Can we agree about *agree*?*

Emmanuel Chemla† and B. R. George‡

DRAFT of December 2014
comments welcome – consult authors before citing

Abstract

This squib attempts to constrain semantic theories of *agree wh* constructions, by broadening the data set and collecting naive speakers' intuitions. Overall, our data suggest relatively permissive truth-conditions for these constructions. They also suggest a previously undiscussed presupposition for *agree wh* and also indicate that *agree wh* is not straightforwardly reducible to *agree that*. Although some accounts suggest differences in truth conditions among different asymmetrical *agree with* constructions and symmetrical *agree* constructions, we do not find any indication of such truth-conditional distinctions. In the course of our exploration of the data, we offer a new approach to distinguishing between truth, falsity and presupposition failure.

Keywords: embedded questions, reducibility, presupposition, quantitative data

1 Embedded questions, reducibility, and *agree wh*

The goal of this squib is to constrain semantic theories of *agree wh* constructions, by broadening the data set and collecting naive speakers' intuitions.

That is, we are concerned with the truth-conditions of sentences like those in (1), and in particular their truth or falsehood in the conditions represented by the pictures like those in (2). The pictures represent on different lines John's beliefs (**j**) and Mary's beliefs (**m**), in which white letters in gray boxes encode ignorance about the color of that letter (e.g., in (2)c, neither John nor Mary knows what colors C and D are).[1]

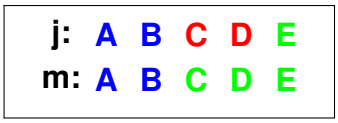
We wish to thank Alexandre Cremers, Paul Egré, Michael Franke, Manuel Križ and Benjamin Spector. The research leading to these results was supported by a 'Euryi' grant from the European Science Foundation ("Presupposition: A Formal Pragmatic Approach"), the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n.313610, ANR-10-IDEX-0001-02 PSL and ANR-10-LABX-0087 IEC.

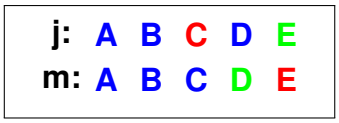
†Ecole Normale Supérieure, Laboratoire de Sciences Cognitives et Psycholinguistique. chemla@ens.fr

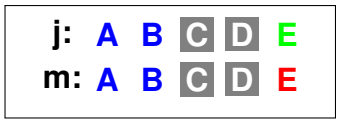
‡Department of Philosophy, Carnegie Mellon University. bergeorge@marblesandunicorns.net

[1] We use blue as the most important color to be distinguished from the others in all the examples of the paper. We hope this will minimize the difficulty for most readers, including those with color vision disabilities. We will offer additional descriptions in the text whenever more needs to be known.

(1) a. John and Mary agree on which letters are blue.
b. John and Mary do not agree on which letters are blue.

(2) a. 

b. 

c. 

One interest of these types of cases is that they bear on the question of the *reducibility* of *agree wh* to *agree that*. That is, of whether the semantics of *agree* ascriptions with embedded questions can be defined in terms of the notion of propositional agreement and the notion of what constitutes an answer to a question. This issue is interesting as an empirically contentious case of the general issue of reducibility for responsive attitude predicates (in the terminology of Lahiri 2002), which is in turn a natural generalization of the question of reducibility for *know wh*.

In this introductory section, we will first discuss in more details the issue of reducibility (§1.1); we will then introduce two more questions that arise for the semantics of agreement: opinionatedness (§1.2) and presuppositions (§1.3). Our goal will be to identify contentious empirical cases which we will test in the remainder of this squib.

1.1 Reducibility and *agree*

In the epistemic and semantic literature on *know wh* constructions, attention has focused on the relationship between *know wh* and *know that*, with an emphasis on the viability of the (intuitively appealing) project of reducing the former to the latter. The natural starting assumption regarding *know wh* ascriptions like (3) is that their truth-conditions are reducible to facts about propositional knowledge of answers – that is, to the kinds of facts reported by statements like those in (4):

(3) John knows which letters are blue.

(4) a. John knows that C is blue.
b. John knows that B and C are the (only) letters that are blue.

A representative statement of the reductive approach can be found in Higginbotham (1996), who says that 'a V like "know" may be mediated, in its use with indirect questions, by the ordinary knowledge relation between agents and propositions', giving the formula (5) as a possible rendering of the conditions under which a person x stands in the *know* relation to a question π.

(5) $know(x, \hat{}\pi) \leftrightarrow (\exists p)(know(x, p)\ \&\ p \text{ answers } \pi)$

That is, *know* relates a subject to a question if it relates the subject to at least one answer to that question. On this view, the main problem in the semantics of *know wh* is the explication of the notion of answer.

A slightly more general statement of the reductive principle, abstracting away from Higginbotham's particular existential quantification schema, might be given by (6):

(6) For every a and every question π, whether a knows π depends only on what answers to π a knows.

Reductive accounts of *know wh*, treating it as analyzable in terms of something like (5), but employing different notions of answerhood, have been popular in the formal semantics literature: such treatments (or treatments which can be reformulated in such terms) include Karttunen (1977), Groenendijk and Stokhof (1984), Beck and Rullmann (1999), Sharvit (2002), and Lahiri (2002), among many others. A recent linguistic challenge to reducibility of *know wh* can be found in George (2011, 2013).

In the epistemic literature, this kind of reducibility has met with more skepticism, having been challenged by Ryle (1945) (for *know how* in particular) and Schaffer (2007) (for question-knowledge generally). Pro-reducibility replies to these lines of thought include those found in Stanley (2011).

The problem of reducibility for *know* is a special case of a more general problem of reducibility for attitudes that are *know*-like in that they can embed both interrogative-type clauses and declarative-type clauses, including *forget*, *learn*, *be sure*, and *agree*.[2] For any attitude R for which both question-oriented and propositional uses derive we could state the following reducibility hypothesis:

(7) For every a and every question π, whether a stands in relation R to π depends only on what answers to π are such that a stands in relation R to them.

