Informationally Redundant Utterances Alter Prior Beliefs about Event Typicality

Abstract Most theories of pragmatics and language processing predict that speakers avoid excessive informational redundancy. Informationally redundant utterances are, however, quite common in natural dialog. From a comprehension standpoint, it remains unclear how comprehenders interpret these utterances, and whether they make attempts to reconcile the 'dips' in informational utility with expectations of 'appropriate' or 'rational' speaker informativity. We show that informationally redundant (overinformative) utterances can trigger pragmatic inferences that increase utterance utility in line with comprehender expectations. In a series of 3 studies, we look at utterances which refer to stereotyped event sequences describing common activities (*scripts*). When comprehenders encounter utterances describing events that can be easily inferred from prior context, they interpret them as signifying that the event conveys new, unstated information (i.e. an event otherwise assumed to be habitual, such as *paying the cashier* when *shopping*, is reinterpreted as non-habitual). Further, we show that the degree to which such inferences are triggered depends on the framing of the utterance. In the absence of prosodic or discourse markers indicating the speaker's specific intent to communicate the given information, such inferences are far less likely to arise. Overall, the results demonstrate that excessive conceptual redundancy leads to comprehenders revising the conversational common ground, in an effort to accommodate unexpected dips in informational utility.

Keywords: experimental pragmatics; redundancy; psycholinguistics; context-dependent implicatures; accommodation; overinformativeness; prosody.

Contents

List of Tables

List of Figures

1 Introduction

Pragmatic theories and theories of language processing typically include constraints against elements which add no new information to the discourse, or are linguistically or informationally redundant (cf. Aylett & Turk 2004, for a theory of phonetic reduction; Cohen 1978, for a computational theory of speech act generation; Grice 1975, for a theory of rational communication; or Jaeger 2010, for a theory of reduction at all levels of linguistic representation, among many others). At the form level, redundancy may include overt mention of, or increased articulatory effort towards producing material that is easily predictable or recoverable in context. In other words, more signal is provided than the comprehender requires to accurately recover the intended phonological, lexical, or syntactic form. Examples of redundancy avoidance at this level include vowel shortening (Aylett & Turk 2004), use of shorter word variants (Mahowald et al. 2013), or omission of optional complementizers (Jaeger 2010).

At the informational or conceptual level, redundancy refers to the explicit mention of information that the comprehender is already in a position to infer automatically, using world knowledge or common ground information, or that is already entailed or strongly implied by the preceding discourse. In other words, more information is provided than needed to recover the intended meaning or world state. In contrast to redundancy avoidance at the form level, constraints against *overinformativeness*, or redundancy, at the informational level have always been somewhat debated (Grice 1975).

There is ample evidence that speakers are routinely overinformative at the informational level, and that speaker overinformativity is frequently tolerated by listeners (Baker et al. 2008, Engelhardt et al. 2006, Nadig & Sedivy 2002, Walker 1993). In this paper, we explore the question of whether there is empirical evidence for constraints against this variety of speaker redundancy, when they might come into play, and how comprehenders may react to and interpret any violations of such constraints (or at least, of their expectations that speakers will be concise). Specifically, we look at cases where the redundancy is at the level of background world knowledge – as opposed to, for example, repeating something that has already been stated, or referring to an object in more descriptive detail than strictly necessary given the physical context.

Utterances such as (1) are at face value redundant, in that they overtly state that "John" *paid the cashier*, which conventionally can be inferred simply on the basis of him having gone *shopping*.

(1) John went grocery shopping. He paid the cashier!

Once it has been established that *John* went grocery shopping, comprehenders' expectations of a world state where the *paying* action has occurred are very high (Bower et al. 1979; cf. Zwaan et al. 1995). A theoretic account of utterance choice which places a constraint on informational redundancy would predict that uttering the second sentence in this context would be marked, at best. Further, it would predict that comprehenders should note this markedness, and possibly penalize it, or at least regard the speaker as somewhat odd. However, in this paper we also show that comprehenders, through pragmatic reasoning about the common ground, can accommodate these utterances by changing their previous beliefs about the likely world state.

1.1 World knowledge

Utterances like the one shown in (1) are redundant on the basis of background world knowledge. As background knowledge is fairly unsystematic and comprehender-specific, and can be difficult to control for, here we use *script*, or *schema* knowledge as a proxy for world knowledge. *Script* knowledge refers to people's implicit awareness of the typical event structures of various stereotyped activities, such as *going shopping* or *going to a restaurant* (Fillmore 2006, Minsky 1975, Schank & Abelson 1977). The former, for example, normally involves events such as *going to a store*, *selecting food items*, and *paying the cashier*. Comprehenders anticipate upcoming events once a script is "invoked" (Zwaan et al. 1995); and when recalling stories based on scripts, have difficulty distinguishing actions that were actually mentioned, and those that were unmentioned but implied by the script (Bower et al. 1979). These findings suggest that events which are strongly associated with a script are almost part of its conventional meaning, and that explicitly mentioning their occurrence is therefore redundant[1].

