sense (justified or not) of embedment. It's possible, in other words, that the

perceived complexity may be experienced as a "stylistic deviation" (Hakemulder & Van Peer, 2015) and as such imply an intention behind the sentence.

Sentences for which there was the widest disagreement between ourselves and our raters provided useful feedback for the potential ambiguities and even rethinking of the degrees of embedment. The most divergent sentence, "I suppose I used to wish for many things that I now don't think worthy of being aspired to." differed by 1.83 levels (Target: 5; Mean: 3.17). It now strikes us as being either level 3 or 4, depending on whether "aspired to" requires a perspective or not: it does indeed seem to be ambiguous. Similarly, the 1.42 difference in "About some puzzles I feel intuitively that it would be fun to think about different ways of solving them." (Target: 4; Mean: 2.58) makes us think that our attributing a mental state to "intuitively" separately from "feel" was incorrect, and that "would be fun" may not require a perspective either. The 1.17 difference in "I used to wish that I could summon in myself the same feeling of happy indifference to the outside world that, I think, is implied by that picture." (Target: 5; Mean: 3.83) seems to hinge on whether "implied" itself implies an embedment; we are now rather inclined to agree with our raters that it does not.

Even sentences with overall agreement but high variability are of interest. For example, "I was uncertain about my other tastes and preferences, like my interest in understanding how music worked, which could be simply a whim or a phase." received an average rating of 3.17, very close to our judgment of 3. However, the standard deviation was 1.40, indicating that different rater took quite different views of it. The "interest in understanding" is, perhaps, highly ambiguous and can be read as one or two mental states; "a whim or phase" can be read as one mental state or none, and it is likely that raters chose each option at different times. Similarly, "Still, I remember being shocked by how jealous I was about them as a child." had a moderate level of agreement (target, 2, mean, 2.90) but a relatively large SD (1.24). Does "jealous" require an embedment? Most of the raters apparently thought so, and we could be convinced of it ourselves.

The embedment judgments are not entirely straightforward, and they are probably influenced by the context of the vignette. In this experiment, the sentences of each vignette were presented together as a vignette, so this context was available. However, our instructions to give a response to each sentence may have weakened the link that we felt as we were constructing them. A future study could further these results in at least two ways: one would be to ask raters to highlight the elements of each sentence that they interpreted as introducing additional embedment. And another would be to examine these sentences in isolation to determine whether there are consistent differences that could be attributed to the context of the vignette.

Discussion

Overall, there was good agreement between our own judgments of embedment and that of the current participants. The great majority of responses were within 1 level (94.2%), but differences did account for 25.54% of the judgments. This discrepancy indicates that determining the level of embedment is not completely straightforward. As the comments showed, some of the disagreements were cases in which a mental state could be postulated even though we (and the majority of other participants) did not find it compelling enough. We may have encountered here what could be generously characterized as a proto-literary situation, in which extremely subtle textual cues are interpreted as implying an attitude or feeling. Because it's been proposed that literature in the canon often embeds mental states by implying rather than spelling them out (Zunshine, 2014, 2015a), such disagreements were of particular interest to the investigators. They demonstrated, among other things, that readers, particularly those whose exposure to literature has been quite extensive, might go the extra mile to intuit a mental state in the text. Call it "Theory of Mind working in overdrive," a professional hazard, perhaps, of those who read literature for a living. Indeed, "expert readers" (Schneider, 2001) have been shown to accommodate complexity more rapidly than nonexpert readers. Similar issues have been raised in relation to "authorial audiences" (Rabinowitz, 1987).

Experiment 2

The vignettes of Experiment 1 were designed to exemplify perspective embedments of various levels and deliberately did so primarily by explicit words denoting mental states such as "know" and "believe." Experiment 2 constitutes a preliminary exploration of the issue of implied (as opposed to explicitly spelled out) mental states, as is common in literary works. We chose sentences from a work commonly taught in American schools, *To Kill a Mockingbird,* to explore whether embedment judgments can be made as easily on a non-constructed text.

Method

Participants
The participants were those of Experiment 1.

Stimuli
After the 84 vignettes, the first 12 sentences of "To Kill a Mockingbird" (Lee, 1960) were presented, with the original paragraph breaks (see Appendix). This text was

chosen as an example of the literature commonly taught in schools in the United States, which can thus be considered as a crossover between literary fiction and popular fiction. (For a discussion of differences between the two, specifically in the context of Theory of Mind, see Kidd and Castano (2013) and Zunshine (2015a).) Hence there was a two-fold rationale for including these sentences in the experiment. First, it would give an indication of whether embedment judgments are more difficult in a work of literary fiction, where we can assume that a wider variety of embedment techniques, some of them very subtle, would be active. Second, it places the degree of agreement for the vignettes in a comparative light, that is, the degree to which it captures significant behavior across judges.

