Demand Characteristics as a Tool for Evaluating the Design of Collaborative Tasks

Ed Baggs
School of Informatics
University of Edinburgh
e.baggs@ed.ac.uk

Abstract

It is proposed that more attention should be paid to demand characteristics in collaborative tasks. The paper focuses on joint problem-solving tasks of the type typically used in dialogue research. The impact of demand characteristics in these tasks—specifically, the presence of discrepancies between how researchers believe a task to be and how it is perceived by subjects—is often difficult to evaluate from published write-ups, because attempts to identify such confounds are typically unsystematic. This need not be the case. Methods exist to evaluate the validity of our descriptions of a given task. In addition, tasks involving dialogue have a unique feature, namely the openness of the exchange between subjects, which allows us to directly observe what kinds of cues subjects make use of while completing the task. We can exploit this openness to evaluate and improve task methodology; this last point is illustrated with some examples from the HCRC map task corpus.

1 Introduction

It is a commonplace to observe that context plays an essential part in conversation, but this is misleading. The word *context* implies that the real business of an interaction is the language used, and that everything else is mere scaffolding. From the point of view of a given individual, however, this is simply not the case. An individual is only ever trying to accomplish a *task*; the language used during a task is at best a means, not an end in itself (cf., Cohen, Levesque, Nunes, & Oviatt, 1990).[1]

[1]Couldn't the task be simply to have a conversation? Perhaps, but even then the goal is not to produce a conversational record for its own sake, but to gain knowledge from other people, or to tell them a story, to pass the time, etc.

On this way of seeing things, conversational transcripts, and other records of the language used during the completion of an experimental task, are traces of what happened during the completion of the task, analogous to a series of footsteps left on a beach. In much of the empirical work carried out on dialogue and interaction, the implicit goal has been to derive general truths about language use from these kinds of linguistic traces taken from experimental data and speech corpora (Schober, 2006). The ultimate goal here seems to be to come up with a general theory of communication, so we'll call this way of doing things the general theory–directed approach. The present paper adopts an alternative, task-directed approach. Here, these linguistic traces are seen as a tool for understanding the *tasks* in which linguistic data originated. In particular, this is proposed as a method for evaluating the internal validity of tasks: is our description of a task consistent with how the task is really perceived by those carrying it out?

A task here is understood in a commonsense way as any (language-involving) goal-directed phenomenon we are interested in explaining; exactly what the nature of a given task is is subject to revision following empirical investigation. What's needed is that, for a given task, we have a good way of assessing what exactly is going on when people carry it out: what specific mechanisms are employed? This is necessary if we want to know how confident we should be about our description of the task of interest, and, ultimately, about the extent to which we are justified in making general conclusions from results specific to the task. Below I propose that the concept of demand characteristics can be adapted as one tool for addressing these issues.

Demand characteristics, on the definition given below, are something common to all psychological experiments as well as to many other situations

Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue, December 16-18, 2013, Amsterdam, The Netherlands.

where someone is following instructions. There are two reasons for narrowing the focus here to experiments on dialogue: 1) I believe the literature on dialogue could only benefit from more attention being paid to task demands and accompanying issues with validity, and 2) dialogue tasks produce data that is particularly useful for developing ideas about demand characteristics themselves, because the open exchange that occurs between the individuals carrying out the task can often allow researchers to reconstruct what was going on as the task was being carried out. Section 4 onwards will be concerned with the second point.

2 Demand characteristics

What are demand characteristics? The concept of demand characteristics is sometimes confused with the more specific 'good subject effect', the idea that subjects want to help the experimenter get useful results, and so behave in the way they think is expected of them. The concept is much deeper than this, however. Ultimately, it is about what tasks look like from the subject's point of view (Kihlstrom, 2002): demand characteristics are the properties of a task situation as perceived by the person carrying out the task.[2] Orne (1962), who introduced the term, wrote:

> 'The subject's performance in an experiment might almost be conceptualized as problem-solving behavior; that is, at some level he sees it as his task to ascertain the true purpose of the experiment and respond in a manner which will support the hypotheses being tested. Viewed in this light, the totality of cues which convey an experimental hypothesis to the subject become significant determinants of subjects' behavior. [...] These cues include the rumors or campus scuttlebutt about the research, the information conveyed during the original solicitation, the person of the experimenter, and the setting of the laboratory, as well as all explicit and implicit communications during the experiment proper. *A frequently overlooked, but nonetheless very significant source of cues for the subject lies in the experimental procedure itself, viewed in the light of the subject's previous knowledge and experience.* For example, if a test is given twice with some intervening treatment, even the dullest college student is aware that some change is expected, particularly if the test is in some obvious way related to the treatment.' [emphasis added]

One technique researchers have used to try to mitigate the confounding effect of subjects' expectations about an experiment is to deceive them as to the true purpose of the task. As Orne was aware, however, the efficacy of such deceptions is hard to assess from subjects' behaviour alone: a subject might appear to be behaving as the experimental manipulation predicts, but we do not necessarily know if this is a spontaneous response that reflects how the subject would behave outside of the laboratory, or if it is a more narrow response to some particular perceived cue in the set-up. And further, there exists a 'pact of ignorance' between subject and experimenter: subjects presumably have no wish for their data to be discarded from the analysis, and researchers do not wish to have to replace subjects, so it is in the interests of neither for the experimenter to probe too hard about what the subject was thinking during the task, lest the data should have to be rejected (Orne, 1969).

A note here on deception. It might be contended that this kind of deception is not relevant to tasks in the cognitive literature on language use, where everything is as it seems, and subjects are merely being asked to solve a problem set by the experimenter; in the map task, considered below, subjects are explicitly given roles as either the giver or follower of instructions, and are then simply instructed to carry out the task between themselves. We cannot assume, however, that things are so straightforward. Some of the most famous psychological experiments of the past sixty years or so—the ones our subjects are most likely to be aware of (such as the Milgram experiment)—*do* involve deception. Moreover, the undergraduate students that volunteer for the deceptive experiments are the same as those that volunteer for the non-deceptive ones. And so we must proceed on the assumption that any task that can be perceived as involving deception is likely to be so perceived. That is, even if we are not trying to deceive, we

[2] Following Kihlstrom, I'll continue to use the term 'subject' in preference to 'participant', as it is a more precise descriptor of the volunteer's role in the systematically designed tasks considered here.

still have to consider the possible presence of deception from the subject's point of view.

Whether a task has confounding demand characteristics or not is not simply an objective property of the task. It should be stated clearly that demand characteristics are specific to the subject, and can be located only in the interaction of the subject with the task as a whole. Demand characteristics overlap, in this sense, with James Gibson's concept of affordances (Gibson, 1979). It is tempting to suggest that demand characteristics are an instance of affordances specific to the laboratory, but this would be misleading. Affordances are *opportunities* for action, perceivable by an organism in the relation between external structure and its own ability to act upon that structure. Demand characteristics, by contrast, are contractual in quality: subjects in an experimental situation have committed themselves to carry out the task the experimenter has set for them; a response might be required even if no meaningful action is perceived (for example, a forced choice might have to be made between two stimulus items that appear the same). Different subjects will perceive a given task differently because they bring different things into the experiment: some will arrive with knowledge that's relevant to the task hypothesis: perhaps they have participated in a similar task before, or they might have had some other experience or training that makes them well-placed to detect the hypothesis. Researchers are generally aware of these problems, and try to avoid, for instance, testing the same subjects on similar tasks, or on different variants of the same task.

Despite this complication—that different cues are available to different subjects—we can still hope to identify properties within a task procedure and set-up that are likely to generate problematic demand characteristics. It may be useful to conceptualize the kinds of cues present in a given task as likely to tilt the resulting behaviour either towards or away from that predicted by the research hypothesis. I'll call these positive and negative demand characteristics, respectively (these labels are intended to be analogous to 'false positive' and 'false negative', rather than to imply good and bad). It then becomes possible to think of the (internal) validity of an experiment as a function of the cues present. This is represented schematically in Fig. 1. Note that if a task produces cues that consistently tilt behaviour one way or the other, then the task falls outside the shaded zone, and the task procedure should be considered insufficiently sensitive to detect the behaviour of interest. Note also that it is not enough for a task to fall within the shaded area for it to be considered *externally* valid—that is, a genuine result may still fail to generalize outside of the task, if the task is a poor model of the phenomenon of interest. Fig. 1 applies only to tasks that might appear to involve deception, or where the true research hypothesis is otherwise hidden from the subject; the situation may be different for non-deception tasks, such as, say, a test designed as a simple evaluation of a person's ability in some area (an IQ test is Orne's example); here, positive demand characteristics may merely serve to increase motivation.