We might further hope that there is a general principle at work, which states that the reducibility hypothesis holds for any choice of R.

In the case of agreement, where at least two individuals are involved, the natural formulations are given in (8):

(8) a. For every plural collection of individuals A and every question π, whether A agree on π depends only on what answers to π are such that A agree that those answers are true.
b. For all a and b, and every question π whether a agrees with b that π depends only on what answer to π are such that a agrees with b that those answers are true.

These are special cases of (7), saying for *agree* something analogous to what (6) says for *know*.

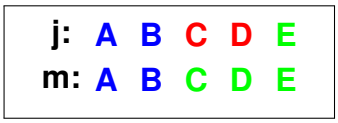
The interest of the reducibility principles in (8) is that they are an interesting test case for the general principle (7), and also that the question of their truth or falsehood is an interesting and important question in what we might call epistemology of agreement, just as the reducibility problem for *know wh* is of considerable interest in the more familiar epistemology of knowledge. Is there something about agreement *wh* that is special to questions, or is it that everything special to agreement can be seen in agreement *that*, while agreement *wh* is just the result of combining

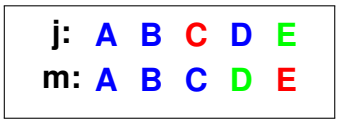
[2]These must be distinguished from embedders like *ask* and *wonder*, which cannot embed propositional clauses and are so analyzed as fundamentally question-oriented, and embedders like *believe*, which cannot embed interrogative-type clauses. For these classes, the question of reducibility, at least in the sense discussed here, does not arise.

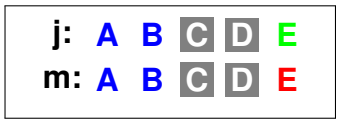
agreement *that* with general reduction principles?

The reason that *agree* is of special interest is that there are live (but not uncontested) proposals in the semantics literature which imply that *agree wh* is not reducible to *agree that*. To see the issue let us return to (2):

(1) a. John and Mary agree on which letters are blue.
b. John and Mary do not agree on which letters are blue.

(2) a. 

b. 

c. 

Compare (b) and (c) of (2). These two have in common that they make the *agree that* ascriptions in (9) true, while they fail to satisfy the agreement ascriptions in (10)

(9) a. John and Mary agree that A is blue.
b. John and Mary agree that B is blue.
c. John and Mary agree that E is not blue.

(10) a. John and Mary agree that A is not blue.
b. John and Mary agree that B is not blue.
c. John and Mary agree that C is blue.
d. John and Mary agree that C is not blue.
e. John and Mary agree that D is blue.
f. John and Mary agree that D is not blue.
g. John and Mary agree that E is blue.

Since answers to the question *which letters are blue?* are plausibly all sentences in (9) or (10), or boolean combinations thereof, we find that (2)b and (2)c put John and Mary in the *agree that* relation with exactly the same answers to the embedded question. That is, if reducibility for *agree* is right, (1-a) should be true in the circumstances represented by (2)b if and only if it is true in the circumstances represented by (2)c.

Is this the case? Here the intuitions diverge, and we find that the literature is divided. Notably, Beck and Rullmann (1999) offer a semantics of *agree wh* in which (1-a) is true in (c) but not in (b), and the same result is derived by one reading of *agree* generated by the semantics of Lahiri (2002). If this empirical generalization, or one like it, is right, then reducibility does not hold for *agree*, and so the general reducibility hypothesis for arbitrary *know*-like attitudes is false.

On the other hand, some sources in the literature reject this empirical picture. The theories of *agree wh* in Egré and Spector (2007, 2014) and George (2011) both treat the two cases discussed above the same, and this reduction-friendly picture is also what is predicted by the brief remark on *agree wh* in Kratzer (2006). If the empirical picture is compatible with these accounts, then *agree* does not present any special challenge to reducibility.

1.2 *agree* and partial information: some competing theories

The contentious cases for *agree wh* – those that are disputed in the literature and that call reducibility into question – all involve partial information. That is, they involve cases in which at least one of the people to whom agreement is being attributed has incomplete beliefs about the question, as in (2)c.

Available accounts of *agree wh* all hold that when both John and Mary are fully opinionated about the color of all letters, (1-a) will be true if John and Mary believe the same letters to be blue, and untrue if there is any letter that one of them believes is blue and the other believes is not blue. Proposals differ in their handling of cases where one or both of the characters lack an opinion about whether one or more of the letters are blue. We will identify three general types of accounts.

1.2.1 Full Alignment (FA) accounts

Beck and Rullmann (1999) offer what we call a Full Alignment (FA) account (this account is also reimplemented by Sharvit 2002 using different theoretical machinery). On an FA account, (1-a) is true iff both characters have the same positive and negative beliefs about whether each letter is blue. That is, whenever one character believes a letter is blue, the other believes that letter is blue as well, and whenever one believes a letter is not blue, the other likewise believes that it is not blue. There is no requirement that the characters be opinionated about any or all of the letters, so on an FA account it is perfectly permissible that the characters be unopinionated, as long as they are unopinionated about the same letters. That is, in (11) on an (FA) account (1-a) will be true in situations (a) and (b), but not in situations (c) and (d). Beck and Rullmann (1999) do not discuss the *agree with* construction.

(11)

a.
j: A B C D E
m: A B C D E

b.
j: A B C D E
m: A B C D E

c.
j: A B C D E
m: A B C D E

d.
j: A B C D E
m: A B C D E

1.2.2 Opinionated Full Alignment (OFA) accounts

Egré and Spector (2007, 2014), George (2011), and Kratzer (2006) all advocate an Opinionated Full Alignment (OFA) picture, implemented in terms of the notion of a strongly exhaustive answer.[3] Descriptively, an OFA account adds to an FA account an additional condition that both characters must be fully opinionated. Thus, on an OFA account, (1-a) would be true in situation (a) of (11), but not in situations (b), (c), and (d).