Utterance (1) introduces a well-known script or event sequence (*grocery shopping*), followed by an informationally redundant event description (*he paid the cashier!*), which references a highly predictable sub-event from the script. In this example, the event described in the second sentence is already strongly implied to have occurred by the preceding invocation of the *grocery shopping* script – given the assumption, shared by most speakers and comprehenders, that people overwhelmingly pay cashiers when they go grocery shopping. Mentioning it explicitly, therefore, is redundant.

[1] Highly inferable events are occasionally used as temporal anchors (*After she entered the restaurant, she*...), and may be used to transition back from interruptions to the script (*She stopped to talk to Brad on the street. She then entered the restaurant*...). However, outside of these contexts, easily inferable script events are usually not mentioned overtly (Bower et al. 1979, Regneri et al. 2010).

1.2 Informational redundancy

While most pragmatic theories do address cases where a speaker may be informationally redundant (Grice 1975, Horn 1984, Levinson 2000 , among many others), they often leave open the question of whether comprehenders do, in fact, perceive (unjustified) redundancy as infelicitous, as well as how they interpret redundant utterances. Most accounts do argue that comprehenders expect speakers to behave rationally – namely, by communicating in a way that is consistent with getting across the intended message (which, furthermore, should be truthful). However, as Grice (1975) notes, it's unclear whether excessive redundancy comes into any real conflict with the goal of successful (truthful, sufficiently informative, relevant, etc.) communication – although comprehenders may wonder what the "point" of excessive information is, and attempt to rationalize unexpected "dips" in informational utility by infusing them with additional pragmatic meaning. Informationally redundant utterances do not clearly interfere with comprehension, as *underinformativeness* or underspecification does, and may aid comprehension in some cases (e.g., object identification; cf. Nadig & Sedivy 2002, Rubio-Fernández 2016)[2]. In this light, it is not straightforwardly clear whether overinformativeness constitutes non-rational speaker behavior, and specifically to what degree this part of the *Quantity* maxim holds: "do not make your contribution more informative than is required."[3]

It is, however, possible that comprehenders perceive excessive information as, at minimum, non-relevant to the discourse (Grice 1975, Horn 1984). The question, then, is whether comprehenders make any particular note of redundancy, simply find it odd or infelicitous, or attempt to accommodate it. If comprehenders do perceive redundant information as irrelevant, then rational speakers should avoid overtly stating conceptually redundant information, except in those cases where this information is intended to communicate a more informative non-literal meaning (or signal an unusual world state). Correspondingly, comprehenders where possible ought to interpret conceptually redundant utterances as either an attempt to convey some non-literal (relevant and informative) meaning, or as reflecting a background world state where the information conveyed can't be taken for granted, and is therefore informative. How comprehenders do in fact react to redundancy has to date only been empirically investigated within the relatively narrow scope of nominal modification, where the evidence, discussed further in Section 2, has largely been equivocal.

[2] This is not to say that comprehension is not in any way impaired by redundancy, and in fact we suspect that it is - but at face value, there is nothing about receiving more information than needed that necessarily hinders one from arriving at the intended meaning of a message.

[3] Grice (1975) explicitly suspected that it did not; later accounts separated the notions of semantic vs. form-based informativeness, which is discussed in the following section.

Ultimately, the question of how comprehenders treat overinformativeness is relevant to a more general theory of human communication, and should be answered to determine the extent to which: a) comprehenders and/or speakers consistently behave in a “Gricean” manner; b) under which conditions they do so, and which deviations from communicative norms are more likely to occur/be tolerated; and c) to what extent comprehenders attempt to resolve apparent violations, and which strategies they use to do so. If comprehenders do not appear to make much of overinformativeness (whether in terms of inferences made, or maxim violations perceived), and there is little evidence that speakers deliberately use overinformative utterances to convey specific non-literal meanings, then it’s questionable to what degree overinformativeness violates communicative norms, in the first place.