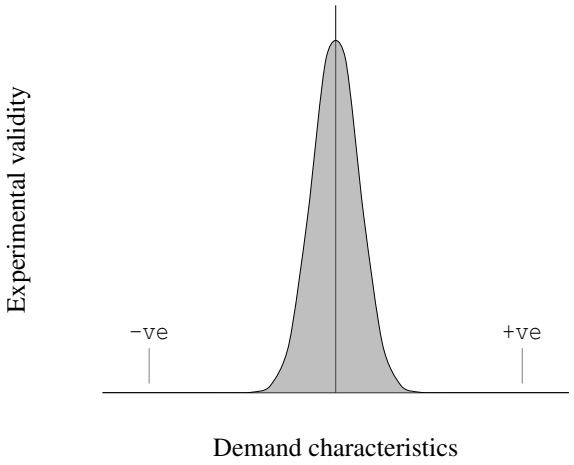

Figure 1: Schematic of the space of possible tasks (in which the research hypothesis is hidden from subjects), showing experimental validity as a function of demand characteristics; validity rapidly declines as demand characteristics push subjects' behaviour towards (positive demand characterstics) or away from (negative demand characteristics) the research hypothesis.

3 Dialogue tasks

There are a variety of ways in which researchers have attempted to study dialogue in the laboratory. I'll here consider one common class of tasks—referential communication games (Yule, 1997)—in which two subjects are recruited to complete a problem-solving task together (I'll ignore versions that use confederates). Routinely, these tasks involve constraints placed on the pairs over how they are allowed to solve the task. Often each member of the pair is given separate materials that they

have private access to and the task is for one member to communicate something about the structure of their materials to the other, using only linguistic expressions.

To reiterate the point at the beginning of this paper: the approach being advocated here is concerned with explaining specific mechanisms involved in the completion of particular tasks. To be clear, by mechanisms here I do not mean internal algorithmic-level descriptions of steps involved in carrying out a task. Instead, I propose to understand a task environment, which includes oneself and other people, as providing a set of possible resources that can be assembled in pursuit of a goal (Wilson & Golonka, 2013). A mechanism, then, is a way of assembling those resources.

That being the case, why should we be interested in these referential communication games? These tasks are not interesting *per se*; they exist because they were devised to advance some general theory about how communication works, not because the researchers who devised the tasks had some inherent interest in this kind of game (for example, early versions of these games explicitly instantiated an information theoretic code model of language as a signal transmitted between an encoder and a decoder; the tasks were employed as a means of disrupting feedback; see Krauss & Weinheimer, 1966). The answer is that we don't currently have a well-developed way of going about the study of collaborative activity that primarily seeks to explain tasks; we do, however, have corpora from existing tasks, such as the map task, below, that can be used as immediate material for developing such an approach. So the following is a preliminary attempt to develop the tools of a task-directed approach, drawing on an existing corpus of data.

4 Demand characteristics in the map task

The HCRC map task (Anderson et al., 1991) is an interesting case in terms of demand characteristics because it was set up not to test a single hypothesis, but to test several hypotheses at once, and to produce a corpus of data that could be used to investigate an open-ended set of research questions. Meanwhile, the concept of demand characteristics, as defined, is only meaningful relative to a single, specific research hypothesis. One might think, then, that the concept would be hard to apply here. Nonetheless, it's easy to identify cues that people are aware of while carrying out the task, and we can talk about these cues in general terms; we can do this by examining the recordings and transcripts from the corpus (available at http://groups.inf.ed.ac.uk/maptask/). Note that the following is not meant to be a discourse analytic discussion of the task. Looking for demand characteristics should be seen as part of the experimental design and evaluation process; it is a way of asking whether our description of the task matches the reality from the subject's point of view. The discussion of the map task here is meant to demonstrate that this can in principle be achieved, in part, by examining the open exchange that goes on as people carry out the task.