Some advocates of OFA accounts, namely Egré and Spector (2007, 2014), also discuss the *agree with* construction, as presented in (12). They associate it with identical truth conditions, but

[3]Lahiri (2002) offers a general implementation of strong exhaustivity (Lahiri's formula (225) in section 3.6.1) that we believe would make similar or identical predictions, given Lahiri's other assumptions about *agree*.

different presuppositions than the plural *agree* construction. (Egré and Spector (2007, 2014) are also the only OFA advocates that discuss presuppositions at all, an issue we come back to in §1.3).

(12) a. John agrees with Mary on which letters are blue.
b. Mary agrees with John on which letters are blue.

1.2.3 Positive Superset (PS) accounts

Finally, Lahiri (2002) treats *agree with* sentences such as (12) as basic. Lahiri's semantics of *agree* and of question-embedding yield what we call a Positive Superset (PS) account as one possible reading of such sentences.[4] On this account, (12-b) is true iff every letter that Mary believes is blue is one that John also believes is blue. Hence, on a PS account, (12-b) is true in situations (a), (b), (c), and (d) of (11), while (12-a) is true only in situations (a), (b), and (c). (1-a) is analyzed as equivalent to the conjunction of (12-b) and (12-a), so it would be true in situations (a), (b), and (c), but not situation (d).

1.2.4 Distinguishing accounts by testing tolerable levels of ignorance

One of our major goals is to sort out what level of ignorance is compatible with the truth of *agree wh* attributions, and our test cases were chosen with this issue in mind. This has broader implications for the architecture of a general theory of questions and attitudes, and of course it closely connects to reducibility, in that the PS and FA pictures are incompatible with reducibility, while the OFA picture is reductive. We further sought to determine whether Lahiri's predicted truth-conditional differences between (12-a), (12-b), and (1-a) were borne out.

1.3 *agree* and presupposition

The presuppositions associated with *agree wh* are under-explored. There is general agreement that there is no factivity effect associated with question embedding under *agree*, but beyond that little is clear. In Beck and Rullmann (1999) the presuppositions (or lack thereof) of *agree wh* are not discussed. In Lahiri (2002), although the accommodation of an opinionatedness presupposition for *agree that* drives some aspects of the semantics of *agree wh*, the presuppositions of *agree wh* receive little direct attention. Among OFA accounts, Egré and Spector (2007, 2014) suggest that *agree wh* has some quite strong presuppositions regarding the opinionatedness of the characters, while George (2011) and the brief remark in Kratzer (2006) do not explore the issue.

Although we did not set out primarily to assess the presuppositions associated with *agree wh* ascriptions, our data reveal certain patterns which we think are reasonably be characterized as presuppositional, and it is our hope that our results may help to shed a little light on the presuppositions of *agree wh*, and their connection with (non-)opinionatedness.

[4]This is something of an extrapolation of Lahiri's brief discussion of agree. Lahiri treats all question-embedding as richly ambiguous, but we believe that this is a faithful rendering of applying Lahiri's implementation of Ans1-style weak exhaustivity to his assumptions about the semantics of agreement.

1.4 Summary and Plan

In order to evaluate both the empirical landscape of *agree wh*, and in particular the merits of competing accounts and the prospects of reductive accounts, we selected a reasonable assortment of picture types involving different kinds of misaligned belief, with an emphasis on those that would help to probe different kinds of mismatch between available accounts, and with an emphasis on cases where one or more of the available accounts made predictions that were incompatible with the reducibility of *agree wh* to *agree that*. In section 2, we describe our experimental design in more detail.

At the level of semantic theory, our results will address a number of issues. First of all, our data help to identify the empirical strengths and weaknesses of the OFA, FA, and PS accounts discussed above. We also observe some apparent presuppositional effects with *agree wh* that, although not really surprising, have not been previously discussed. More abstractly we note that the patterns observed are incompatible with a traditional reductive picture of the realtionship between *agree wh* to *agree that*.

At the methodological level, we suggest a new way to distinguish between truth, falsity and presuppositional failure. Concretely, we used the judgments given to positive and negative sentences in (1) to algorithmically cluster the pictures into three sets. The sets we obtain largely correspond to true situations ((1-a) is true and its negation (1-b) is false), false situations ((1-a) is false, (1-b) is true) and presupposition failure (where the distinction between (1-a) and (1-b) is less clear).

The plan of the paper is as follows: having laid out the preliminaries above, we describe our experimental method in (§2) and our results (§3). We immediately raise and discuss the possibility of a low level interpretation of our results (§4.1). We argue that such an interpretation is unlikely (although not impossible) and thus move on to discuss the implications for competing theories (§4.2), presuppositional matters (§4.3), and reducibility (§4.4). Finally, in §5, we briefly summarize what we have done, and suggest some problems for future research.

2 Experimental method

2.1 Task

We asked participants to provide truth-value judgments ('not true' vs. 'true') about sentences (including (1-a) and its negation (1-b)) based on visual representations of John's and Mary's beliefs about the color of five letters. The visual code that we explained above was explained to participants through a list of simple examples.

2.2 Participants

We recruited 131 participants through Mechanical turk. We excluded three participants from the analyses because they were not native speakers of English. Our analyses are thus based on four groups of 32 participants: each group was tested using one of the four constructions in (13) for

the target sentence, and their corresponding negations in (14). We obtained no relevant difference between these constructions and therefore present all the results aggregated.

(13) a. John and Mary agree on which letters are blue.
b. John and Mary agree about which letters are blue.
c. John agrees with Mary about which letters are blue.
d. Mary agrees with John about which letters are blue.

(14) a. John and Mary do not agree on which letters are blue.
b. John and Mary do not agree about which letters are blue.
c. John does not agree with Mary about which letters are blue.
d. Mary does not agree with John about which letters are blue.

2.3 Stimuli and procedure

After a set of instructions, a sentence was presented on the screen below an image. The sentence-picture items can be divided in two groups: control and target items. We describe these two groups of items sequentially, but all items were presented in random order to the participants. We start with a bit of terminology to describe our conditions.