Procedure

Participants gave the same kind of embedment judgments they had made for the vignettes. They were instructed to complete these sentences even if they did not complete the vignettes. As mentioned above, all participants did complete the vignette task, even if they voluntarily used uncompensated hours to do so.

Results

As a first step in analyzing these paragraphs from an established work, we had to assign our own judgments of the embedment level for these sentences. There was rough agreement between the first two authors, with the exception that the second author felt that the mental states of the narrator had to be factored into the embedment, given the prominence of a very particular narratorial voice in the paragraphs under discussion. In the vignettes, we had decided that this level of embedment was constant and therefore would not contribute to our score. It may be that this exclusion was easier when we were the authors of the text in question and the embedments were signaled by mental state verbs. Excluding the narrator's perspective, however, was more consistent with the instructions and training of the participants than including it would have been, so the ratings by the second author were reduced by one to reflect this decision. This allowed the two to agree on the assigned level of embedment. This issue was discussed in the training session as well.

Numerical assessments were made as before by the participants.

Three participants treated the sequence with a semicolon as one sentence rather than two, as intended. Their judgment for that one sentence was used for both.

The target and participant results for the 12 sentences are shown in Table 2. In half the cases, the responses were essentially the same as the target, while 3 cases were above the target and 3, below.

Table 2. Judgments for the first 12 sentences of "To Kill a Mockingbird." Target is our judgment; mean and standard deviation (SD) are computed across the 12 participants. Mode is the most frequent response across the 12 participants. "Diff/mean" is the mean subtracted from the target level. "Diff/mode" is the mode subtracted from the target level.

Sentence	Target	Mean	SD	Mode	Diff/mean	Diff/mode
1	0	0.33	0.49	0	−0.33	0
2	2	2.67	0.65	3	−0.67	−1
3	0	0.17	0.39	0	−0.17	0
4	0	0.17	0.39	0	−0.17	0
5	1	1.08	0.29	1	−0.08	0
6	1	1.58	1.08	1	−0.58	0
7	1	2.00	0.85	2	−1.00	−1
8	3	2.08	0.90	2	0.92	1
9	2	1.92	0.67	2	0.08	0
10	2	1.00	1.04	0	1.00	2
11	1	0.92	0.90	0	0.08	1
12	0	1.25	0.97	2	−1.25	−2

As the difference from our judgment increased, the disagreement among the participants also increased. There is a larger standard deviation as the mean distance from the target value increased. As with the vignettes, this is to be expected as the ratings could not go below zero, so disagreements were restricted at the lower end of the scale. It may also be that the judgments are more variable because they are more difficult and thus open to increasing numbers of interpretations. If the latter is the case, we might expect that there would be vanishingly few sentences at the higher levels that would receive perfect agreement. This question is open to further exploration, but it's possible that the "rules" for an exact classification on the higher levels may elude us, and that the reader, as trained and sophisticated as she may be, may find that a synthetic grasp or "intuition" is what most people are working with when it comes to the more subtle examples.

The overall level of agreement between our judgments and the participants was lower for the literary extract than for the vignettes. An average of 49.3% of the judgments on the Mockingbird sentences agreed exactly, while 74.5% agreed for the constructed vignettes. Cohen's kappa ranged from −0.04 to 0.65, with a mean of 0.30. This is in the range of "fair" agreement (Landis & Koch, 1977). Fleiss' kappa for all raters (including the authors) was 0.24. This indicates poor agreement.

Discussion

The level of agreement between our judgments and those of our raters for the sentences from "To Kill a Mockingbird" was lower than that for our constructed vignettes. The primary reason for the decline is likely to be the use of less explicit means of indicating mental states in the novel than we allowed ourselves in the vignettes. Interpretations that included greater involvement of the narrator as well as an implied reader and/or author were also possible. A general comment came from one participant: "This was tricky. I tried to separate thoughts or feelings from speech or actions." Another felt that having two discussants in sentence 6 gave four levels of embedment, two for each of them. Similarly, for sentence 8, the "idea" that two of them shared implied to additional levels of embedment for a total of 3 (which was, indeed, the target level). In the future, to anticipate this particular concern – that is, whether a thought or a feeling shared by several characters should be counted as one mental state or as several mental states experienced by each character individually – researchers may want to introduce subjects to Alan Palmer's concept of "intermental unit" (Palmer, 2010). As long as several characters form such a unit (a state of affairs which typically does not last long), they should count as one entity. For example, in "the jury was shocked," the members of the jury constitute a single collective mind with a single perspective and should not be viewed as twelve separate mental states.