In this task, an instruction giver sits in front of a map with a predefined route drawn on; the goal is to communicate this route to an instruction follower who can't see the instructor's map, and for the follower to reproduce that route on their own map. Subjects were told this goal explicitly: 'Subjects were told that the goal of the task was to enable the Giver's route to be drawn on the Follower's map, that the Giver's and Follower's maps might be different in some respects, and that both participants could say whatever was necessary to complete the task, but that neither could use gestures.'

Examples of the maps can be seen in Figures 2 and 3. The instructor had the map on the left; the follower's completed map is shown on the right. I'll here look at three exchanges that illustrate some effects of demand characteristics in this task.

The first exchange (from a pair coded as q1nc2 in the corpus) I present as evidence that the constraints on communication described in the instructions given to subjects are only partly true as a description of what actually happened in the task. Specifically, the rule that 'neither could use gestures' can only have been partly followed (g is the instruction giver, f the follower; I have added the comment and punctuation):

> g — and you should be kind of ehm two and a half inches away from the right-hand side of the page just now
> f — oh [uhh...] no
> g — no
> g — where are you?
> f — my inches must be different from

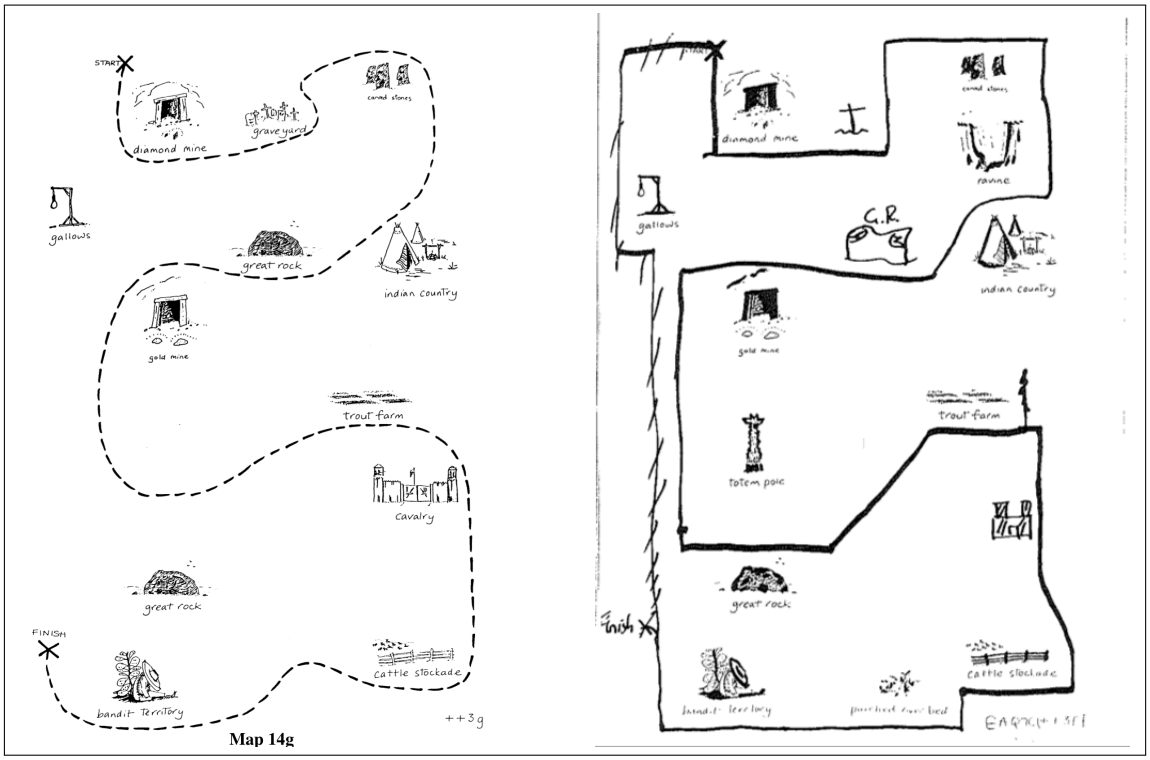

Figure 2: Map task conversation q7ec1—the route giver's map is on the left, the completed follower's map on the right; in this trial, the pair could make eye-contact; they were both male, and knew each other beforehand; recording duration 5'58" (map images are copyright Human Communications Research Centre 2007, and are available under a creative commons licence, cc-by-nc-sa)