Terminology

For each character, each letter could be of the three following types: (**T**) of the target color (the color mentioned in the sentence), (**N**) of a nontarget, randomly chosen color, or (**U**) of an uncertain color for this character. The status of a given letter for both characters could then be described as a pair. For instance, the letter A could be **T**-**U**, which would mean that one character believes A is of the target color while the other had no opinion about the color of A, or it could be **N**-**N**, which means that each character believes that A is of some randomly chosen non-target color (possibly a different one).

2.3.1 Control items: *opinion* sentences

Control items were used to test whether participants understood and took into account the visual codes established to represent beliefs (and absence of belief). These items combined sentences of the form of (15-a) and (15-b) with three types of images. With blue as the target color and B as the target letter as in (15) then John thought B was blue for Ctrl-1, B was some other color for Ctrl-2, and John had no opinion about the color of B for Ctrl-3, see Table 4 for illustration and results (using the terminology introduced above, according to John B was either **T**, **N** or **U**, respectively). Each participant saw 4 positive and 4 negative versions of each of these three control conditions, in which the color and letter were chosen at random on each trial.

(15) a. John has an opinion about whether B is blue.
b. John does not have an opinion about whether B is blue.

2.3.2 Target items: *agree wh* sentences

Participants were presented with *agree wh* sentences, as described in (13), paired with a pseudo-randomly generated image that would fit one of 13 conditions. These conditions are described in Tables 1, 2 and 3 mostly by way of illustration and by indicating the number of letters of different types.

The conditions were selected to present reasonable range of combinations of **T**-**T**, **T**-**U**, **N**-**N**, **N**-**U**, and **U**-**U**. Less emphasis was placed on **T**-**N**, because all available theories agreed that even a single **T**-**N** letter was incompatible with agreement, but one condition of this kind was included for completeness. We placed a particular emphasis on picking conditions in which the theories differed, in which one or more theories made suspicious or disputed predictions, or in which predictions diverged from what one would expect under reducibility. We avoided cases in which either characters beliefs involved exactly one letter of the target color, to minimize the confounding effect of the plural inflection of *which letters*.

Name and illustration		Description (number of letters of each type)				
		T-T	N-U	U-U	N-N	Comments
1a	j: A B C D E m: A B C D E	2			3	Fully opinionated; identical beliefs.
1b	j: A B C D E m: A B C D E	2		3		Both uncertain; identical beliefs.
1c	j: A B C D E m: A B C D E	2		2	1	Both uncertain; identical beliefs.
1d	j: A B C D E m: A B C D E	2	3			One uncertain about other's Ns.
1e	j: A B C D E m: A B C D E	2	2		1	Each uncertain about one of the other's Ns.
1f	j: A B C D E m: A B C D E	2	1		2	One uncertain about one of other's Ns.

Table 1: For all these conditions, if one character believes that a letter is of the target color, the other one also does. And there were two letters of the target color.

The sentence could be either positive or negative. We thus obtained 2(positive/negative) $\times$ 13(conditions) $=$ 26 such target cases. Each of these conditions was presented three times to each participant when the two characters played symmetrical roles (1a,b,c,e; 2a; 3a,b,c), or 6 times to vary the role played by each character in cases in which that would make a difference (1d,f;2b,c,d).[5]

[5]There was also some structure to the three repetitions. (i) In one repetition the colors were assigned so that target colors were attributed to the first letters, unknown colors to the next letters, and the non-target colors to the final letters; (ii) in another repetition the colors were assigned in the opposite manner (non-target, unknown, target); (iii) and in

Name and illustration	Description (number of letters of each type)					
	T-T	N-U	T-U	N-N	T-N	Comments
2a j: A B C D E m: A B C D E	1		2	2		Each uncertain about one of the other's Ts.
2b j: A B C D E m: A B C D E	2		1	2		One uncertain about one of the other's Ts.
2c j: A B C D E m: A B C D E	2	2	1			One uncertain about other's Ts and Ns.
2d j: A B C D E m: A B C D E	2			2	1	Fully opinionated; direct disagreement.

Table 2: For all these conditions, there is a letter that one character believes it is of the target color, while the other one believes it is of some other color or is uncertain about its color

Name and illustration	Description			
	T-T	N-N	U-U	Comments
3a j: A B C D E m: A B C D E		5		Fully opinionated; identical beliefs; no Ts.
3b j: A B C D E m: A B C D E		2	3	Both uncertain; identical beliefs; no Ts.
3c j: A B C D E m: A B C D E			5	Both uncertain; totally unopinionated.

Table 3: For all these conditions, there was no letter believed to be of the target color

Each participant thus saw 2(positive/negative) $\times$ [8(symmetrical conditions) $\times$ 3(repetitions) + 5(asymmetrical conditions) $\times$ 3(repetitions) $\times$ 2(counterbalancing the roles)] = 108 target items.

3 Results

3.1 Control 'opinion' results

In Table 4 we report the results for the control items. The representation we adopt is such that true and false sentences can be detected easily: red and right indicate truth, while blue and left indicate falsity. Participants accurately identified *John has an opinion about whether X is blue* as true in the first two cases (Ctrl-1 and Ctrl-2) and false in the third one (Ctrl-3). This can be seen from the length of the red line which represents the average percentage of *true* responses to

another repetition the positions of the colors were randomly assigned (so A, C and E could be in the target color, without B or D being in the target color).

these positive sentences (percentages being calculated per subject first). The blue bars represent the average percentage of *true* answers to the negative version of the sentence (*John does not have an opinion about whether X is blue*). We can see that the pattern is appropriately reversed here: the negative sentence was judged true in case the positive sentence was judged false and *vice versa*.

Statistically, we ran logit mixed-model analyses with subject and experimental item entered as random factors without including a slope, for the sake of simplicity. These analyses confirm that Ctrl-1 and Ctrl-2 (grouped together for this analysis) received more true responses than Ctrl-3 for the positive sentence ($\beta = 5.3, z = 24, p < .001$) and fewer for the negative sentence ($\beta = 5.8, z = -24, p < .001$).