Another comment was that speaking seemed to count for thinking in literature, so that sentence 12 (see Appendix) should have 2 levels of embedment (opposed to our judgment of 0). Note that this was a sentence that the second author would also rate as having a higher embedment level, responding to the presence of the narratorial voice. Our guess is that while the verb "said" may contribute to the subtly elegiac mood of the sentence, it does not account for it. In other words, speaking may count toward embedment in literature, but this effect would always depend on the context; we would advise against a hard-and-fast rule such as "speaking equals thinking."

General discussion

The results from the present experiment indicate that embedment levels do elicit fairly consistent judgments even from judges who have received only minimal training in making those judgments. The disagreements presumably had multiple causes, but the present study did not allow for extensive debriefing in which the source of the disagreements might have been further explored. Nonetheless, the sizable agreement does indicate that embedment can be judged with very little

training. We would expect that further discussion about the source of disagreements would lead to better training and greater initial agreement about the embedment for different sentences. We can also expect that genuine disagreements will exist even after discussion, reflecting the role of differences of perception in what constitutes or can be recognized as a mental state when the reader faces complex situation of nesting. It is important to keep in mind that complex embedments emerge out of the *experience* of reading – as opposed to just being there in the text, unaffected by who opens the book and when. Our experiments have demanded an abstract, analytical, and sophisticated articulation of what, in the process of reading, may exist as a quick first apperception or "intuition" of what a character's mind may be like. Our protocols, meanwhile, by demanding increasingly fine-tuned determinations would be facing here what could be labeled as the challenges and exigencies of literary readings.

Another type of factor that might account for differences, one which was brought up by some of the participants, is that of linguistic and cultural differences. For instance, the same sentence from an eighteenth-century Chinese novel, featuring one character being aware of another character's intention of placating the third character's feelings would be experienced somewhat differently by an English speaker, who can only read it in translation, than by a native speaker of Chinese. In addition to processing the third-level embedment, the Chinese reader may also notice a classical allusion contained in the phrase describing the third character's feelings and thus factor in the intentions of the author, increasing the levels of embedment. Moreover, a scholar of Chinese literature, who happens to be an expert in the field, may construct the intentions of the author yet differently, while still operating on the third-to-fourth level of embedment (Zunshine 2015a).

The present experiments measured levels of embedment in sentences and short vignettes, and so the question arises whether such stimuli adequately model what is going on in works of fiction. There, some individual sentences certainly feature high levels of embedment, generated by either explicitly spelled out mental states or tacitly implied mental states. Taking as examples sentences culled from E.M. Forster's *Howards End*, one can see that in, "Ought Margaret to know what Helen knew the Basts to know?," third-level embedment is created by explicitly spelled out nested mental states, while, in contrast, in, "One may as well begin with Helen's letters to her sister," it is created by tacitly implied mental states.

Still, even on the sentence level, there are a number of important differences. First, in actual works of fiction, it's not to be expected that every sentence would contain third-level embedment. (For a notable exception, consider Henry James, who embeds three or four mental states in nearly every sentence.) Second, the context beyond the sentence may influence the experience of embedment. For instance, a perceptive reader may question the likelihood of Forster's narrator

proffering a crude nesting along the lines of "Ought Margaret to know what Helen knew the Basts to know?" and double-check its source in the text. That source turns out to be Tibby, a young man whose "attention wandered when 'personal relations' came under discussion." This means that we now may want to factor in the narrator's ambivalent attitude toward Tibby's lack of interest in "personal relations," an attitude that we have to infer, and may disagree about (for instance, some readers may decide that the narrator views Tibby's way of thinking sympathetically, while others, that he finds it off-putting). But even while disagreeing, we would still be nesting mental states, as in: "the narrator is *amused* that Tibby *doesn't really care* whether Margaret *knows* what Helen *knew* the Basts to *know*." That is, to make sense of a complex social situation implied by the sentence, "Ought Margaret to know what Helen knew the Basts to know?", we have to process high-level embedments: quickly, intuitively, without being aware of it, and often factoring in contextual cues from elsewhere in the text.