yours 'cause I'm not even halfway across the page
f — I should be away at the other s– side of the page?
g — you should be kind of at the right-hand side
f — how l– how big's your page?
g — er
f — is it that size? [f shows the back of her map to g]
g — uh-huh
f — uh-huh

This exchange in fact comes from a no-eye-contact trial, in which there was a barrier between the pair. The follower can be heard on the recording wielding the page. What's not seen in the transcript is that the instructor breathes in, perhaps apprehensive about what has just happened, as if she is worried that they have just broken the rules and so will have to be ejected from the experiment. Of course, by normal standards, this is a perfectly sensible thing to do: showing something to someone to confirm that you're both talking about the same thing. (Even more sensible would be for the instructor to pass her map over the barrier for the follower to copy out the route directly. None of the participants did this, of course; they would have been ejected.) Here, then, is one instance of gesturing that found its way into the corpus. Video recordings of the sessions (not available online) no doubt contain countless other instances, particularly if we consider facial expressions as gestures.

The lesson here is perhaps that if you want your subjects to behave towards one another in a specific way, it is not reasonable to place the burden of maintaining that behaviour on the subjects themselves. The subjects did not have visual access to each other's maps. This was more or less guaranteed by the layout of the furniture in the laboratory. They did, however, have continual access to each other's gestures, and to their own ability to produce gestures. Given how ubiquitous gesturing is in life outside the laboratory, it would seem to require considerable effort to deliberately suppress this behaviour.

The second exchange is from the pair whose maps are shown in Fig. 2. This exchange con-