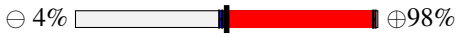
Situation		Opinion claim John has/does not have an opinion...	Correct answer	Results
Ctrl-1	j: A B C D E m: A B C D E	... about whether A is blue		
Ctrl-2	j: A B C D E m: A B C D E	... about whether E is blue		
Ctrl-3	j: A B C D E m: A B C D E	... about whether C is blue		

Table 4: Results for control sentences. The length of the red bar corresponds to the percentage of true answers given to the positive sentence (also given at the end of this bar), the length of the blue bar spreading in the opposite direction corresponds to the percentage of true answers given to the negative sentence (also given at the end of this bar).

3.2 Target 'agree wh' results: difference between constructions

We found no difference for sentences such as *John and Mary agree...* and *John agrees with Mary ...* sentences, no effect of which character was giving a particular role in asymmetrical cases, no effect of which character was mentioned first in the sentence and no effect of the preposition (agree *on* which or agree *about* which): in mixed-model analyses with subject, condition and polarity of the sentence as random factors we obtain all $|z|$s < 1.2, ns.

This null-result is of theoretical importance for Lahiri (2002) predicts that these constructions could differ in conditions 2b, 2c and 2d as presented in Table 2 (results in Table 6). As discussed in (12), 'Mary agrees with John on/about which letters are blue' is predicted to be true in these situations, while 'John agrees with Mary on/about which letters are blue' and 'John and Mary agree on/about which letters are blue' are predicted to be false. Yet, each of these sentences in each of these three situations were judged false between 88% and 90% of the time and their negations were judged true between 81 and 87% of the time. Hence, all of these sentences are best understood as false, including 'John agrees with Mary on/about which letters are blue' which does not differ from the other.

Concretely, in the absence of observable differences, we collapsed the results along these dimensions for the following analyses.

3.3 Target 'agree wh' results: difference between conditions

Results for the conditions involving the *agree wh* target sentences are presented in Tables 5, 6 and 7, following the presentations of the conditions in Tables 1, 2, 3, respectively. Impressionistically, we can see that conditions in Table 1/5 give rise to true judgments for the positive sentences and false judgments for their negations ($\beta = 8.9, z = 36, p < .0001$, with subject, experimental item and conditions as random factors). Conditions in Table 2/6 give rise to the reverse pattern ($\beta = 1.8, z = -47, p < .0001$). Finally, conditions in Table 3/7 yield somewhat intermediate ratings for positive and negative sentences alike (positive sentences actually receive more true responses, $\beta = 1.8, z = 18, p < .001$, but to a smaller extent than for the set of conditions, as revealed by a significant interaction between the two effects: $\beta = -6.6, z = -29, p < .001$). These results suggest that the (positive) sentence is true in the first set of conditions (Table 5), false in the second (Table 6), and a presupposition failure in the last set (Table 7).

At this point, however, these results rely on an arbitrary classification of the conditions. In other words, so far we could have grouped the conditions into three different sets as it fits our needs, and then run the above computations, which would simply confirm that our grouping was effective. We would rather find an objective way to group the conditions into different groups first. In fact, this is what we have been doing all along. The grouping that we have used so far is itself the result of an objective analysis of the data, we used it right away for simplicity. In the following section, we thus present this methodology which delivers an objective analysis of results of the kind we present, allowing to categorize sentences as true, false or presupposition failure.

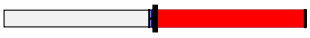
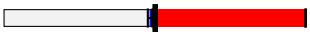
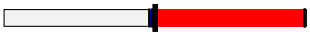
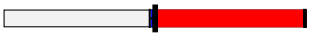
Name and illustration		Predictions			Results
		OFA	FA	PS	
1a	j: A B C D E m: A B C D E	👍	👍	👍	⊖ 3%  ⊕99%
1b	j: A B C D E m: A B C D E	👎	👍	👍	⊖ 3%  ⊕99%
1c	j: A B C D E m: A B C D E	👎	👍	👍	⊖ 2%  ⊕98%
1d	j: A B C D E m: A B C D E	👎	👎	👍	⊖ 3%  ⊕99%
1e	j: A B C D E m: A B C D E	👎	👎	👍	⊖ 2%  ⊕100%
1f	j: A B C D E m: A B C D E	👎	👎	👍	⊖ 3%  ⊕98%

Table 5: For all these conditions, if one character believes that a letter is of the target color, the other one also does. And there were two letters of the target color. The OFA, FAand PScolumns schematize the predictions of Egré and Spector (2007, 2014), Beck and Rullmann (1999), and Lahiri (2002) respectively. See explanations of the bars in Table 4, but in short: the more to the right, the more the (positive) sentence was judged true in this situation. Predictions for the *J and M agree* sentences are shown. Predictions for *agree with* sentences are the same as these when available, but *agree with* is not addressed by Beck and Rullmann (1999).

Name and illustration		Predictions			Results
		OFA	FA	PS	
2a	j: A B C D E m: A B C D E				⊖ 88%  ⊕6%
2b	j: A B C D E m: A B C D E				⊖ 82%  ⊕12%
2c	j: A B C D E m: A B C D E				⊖ 83%  ⊕13%
2d	j: A B C D E m: A B C D E				⊖ 91%  ⊕6%

Table 6: For all these conditions, there is a letter that one character believes it is of the target color, while the other one believes it is of some other color or is uncertain about its color. See explanations of the bars in Table 4, but in short: the more to the right, the more the (positive) sentence was judged true in this situation. Predictions for the *J and M agree* sentences are shown. Predictions for *agree with* sentences are the same as these when available, except that in some cases Lahiri (2002) predicts variable truth value for the *agree with* cases, which is indicated with with the parenthetical ( /) below the main prediction.