Context remains crucial when readers deal with embedments on the level of paragraphs and chapters. For instance, the literary plot device known as dramatic irony, which can work across chapters, is often instantiated as a third-level embedment. Thus in chapter seventeen of *Howards End*, the Schlegels' lease expires and Margaret starts searching for a new home for her family. From that time on, throughout Margaret's various conversation with the Wilcoxes, her marriage to Henry Wilcox, and her sister's involvement with the Basts, one important aspect of the readers' engagement with the novel is their *awareness* that Margaret *doesn't know* (while the Wilcoxes do know) that the late Mrs. Wilcox had *wanted* her to inherit Howards End.

This third-level embedment is essential to our understanding of Forster's novel. Yet studying it in the laboratory would effectively require cognitive scientists to practice the kind of close textual reading expected in literary studies. Hence the particular challenge of the interdisciplinary inquiry, as is exemplified by our approach. On the one hand, being realistic about what could or could not be undertaken in a lab, we are particularly interested in assessing the efficacy of the sentence-level analysis, even if it means foregoing a potentially more sophisticated investigation of paragraph- and chapter-long excerpts.

On the other hand, we believe that literary critics already study a broad variety of perspective embedments, as they uncover, construct, and debate nested motivations and attitudes of characters, implied authors, narrators, readers, and other critics. They do it, however, without being aware of the deep cognitive underpinnings of the process, that is, without realizing that they themselves are nesting their and their interlocutors' mental states while discussing nested mental states that they encounter in fiction (compare to Rabinowitz & Bancroft, 2014). (This is not to diminish or ignore the self-awareness that constitutes an integral part of

critical endeavor; it is merely to point out that this particular conceptual framework is new to literary analysis and has not yet become a familiar aspect of critical self-reflection.) What we hope to achieve, therefore, is to bring into the open some of the unspoken cognitive principles that underlie existing practices of literary analysis, while also suggesting that the sentence-level study of embedment may be most accessible for experiments conducted by cognitive scientists.

Important as the issue of self-reflection is, key payoffs accruing to the field of literary criticism from becoming aware of the omnipresence of multiply-embedded mental states in fiction lay elsewhere. Some of those payoffs are pedagogical; that is, they furnish students with new effective ways of engaging with the style of individual writers (Zunshine 2014). Others are interpretive; that is, they offer scholars new insights into the ideological work done by novels, for writers can portray some characters as capable of nesting more mental states than others – a difference often underwritten by considerations of class, race, and gender (Zunshine 2015a). Yet other payoffs pertain to educational policy; that is, attention to levels of embedment in fiction can be used to rebuff some of the unfounded claims made by proponents of the "Common Core Standards Initiative" about the place of literature in the k–12 curriculum (Zunshine, 2015b).

The broad agreement about the levels of embedment in individual sentences demonstrated by our experiment shows that sentence-level embedment of mental states is a real phenomenon that can be reliably assessed in a laboratory setting. Were we to speculate what might account for this agreement if embedment were not at play, we could say that by merely counting the words indicative of mental states in each sentence (such as feel, want, hope, fear, think, suspect, etc.), participants might have come up with similar results. Although our participants' comments strongly suggest that this was not likely to have been the case, to anticipate this objection in future experiments, we suggest adding several control sentences to the set, which would embed mental states in parallel rather than recursive fashion, and discussing the difference in the preceding training sessions. For instance, "I wanted her to know that I didn't like her approach and that I knew what she was hoping for" contains two parallel embedments, a third-level one (I *wanted* her to *know* that I *didn't like* her approach) and fourth-level one (I *wanted* her to *know* that I *knew* what she was *hoping for*). If participants merely count the "mental state" verbs, they will say that this sentence embeds mental states on the fifth level, which would be incorrect. Knowing that such control sentences are included in the set would presumably encourage them to be on the lookout for parallel embedments and thus go significantly beyond merely counting the mental state words.

Note, however, that this precaution applies only to the synthetic examples, such as the ones used in Experiment 1, because in actual works of fiction, mental states are often implied rather than stated explicitly. Along the same lines, we

believe that the method outlined in this study is valid as long as researchers are aware of its limitations when applied to works of fiction. That is, one must be aware that in fiction, sentence-level embedment constitutes only one subset of embedments and that (for instance) plot-level embedment of mental states may influence the perception of what is going on in a particular sentence. (For a related discussion see Van Duijn et al. (2015).)