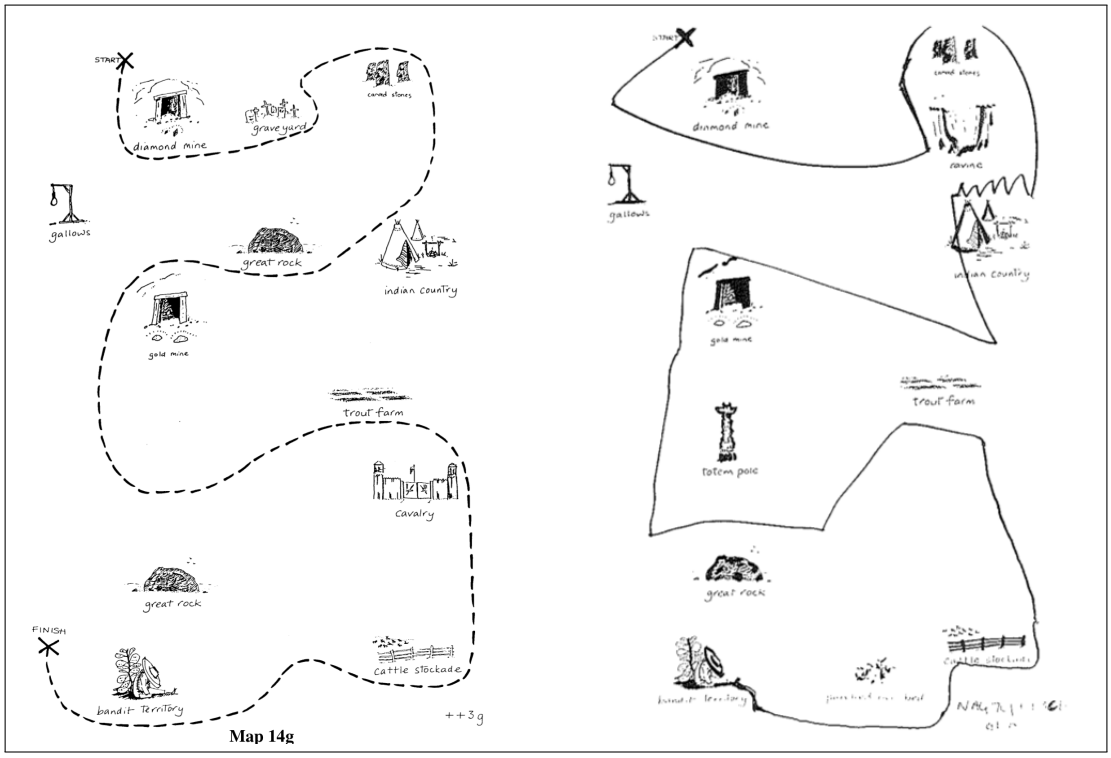

Figure 3: conversation q7nc1—no eye-contact, both female speakers, knew each other beforehand, recording duration 6'44"

tained a false finish, hence the crossed-out line on the left. The completed follower's map also features some extra landmarks, which the instructor insisted be drawn in (the initial maps differed in the placement of some landmarks). These both suggest that the pair were motivated to perform the task well.

The recording of this pair also reveals another aspect of dialogue tasks which is absent from non-dialogue tasks. It's clear from the recording of this exchange (though again, not the transcript) that the instructor is trying to make the follower laugh as they complete the task. He repeatedly instructs the follower to 'hang a left', instead of the more mundane 'turn left', and does so with audible delight. At the point above the Indian country on the right hand side:

> g — until you get to the indian country
> then you do a wee chicane
> g — turn left above the indian country

Between these two utterances the follower can be heard chuckling. It seems fair to say that instructor is willing to sacrifice some precision here in favour of making the task more enjoyable. Here is a demand characteristic peculiar to tasks that allow interaction: a joint task is also a social activity between subjects. Whether this is something to be concerned about will depend on the research question we are interested in answering.

Finally, look at Fig. 3. This is the same map as in Fig. 2, completed by a different pair. There is a salient feature on the instructor's map towards the top, where the route makes an 'S' curve around the graveyard. The instructor in Fig. 3 draws this to the follower's attention and tells the follower to go 'back towards the right' (this pair started at the finish point, hence 'right' and not 'left'). This bend can be seen on the follower's map in Fig. 3. However, none of the other completed versions of this map (each map was completed by eight different pairs) features this curve. The goal of the task as interpreted by the pairs seems to have been to avoid hitting the landmarks. It is worth emphasizing this because it conflicts with the assumption that the goal defined by the instructions—'to enable the Giver's route to be drawn on the Follower's map'—is well defined. Anderson et al. assume it is, and that this allows for an objective measure of communicative success: 'Because the correct solution to the problem is well defined, *successful communication can be measured* in terms of the extent to which the achieved route

corresponds to the model.' [emphasis in original] If people are partly using a landmark-oriented strategy, then the standard measure of success (absolute deviation from the path) is strictly measuring a different thing from what subjects are alert to: it measures whether the path is in the right place in absolute terms, not whether it is in the right place relative to the landmarks.

In summary, these exchanges provide evidence for three properties of the task not acknowledged in the original description in Anderson et al. (1991): 1) The instruction that subjects cannot use gestures creates an artificial burden on subjects to monitor their own behaviour. 2) The task has properties not present in individual problem-solving tasks: participants here are sometimes attempting to amuse each other; this may introduce a discrepancy between the overall goal of the task as the subjects see it and the task as the researchers assume it to be. And 3) the route, as interpreted by most pairs in the map task, is landmark-oriented, and not absolute, as assumed by the researchers. This partially undermines the claim that the task has an objective measure of success. In general, we might want to consider that objective measures of communicative success are a fiction; communicative success can only be defined relative to the goal from the point of view of whoever is trying to accomplish it. Any research question that hinges on communicative success should be alert to such discrepancies between the thing measured and the tool used to measure it. Indeed, anyone using task corpus data to investigate a specific research question should try to evaluate the demand characteristics of the task relative to that question. These three observations can all be used to make better sense of the behaviour in this particular task.