Name and illustration		Predictions			Results
		OFA	FA	PS	
3a	j: A B C D E m: A B C D E				⊖ 12%  ⊕71%
3b	j: A B C D E m: A B C D E				⊖ 25%  ⊕54%
3c	j: A B C D E m: A B C D E				⊖ 27%  ⊕47%

Table 7: For all these conditions, there was no letter believed to be of the target color. See explanations of the bars in Table 4, but in short: the more to the right, the more the (positive) sentence was judged true in this situation. Predictions for the *J and M agree* sentences are shown. Predictions for *agree with* sentences are the same as these when available.

3.4 Clustering analyses

Facing this type of data on positive and negative sentences in different conditions, a traditional truth conditional investigation needs to differentiate the 'true', the 'false' and the presupposition failure situations. Based on the ratings we obtained for positive and negative sentences, we ran a blind clustering algorithm which categorized situations in three groups. We applied the k-mean algorithm (using the *kmeans* function from the R software) to sort the data in three groups. The algorithm was given as input for each condition and each participant the pair of their mean answers to the positive and to the negative versions of the target sentence in this condition. The algorithm then categorized conditions according to these pairs of ratings it received. Unsurprisingly at this point, the three groups of conditions described in Tables 5, 6 and 7 were extracted.[6]

We also ran the algorithm adding the control *opinion* conditions to the set of conditions to be categorized. The target conditions were distributed in the same way across the three sets. Interestingly, the Ctrl-1, Ctrl-2 'true' conditions were categorized in the first group, and the Ctrl-3 'false' condition were categorized in the second group. This new clustering thus provides (or confirms) labels for these sets: the first set contains conditions leading to 'true' answers, and the second set contains conditions leading to 'false' answers.[7,8,9]

[6]Some algorithms exist to determine *a priori* how many groups should be formed. However, these algorithms rarely provide a clear answer (we ran four such algorithms and obtained: 2, 3, 4 and 8 as possible optimal number of clusters). These algorithms are best suited in situations where there is no *a priori* reason to choose a particular number of clusters. In our case, we had a priori reasons to look for three clusters: we wanted to sort the situations into true, false and presupposition failure.

[7]Labels 'true' and 'false' correspond to the positive versions of the sentence.

[8] If the algorithm is not given pairs of ratings but is applied to positive and negative sentences separately, then it delivers three groups as follows:

- conditions in Table 5, Ctrl-1 and Ctrl-2 lead to the positive sentence being classified in group 1, and the negative sentence in group 2;
- conditions in Table 6 and Ctrl-3 lead to the reverse: positive in group 2, negative in group 1;
- conditions in Table 7: positive in **group 3**, negative in group 2.

Overall, the pattern is the same. The first group recovered by the algorithm corresponds to true sentence-situation pairs, the second group to false sentence-situation pairs, and the third group corresponds to judgments obtained for positive sentence in case of presupposition failure. So, positive sentences seem to be affected by presupposition failure more visibly in this paradigm and to give rise to a new category. Negative sentences just get judged as the false sentences (by means of local accommodation, it should be judged 'true').

[9]We did not have controls to label the third category as 'presupposition failure.' However, we may compare the current study to the investigation of Križ and Chemla (2014), which explores more systematically several ways to detect truth value gaps. It is applied to so-called 'homogeneity' of plural definite descriptions. In a pilot experiment, Križ and Chemla found a similar pattern of responses: true situations led to clearly true responses for the positive sentence and clearly not-true responses for the corresponding negative sentence, the reverse was true for false situations, and situations in which the homogeneity assumption was not met led to somewhat intermediate types of answers for both positive and negative sentences. This is not entirely sufficient for several reasons. First, this was done in a different setting with different participants. Second, homogeneity may not be presuppositional to begin with. But note that there may actually be different *kinds* of presupposition, leading to different reactions for presupposition failures anyway. In trying to think about presupposition controls to add, we thought about definite descriptions, factive verbs and other presupposition triggers, but quickly realized that they may not form a uniform class. What is crucial for the current

4 Discussion

4.1 A low level interpretation of the results

In experimental situations, one major concern is that results may not correspond to the semantic phenomenon that we are after (truth conditions), but may instead be a by-product of a specific strategy that could, e.g., help the participants carry out the task efficiently. Given our results, one particular strategy comes to mind. Imagine that upon encountering a sentence such as *John and Mary agree on which letters are blue*, our participants looked for blue letters in the display, and asked for each of them whether the other letter in the same column is also blue. If the answer is yes for all blue letters, they would classify the sentence as true, and otherwise they would classify it as untrue. Negative sentences would be treated in a similar fashion. If participants uniformly applied this procedure, it would yield our two sets of true and false conditions. One could even try to account for our 'presuposition failure' cases in terms of such a strategy: In these cases, the search may be made void by the fact that no blue letter is found, and, in the absence of a blue letter, participants would be puzzled and report noisy judgments, which might be misidentified as an indication of 'presupposition failure'.

It would be most convincing to either confirm our data using a different methodology that would not be subject to the same strategical approach or to explore independent ways of rejecting this interpretation. For instance, one could reason about the possible distinctive consequences of this strategy on some other dimension, e.g., response times, and try to evaluate whether actual response time patterns raise or decrease the plausibility of this interpretation. At this point, we will provide only partial counter-arguments. First, it should be highlighted that we have no positive evidence that this strategy is being employed. Second, the results obtained accord reasonably well with intuition. Finally, we found evidence that in presupposition failure cases, positive and negative sentences were not treated alike (see footnote 8). In other words, responses are sensitive to a linguistic aspect of the sentence (the presence of a negation) in cases where the *picture* makes the described strategy void. Accounting for such an effect thus requires to complement the (mostly-)picture-based strategy with a sentence-based component. The attempt of explaining away the current facts by appealing to strategies is therefore not complete.