It follows, then, that in a hypothetical interdisciplinary study that would one day, we hope, undertake the analysis of longer fictional selections, high agreement rates on individual sentences are somewhat unlikely, because of the necessary complicating role of large-scale (i.e., paragraph, chapter, and cross-chapter) embedments of mental states in the perception of the sentences. This word of caution fully applies to sentences containing third-level embedments. In works of fiction – as opposed to the synthetic examples used in Experiment 1 – one doesn't have to go to the fourth and fifth level to encounter ambiguity and disagreement. Sentences featuring mental states nested on the third level have plenty of potential for perennially conflicting interpretations. We saw it with the deceptively straightforward "Ought Margaret to know what Helen knew the Basts to know?", and the effect can be even stronger when nested mental states are implied rather than spelled out explicitly. The measurement of embedment thus provides new insight into literature, while the areas of disagreement open further lines of discussion for literary criticism.

Acknowledgements

This research was supported by a grant from the Teagle Foundation to Yale University and Haskins Laboratories, and by NIH grant HD-01994 to Haskins Laboratories. We are especially grateful to three co-authors who added to the second authors' stock of vignettes: Robert F. Barsky, Peter Steiner, and Eugenia Kelbert. We thank Vilena Livinsky and Argyro Katsika for running the studies, Katherine M. Dawson and Kevin Roon for technical assistance, and Xhuliana Agolli, Yulia Greyman, Stephen J. Frost, Hannah M. King and two anonymous reviewers for helpful comments. This article is dedicated to the memory of Michael Holquist.

References

Bakhtin, M. M. (1981). *The dialogic imagination* (C. Emerson & M. Holquist, Trans. M. Holquist Ed.). Austin: University of Texas Press.

Bowes, A., & Katz, A. (2015). Metaphor creates intimacy and temporarily enhances theory of mind. *Memory and Cognition*, 43, 953–963. doi: 10.3758/s13421-015-0508-4

Brooks, P. (1992). *Reading for the plot: Design and intention in narrative*. Cambridge, MA: Harvard University Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychosocial Measurement*, 20, 37–46. doi: 10.1177/001316446002000104

Dunbar, R. I. (1996). *Grooming, gossip, and the evolution of language*. Cambridge, MA: Harvard University Press.

Fleiss, J. L., & Chilton, N. W. (1983). The measurement of interexaminer agreement on periodontal disease. *Journal of Periodontal Research*, 18, 601–606. doi: 10.1111/j.1600-0765.1983.tb00397.x

Hakemulder, F. (2000). *The moral laboratory: Experiments examining the effects of reading literature on social perception and moral self-concept*. Amsterdam: John Benjamins Publishing. doi: 10.1075/upal.34

Hakemulder, F., & Van Peer, W. (2015). Empirical stylistics *The Bloomsbury Companion to Stylistics*. London: Bloomsbury Academic.

Hirschfeld, L. A. (2013). The myth of mentalizing and the primacy of folk sociology. In M. Banaji & S. Gelman (Eds.), *Navigating the social world: What infants, children, and other species can teach us* (pp. 101–106). New York: Oxford University Press. doi: 10.1093/acprof:oso/9780199890712.003.0019

Hogan, P. C. (2003). *The mind and its stories: Narrative universals and human emotion*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511499951

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111, 228–238. doi: 10.1037/0096-3445.111.2.228

Kidd, D. C., & Castano, E. (2013). Reading literary fiction improves Theory of Mind. *Science*, 342, 377–380. doi: 10.1126/science.1239918

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. doi: 10.2307/2529310

Lee, H. (1960). *To kill a mockingbird*. Philadelphia: J. B. Lippincott and Co.

Miall, D. S. (2006). *Literary reading: Empirical and theoretical studies*. New York: Peter Lang Publishing.

Miall, D. S., & Kuiken, D. (2001). Shifting perspectives: Readers' feelings and literary response. In W. van Peer & S. Chatman (Eds.), *New perspectives on narrative perspective* (pp. 289–301). Albany, NY: State University of New York Press.

Oatley, K. (2012). *The passionate muse: Exploring emotion in stories*. New York: Oxford University Press.

Palmer, A. (2010). *Social minds in the novel*. Columbus, OH: Ohio State University Press.

Rabinowitz, P. J. (1987). *Before reading: Narrative conventions and the politics of interpretation. The theory and interpretation of narrative series*. Ithaca, NY: Cornell University Press.