To repeat, the purpose of this discussion is to demonstrate how we can take advantage of the open exchange of dialogue to evaluate the suitability of an experimental methodology for addressing a given research question, and to improve that methodology in subsequent versions of the task. This evaluation can be done in a rigorous way: produce a description of the task goal, then look for counter-evidence that that's what the goal is from the subject's point of view; describe your dependent measures, then look for counter-evidence that these are measuring what you think they are measuring; and so forth. To be sure, this is not guaranteed to detect every possible confound, but it can surely detect some.

5 Detecting demand characteristics

What we are interested in here, is detecting demand characteristics in situations where the cues are not well understood and where unknown confounding cues may be present. Orne (1969) described three main methods for doing this. He called these methods 'quasi-controls'. All of his methods seek to recruit the subject as a co-investigator. Orne was interested in hypnosis; he developed the concept of demand characteristics in order to ask questions such as this: are hypnotized subjects really under the control of the hypnotist, or might they merely be behaving in the manner they think they're expected to, because of the peculiarities of the situation? The techniques may be partly applicable to dialogue research too.

The first method is simple post-test inquiry: ask the subject what they thought they were doing. Such inquiries are presumably widely conducted nowadays, but are less commonly reported. It is not clear why this should be the case. These questionnaires are in part suspect, of course, because of the pact of ignorance mentioned above: research participants do not wish to be ejected from the analysis, and so, if they did in fact suspect some deception, they have an incentive to keep this to themselves. But this would still yield a set of responses consistent with the deception being valid, and even this kind of thing is not widely reported. One reason why researchers may omit the questionnaire data from the write-up is that it's seen as too difficult to summarize. If this is the case, though, then this too should be reported: if subjects do not in fact have a common idea of what it is they are doing, this may undermine an unstated assumption of the researchers, who presumably intend the task to be perceived in a uniform way. More diligent reporting of the kinds of things people say after a task should be encouraged.

Orne's second quasi-control method he called the 'non-experiment'. Here, subjects are shown the materials and the set-up, but not actually asked to carry out the task. Instead, they are asked to guess how others would respond if given these materials and asked to complete the task. This method may be of potential use in dialogue research. A possible shortcoming is that dialogues are unpredictable from the standpoint of any one participant, and perhaps even to a pair of non-

participants: each member of a pair has only a partial perspective on what the task is. A pair might only be able to work out how they would perform a task by actually doing the task. Similarly, Orne noted that the non-experiment cannot be sensitive to cues that subjects themselves are not consciously aware of. Still, the method could allow researchers to see what kinds of approaches people are inclined to take going into a task.

The third method—simulation—is perhaps more specific to the kinds of question Orne was interested in. Here, 'simulating' subjects are recruited and asked to behave as if they are real subjects, that is, they're told they're in an experiment involving hypnosis, and are asked to behave as if they are actually hypnotized; there is an experimenter who is blind to who are the simulators and who are the 'reals'; the simulators' task is to make the experimenter believe they are genuinely hypnotized. It's harder to imagine where this simulation method could be applied in dialogue research.

The discussion in this paper suggests a fourth method. Dialogues have an inherent feature that allows the researcher to look in and infer directly what kinds of demand characteristics people are sensitive to: dialogues are open, in the sense that they consist of behaviours that can be observed from the outside, rather than solely of internal mental behaviour that has to be inferred by proxy. The openness of a dialogue means that subjects can be used as their own quasi-controls. The brief discussion of the map task above is intended to demonstrate the plausibility of this method. Granted, this method involves some uncertainty; it depends on inference on the researcher's part: the researcher is looking for counter-evidence that the task is perceived by the subject in the manner intended. But the method is valuable if it allows us to detect at least some potential confounds that we would otherwise be ignorant of.