More concretely, strategies are likely to develop across the course of an experiment, and one may think that this is particularly so with an experiment like the current one which uses the same sentences repeatedly. We therefore checked that our results were not the result of a strategy that developed late in the course of the experiment in two ways.[10] First, we computed the average rating in all conditions for positive and negative sentences based on the first 10 answers only. The Pearson correlation cœfficient between these values and the average across the whole experiment was 98%, thus suggesting that everything was in place after a couple of trials and before strategies could be developed. Second, we ran our clustering analyses on the results obtained for the first trials. We

purposes then is that the third group behaves differently than clearly true and clearly false cases, making it a 'not-true and not-false' group, and therefore plausibly a presupposition failure group.

[10]Thanks to Michael Franke for suggesting analyses of early responses.

could not use the first 10 trials of each participant for this analysis because that would have left too many empty cells. Instead, we used the first *15* answers and put together participants in 10 groups of approximately 13 participants. Based on the first answers of these 10 'super-participants', we obtained exactly the same clustering of the conditions. Overall then, these additional analyses reveal that our results are obtained even if we focus our attention to the beginning of the experiment, when 'strategies' are less likely to have developed.

Overall, the existence of strategies is a worry we do not want to treat lightly. Our hope is that even if noise was introduced in our data through this mean, it should not prevent the questions we have tried to raise to reach their audience. We also hope that others could judge our empirical claims by themselves or complement them with other investigations. We will continue the discussion in the hope that the current counterarguments give sufficient credibility to the current results.

4.2 Consequences for theories of *agree wh*

Among OFA, FA, and PS, our results do not pick out a clear winner.

PS's predictions for *agree with* in particular, are not borne out. PS predicted that in conditions 2b, 2c, and 2d, one of two possible *agree with* sentences would be true, but all *agree with* sentences came out false in these conditions, just like the symmetrical agreement sentences did (see section 3.2). If we put aside *agree with* sentences, on the other hand, PS is a clear winner, with predictions that match the data in all cases that produced a clear true or false result. Likewise, a hypothetical competitor that used PS's semantics of symmetrical as a semantics for *agree with* as well would be a perfect match for all conditions (again, except the ones we have been analyzing as cases of presupposition failure). Call such a hypothetical theory a Positive Alignment (PA) theory.

Although PA unambiguously outperforms OFA and FA, it does not appear to make the right predictions for the conditions in Table 7, if we are right in analyzing these in terms of presupposition failure. However, this may reasonably be regarded as an independent issue: these conditions seem to run afoul of an unnoticed (or in any case unstated) presupposition, but such a presupposition could plausibly be built into most accounts that made PA-type predictions, and as discussed below, it may well be the result of a general effect that is not specific to *agree wh* at all.

4.3 Consequences for *agree* and presupposition

The new analysis highlights the importance of presuppositions, as failure of a (possibly defeasible) presupposition is a natural explanation of the Table 7 results.

The results do not align with the presuppositions predicted by Egré and Spector (2007, 2014) (the only account to explicitly treat the presuppositions for this type of *agree wh* sentence): Egré and Spector derived presuppositions of complete opinionatedness, so their approach would have yielded presupposition failure for 1b, 1c, and 2a with all sentence types.

Instead, we seem to have what could be thought of as an existential import presupposition. Such a presupposition would not be surprising, as it seems closely connected with two more general kinds of proposed presuppositions. One natural way to analyze this presupposition would be

as an example of the more general theme of existential import associated with universal quantification. It seems natural (and consistent with Lahiri's approach), to implement a PS/PA account with universal quantification over possible answers (considered, for these purposes, as propositions of the form χ *is blue*, for some letter χ), giving the truth-conditions of an *agree wh* ascription by insisting that for every answer that either character believes, the other believes it as well. If as is alleged for many instances of universal quantification, the *every answer that either character believes* is understood as coming with a presupposition that the restrictor of the quantification is nonempty, then we will derive a presupposition that fails exactly in the cases given in Table 7.

Alternatively, a *wh* question is sometimes regarded as presupposing that at least something has the property in question. If this is right, then the question *Which letters are blue?* would presuppose the existence of blue letters. Now, what we find with *agree wh* embedding is not a presupposition that some of the letters actually *are* blue. but rather a presupposition that some of the letters are thought to be blue. Thus, it cannot be a matter of simply inheriting the nonemptiness presupposition from the question. But the predictions for *agree wh* would depend on one's exact semantics of questions, and on one's account of the results of embedding a presupposition-laden question, so it is not obvious that the existential presupposition of the question couldn't be implicated in the presupposition we have here. In any case, although we have no particular account of how to derive the presupposition in this way, the similarity between the presupposition we observe and the plausible existential presupposition for the question at least suggests another possible line of attack.

4.4 Consequences for embedded questions and reducibility

If reducibility (at least in any standard version) held for *agree*, we would expect (1-a) to have the same truth value in situations 1f, 2b and 2d,[11] and similarly we would expect the same truth value in situations 1b, 1d and 2c.

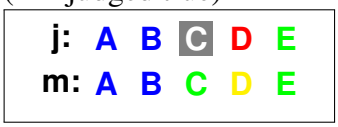
If we assume, in addition to reducibility, that the exact number of positive and negative cases involved is unimportant, then we should expect the same results for 1c, 1e and 2a as well. In all these cases, we see that the prediction of reducibility is not born out: 1f is judged true while 2b and 2d are judged false, 1b and 1d are judged true while 2c is judged false, and finally 1c and 1e are judged true while 2a is judged false.

Let us work through the case of 1f and 2b explicitly (the other cases are analogous). These conditions are nearly identical. In both, there are two letters that both John and Marry believe to be blue, and two that they believe are not blue. They differ only in that, in 2b there is one letter about which one character is unopinionated, but which the other believes is blue, while in 1f, there is again a letter about which one has no opinion, but here the other believes that that letter is not blue.

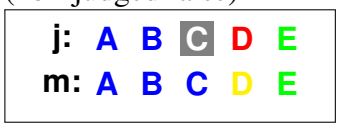
(16) John and Mary agree on which letters are blue.