Rabinowitz, P. J., & Bancroft, C. (2014). Euclid at the core: Recentering literary education. *Style*, 48(1), 1–34.

Schneider, R. (2001). Toward a cognitive theory of literary character: The dynamics of mental-model construction. *Style*, 35, 607–639.

Schneider, R. (2013). The cognitive theory of character reception: An updated proposal. *Anglistik*, 24(2), 117–134.

Scholes, R., Phelan, J., & Kellogg, R. (2006). *The nature of narrative* (Revised and expanded, 40th anniversary edition ed.). New York: Oxford University Press.

Sperber, D., & Wilson, D. (2002). Pragmatics, modularity and mindreading. *Mind and Language*, 17, 3–23. doi: 10.1111/1468-0017.00186

van Duijn, M. J., Sluiter, I., & Verhagen, A. (2015). When narrative takes over: The representation of embedded mindstates in Shakespeare's Othello. *Language and Literature*, 24(2), 148–166. doi: 10.1177/0963947015572274

van Peer, W., Hakemulder, F., & Zyngier, S. (2007). Lines on feeling: foregrounding, aesthetics and meaning. *Language and Literature*, 16(2), 197–213. doi: 10.1177/0963947007075985

Whalen, D. H., Zunshine, L., & Holquist, M. (2012). Theory of Mind and embedding of perspective: A psychological test of a literary "sweet spot". *Scientific Study of Literature*, 2(2), 301–315. doi: 10.1075/ssol.2.2.06wha

Whalen, D. H., Zunshine, L., & Holquist, M. (2015). Perspective embedding affects reading time: Implications for the reading of literature. *Frontiers in Psychology*, 6(1778), 1–9.

Zunshine, L. (2006). *Why we read fiction: Theory of mind and the novel*. Columbus: Ohio State University Press.

Zunshine, L. (2011). Style brings in mental states. *Style*, 45, 349–356.

Zunshine, L. (2014). Theory of Mind as a pedagogical tool. *Interdisciplinary Literary Studies*, 16, 89–109. doi: 10.5325/intelitestud.16.1.0089

Zunshine, L. (2015a). From the social to the literary: Approaching Cao Xueqin's The Story of the Stone (紅樓夢) from a cognitive perspective. In L. Zunshine (Ed.), *The Oxford handbook of cognitive literary studies* (pp. 176–196). Oxford: Oxford University Press.

Zunshine, L. (2015b). The secret life of fiction. *PMLA*, 130, 724–731. doi: 10.1632/pmla.2015.130.3.724

Appendix. First three paragraphs of "To Kill a Mockingbird" (Lee, 1960: 9)

Our sentence numbers (which were not presented to the participants) are shown before each sentence in brackets.

[1] When he was nearly thirteen, my brother Jem got his arm badly broken at the elbow. [2] When it healed, and Jem's fears of never being able to play football were assuaged, he was seldom self-conscious about his injury. [3] His left arm was somewhat shorter than his right; [4]when he stood or walked, the back of his hand was at right angles to his body, his thumb parallel to his thigh. [5] He couldn't have cared less, so long as he could pass and punt.

[6] When enough years had gone by to enable us to look back on them, we sometimes discussed the events leading to his accident. [7] I maintain that the Ewells started it all, but Jem, who was four years my senior, said it started long before that. [8] He said it began the summer Dill came to us, when Dill first gave us the idea of making Boo Radley come out.

[9] I said if he wanted to take a broad view of the thing, it really began with Andrew Jackson. [10] If General Jackson hadn't run the Creeks up the creek, Simon Finch would have never paddled up the Alabama, and where would we be if he hadn't. [11] We were far too old to settle an argument with a fist-fight, so we consulted Atticus. [12] Our father said we were both right.

Authors' addresses

D. H. Whalen
Program in Speech-Language-Hearing Sciences
City University of New York
365 Fifth Ave.
New York, NY 10016
USA

dwhalen@gc.cuny.edu.

Lisa Zunshine
Department of English
University of Kentucky
USA

Evelyne Ender
French Program
Johns Hopkins University
USA

Eugenia Kelbert
Department of Slavic Literatures and Cultures
Universität Passau
Germany

Jason Tougaw
Department of English
Queens College
USA

Robert F. Barsky
Department of French and Italian
Vanderbilt University
USA

Peter Steiner
Emeritus, Department of Slavic Languages and Literature
University of Pennsylvania
USA

Michael Holquist
Emeritus, Department of Comparative Literature
Yale University
USA