It may be useful here to say something about how, specifically, one should go about attempting to detect demand characteristics for a given set of data. First, it must be reiterated that demand characteristics are not a property of the task, but of how the task is perceived by an individual subject, relative to a research hypothesis. Specifying the hypothesis is a prerequisite before you can look for potential confounds. In general it is not possible to be very precise about exactly what to look for: this will depend on the nature of the hypothesis under consideration. But we can say, in terms of the schematic depicted in Fig. 1., that in order for a study to be valid, the task should produce demands that fall in the neutral space in the middle. That is, there should ideally be nothing about the task set-up itself that is misleadingly pulling behaviour either towards, or away from, the behaviour predicted by the research hypothesis. Non-neutral demand characteristics are a threat to validity: they cast doubt on our ability to attribute behaviour to something about the psychology of the individual subject; and raise the possibility that that behaviour should in fact be attributed to the task set-up. A write-up of the study should then seek to provide the following:

1. a clear statement about what the researchers believe constitute neutral conditions for the task under investigation

2. details of attempts to establish that neutral conditions did in fact prevail for the subjects engaged in the task, and

3. details of potential confounds which the researchers were unable to rule out from the available data.

These steps should be seen as a valuable part of the experimental design and evaluation process.

Finally, it must be admitted that these proposals are not especially novel. Some published studies on dialogue do make use of some of these methods. In particular, I'll note that in Schober and Clark (1989)—a study of how well over-hearers to a referential communication game are able to make sense of a discussion they're not part of—the authors include substantial discussion, under the heading 'Subjective commentary', of both questionnaire data (for experiment two), as well as inferences drawn from analysis of the recordings (for experiment one); that is, they made use of both inquiry and openness to evaluate the experimental design. For someone reading this paper with an eye to how the task looked form the subject's point of view, these discussions are extremely useful.

6 Implications for future work

At the beginning of this paper I made a distinction between general theory–oriented and task-oriented approaches to the study of language use. The discussion about demand characteristics in the

map task has arguably been consistent with either approach. I believe, however, that it is worth trying to pursue an alternative task-directed methodology that makes a strong claim to distinguish itself from the general theory–directed programme. The strong claim is this: the practice of producing language corpora from tasks as a method of studying 'dialogue' is misguided; coprus data can *only* be used as a means of evaluating the task. The reasoning here is as follows. If we want to draw general conclusion from observing specific tasks, then we need to be confident both that our description of the task is correct, and that the task itself is representative of the phenomenon we wish to model. In the case of dialogue, and the tasks used to model it, neither of these is necessarily true. Indeed, it's not clear what the scope of 'dialogue' is at all. It is clear, however, that we cannot judge *what* a task is representative of until we have a good understanding of the task itself; we have to know where to position the task on Fig. 1. One way of doing this is by appeal to quasi-control techniques for discovering demand characteristics.

A possible implication here is that the goal of a psychology of language use should not be to produce a general theory of communication at all; the goal should instead be to identify the mechanisms involved in the completion of specific tasks. This might appear a pessimistic conclusion. But it can perhaps be argued that a more modest scope has the potential to produce more tractable research questions than those commonly asked at present, and may be the only way to carry out a genuinely incremental psychology of language use.

7 Acknowledgments

This work is supported by an EPSRC grant, and by funding from the European Commission through the JAMES project (FP7-ICT-270435); I thank Holly Branigan and Jon Oberlander for their comments and discussion, and the reviewers for their useful remarks.

References

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., et al. (1991). The HCRC map task corpus. *Language and speech*, *34*(4), 351–366.

Cohen, P. R., Levesque, H. J., Nunes, J. H., & Oviatt, S. L. (1990, November). Task-oriented dialogue as a consequence of joint activity. In *Proceedings of the Pacific Rim International Conference on Artificial Intelligence.* Nagoya, Japan.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton-Mifflin.

Kihlstrom, J. (2002). Demand characteristics in the laboratory and the clinic: Conversations and collaborations with subjects and patients. *Prevention & Treatment*, *5*(1).

Krauss, R. M., & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, *4*(3), 343–346.

Orne, M. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*(11), 776–783.

Orne, M. (1969). Demand characteristics and the concept of quasi-controls. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 143–179). Academic Press.

Schober, M. (2006). Dialogue and interaction. *Encyclopedia of language and linguistics*, *2*, 564–571.

Schober, M., & Clark, H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*(2), 211–232.

Wilson, A. D., & Golonka, S. (2013). Embodied cognition is not what you think it is. *Frontiers in psychology*, *4*(58).

Yule, G. (1997). *Referential communication tasks*. Mahwah, NJ: Lawrence Erlbaum Associates.