[11] This is because John and Mary agree to the same possible answer propositions in all three of these, including propositions about letters being blue and not being blue, and all Boolean combinations of such propositions.

a. (1f - judged true)



b. (2b - judged false)



That is, in both cases, the propositions about which letters are blue to which John and Mary stand in the *agree that* relation are exhausted by the following (and those propositions that follow from them):

(17) A is blue.
B is blue.
D is not blue.
E is not blue.

That is, the *agree that* facts are identical for these two situations. In spite of this, the judgments for 1f fall into the TRUE cluster, while the judgments for 2b fall into the FALSE cluster. Thus, the *agree that* facts are not enough to determine the truth value of an *agree wh* attribution.

Thus, our data appear to provide a number of counterexamples to one straightforward notion of reducibility. But at this point, a cautionary note is in order — the notion of reducibility sketched above does not take presuppositions into account. For example it is true that in 1f and 2b the true *agree that* facts are the same, but it may be that some *agree that* attributions that are simply false in one suffer presupposition failure in the other. In particular, it is not implausible that in 4b, (18) is false, while in 1f it suffers presupposition failure.

(18) John and Mary agree that C is blue.

So we cannot rule out the possibility of a more involved notion of reducibility that take presupposition into account in some more involved way (indeed, Lahiri's account of *agree wh* does something like this), but the kind of reducibility that is commonly assumed will not do.

A second point is that the departure from reducibility required by our data may be a relatively modest one. George (2011), in rejecting reducibility for *know*, proposes the following weaker constraint:

(19) If R relates an individual a to a question π, then R relates a to at least one p such that p answers π.

This is, in essence, one direction of Higginbotham's biconditional, given in our initial discussion of reducibility. George's conjecture (19) is inconsistent with the predictions of PS and FA for situation 3c, since in that situation John and Mary do not agree to any standard kind of answer proposition (be it positive or negative, exhaustive or partial), but the PS and FA theories predict that the *agree wh* sentence should nevertheless be true. Our assessment has been that in this case there is presupposition failure - in fact, the observed presupposition failure denies truth to the the *agree wh* attribution in exactly those cases where PS and FA come into tension with George's

conjecture. So, although the two uses of *agree* are not related by a simple reduction relation, they are not wholly unconnected, and obey general connecting principles like the one given in (19).

5 Conclusion

We set out to explore the truth-conditions of *agree wh* sentences, with a particular eye towards presupposition and (non-)reducibility. Our data generally favor a PS account along the lines advocated by Lahiri (2002), except in the area of certain asymmetries involving the *agree with* construction as compared with symmetrical agreement, where we find no evidence for any difference, and in particular find our results at odds with Lahiri's specific truth-conditional claims. Our data further suggest a previously undiscussed presupposition for *agree wh*, and also support the view that there is in fact a non-reducibility effect for *agree*. Along the way, we have seen a novel application of k-means clustering to the differentiation of presupposition failure from truth and falsehood (see Križ and Chemla 2014 for discussion about methods to detect truth value gaps).

There are a number of problems for future investigation. One obvious area for further work is the investigation of a wider variety of *agree wh* sentences using a wider variety of judgment tasks, in order to determine how robust and generally applicable our observations are. It would also be valuable to experimentally explore other non-reducible descriptions in the literature, such as the claims about *know* and *forget* in George (2011), the proposal for *know* in Spector (2005), and the account of *tell* in Heim (1994). Finally, the prospect of non-reducibility suggests that we need a more complex account of what is common to the meaning of the two *agree*s, and, more generally, what principle links the meanings of the two uses of a question-embedding propositional predicate like *agree* or *know*. The treatment of *agree* in Lahiri (2002) suggests one possible approach, and the approach to non-reducibility in George (2011) offers another attempt, but, as the space of non-reducibility phenomena is still under-explored, a really compelling general theory has yet to emerge.

References

Beck, S. and H. Rullmann (1999). A flexible approach to exhaustivity in questions. *Natural Language Semantics 7*, 249–298.

Egré, P. and B. Spector (2007). Embedded questions revisited: *an* answer, not necessarily *the* answer. Presentation at MIT Ling-Lunch Seminar and Journées Sémantique et Modélisation.

Egré, P. and B. Spector (2014). A uniform semantics for embedded interrogatives: *an* answer, not necessarily *the* answer. Forthcoming.

George, B. R. (2011). *Question Embedding and the Semantics of Answers*. Ph. D. thesis, University of California, Los Angeles.

George, B. R. (2013). Knowing-'wh', mention-some readings, and non-reducibility. *Thought: A Journal of Philosophy 2*, 166–177.

Groenendijk, J. and M. Stokhof (1984). *Studies on the Semantics of Questions and the Pragmatics of Answers*. Ph. D. thesis, University of Amsterdam.

Heim, I. (1994). Interrogative semantics and Karttunen's semantics for *know*. In R. Buchalla and A. Mitwoch (Eds.), *IATL 1*, pp. 128–144.

Higginbotham, J. (1996). The semantics of questions. In *The Handbook of Contemporary Semantic Theory*, pp. 361–383. Blackwell.

Karttunen, L. (1977). Syntax and semantics of questions. *Linguistics and Philosophy 1*, 3–44.

Kratzer, A. (2006). Exhaustive questions. Linguistics Colloquium, Massachusetts Institute of Technology.

Križ, M. and E. Chemla (2014). Finding truth value gaps. Ms. University of Vienna & LSCP.

Lahiri, U. (2002). *Questions and Answers in Embedded Contexts*. Oxford University Press.

Ryle, G. (1945). Knowing how and knowing that. *Proceedings of the Aristotelian Society 46*, 1–16.

Schaffer, J. (2007). Knowing the answer. *Philosophy and Phenomenological Research 75*, 383–403.

Sharvit, Y. (2002). Emedded questions and 'de dicto' readings. *Natural Language Semantics 10*, 97–123.

Spector, B. (2005). Exhaustive interpretations: what to say and what not to say. presented at the LSA workshop on Context and Content.

Stanley, J. (2011). *Know How*. Oxford University Press.