1.3 How might comprehenders react to informational redundancy?

In this section, we speculate in more detail how comprehenders might react to specific instances of informational redundancy, or *overinformativeness*. We distinguish three theoretical possibilities. Specifically, we will consider what might happen when a comprehender encounters one of our experimental utterances (normally embedded within a larger context):

(2) John just came back from the grocery store. **He paid the cashier.**

1.3.1 Hypothesis 1: No inference

The first possibility is that comprehenders do not find informational redundancy particularly marked, as it does not necessarily interfere with interpreting the intended message – or, at most, find redundant utterances slightly odd or suboptimal, as has been found in some studies (Davies & Katsos 2010). It’s both likely that comprehenders do not expect speakers to exhibit entirely rational communicative behavior at all times, and that conversational maxims, if they have little or no effect on the comprehender’s ability to understand the basic meaning of an utterance, are followed only insofar as the speaker feels like it. Similarly, comprehenders may note redundancy in speech, but not draw any inferences regarding some intended meaning or background world state. They might instead ascribe the redundancy to some kind of speaker error: perhaps the speaker is stalling for something else to say, having production difficulty, or is simply not communicating very well in that particular instance. In the case of our utterance (2), in this scenario, we might expect that comprehenders would interpret the utterance literally, and make no more of it than stated; i.e., they would take away the message that on some particular instance, John paid the cashier, and perhaps the speaker described it in a bit more detail than strictly necessary.

1.3.2 Hypothesis 2: Non-detachability

If comprehenders do expect speaker utterances to always have a certain level of informational utility, then they may attempt to resolve the provision of excessive or unnecessary information by drawing pragmatic inferences, regarding what they think the utterance means or signifies from the speaker's perspective. These pragmatic inferences would then serve to increase the informational utility of the utterance, and allow comprehenders to maintain the belief that the speaker is being cooperative – since assigning an "informative" pragmatic meaning to an apparently redundant utterance in effect removes the redundancy. In the case of utterance (2), comprehenders might conclude that John's *cashier-paying* is being announced due to its being unusual or unexpected, and that John can't therefore typically be counted on to pay the cashier. This reaction should occur as long as the background and linguistic context is basically consistent with that interpretation, and, as in the case of most pragmatic inferences, should be unaffected by changes to the utterance which do not alter its semantic content (generally referred to as *non-detachability*; Grice 1975), such as prosodic and/or discourse markers which do not change the truth conditions of the sentence - i.e., the inference should be attached to the semantic content, not the specific linguistic form of the utterance.

1.3.3 Hypothesis 3: Form sensitivity

The third possibility is that, as in Hypothesis 2, comprehenders react to a statement of *John's* having paid the cashier by inferring that *John* must be a habitual non-payer. However, as the inferences we are concerned with are based, in a sense, on the specific *form* of the utterance (i.e., too much signal is used to communicate something that would have already been understood), it is possible that such inferences may be relatively sensitive to how exactly the utterance is expressed[4]. In particular, we suspect that expending extra articulatory effort on expressing our already redundant utterance would increase the strength of any pragmatic inferences drawn (or even cause inferences to be drawn where none would be otherwise). Fundamentally, a greater show of *intentionality*, in apparently violating a maxim, provides more evidence that the maxim is not simply being violated due to a speech error or difficulty in utterance planning, and signals to the comprehender that more should

[4] The principle of *non-detachability* generally does not hold for Manner implicatures (Grice 1975), which are conceptually similar to the inferences we look at here; for space reasons, we will not go into more detail on this type of implicature, however.

be made out of the apparent violation. This also echoes Wilson & Sperber (2004)'s stated basis for their Communicative Principle of Relevance: "by producing an ostensive stimulus, the communicator therefore encourages her audience to presume that it is relevant enough to be worth processing" (an *ostensive stimulus* being one that is "designed to attract an audience's attention and focus it on the communicator's meaning"). In the case of our utterance, what we would predict in this case is that the more obvious effort is expended on producing the utterance (whether in the form of prosodic stress, or another attention-drawing signal of relevance and intentionality), the stronger the inference. To note, some amount of form sensitivity is not necessarily incompatible with the second hypothesis (*non-detachability*), but the complete absence of an inference would be.

In this paper, we present three experiments, run concurrently on the same population, which test whether informationally redundant event descriptions lead comprehenders to alter initial beliefs about how predictable, or habitual, the event in question is, on the premise that less habitual events are more likely to be mentioned. We predict, consistent with the first and third scenario outlined above, that informationally redundant event descriptions should generate what we term *non-habituality inferences*, where comprehenders resolve the apparent dip in informational utility by concluding that the usually predictable and habitual event described is in fact relatively *non-habitual*/non-predictable, as this would justify its mention. For example, in our scenario involving *shopping* and *paying the cashier*, a possible inference would be that "*John*" does not pay habitually (e.g., has someone else pay for him, is a habitual shoplifter, or gets free groceries). We first look at these utterances in discourse contexts which implicitly support a *non-habituality* inference, through semantically vacuous prosodic or discourse markers which serve to draw the listener's attention to the information being conveyed[5]. The first experiment uses implicit exclamatory prosody (the marker "*!*") to signal that the utterance is an intentionally conveyed, important, and relevant piece of information. The second experiment uses the discourse marker "*oh yeah, and...*" to do the same, while avoiding the surprise conventionally implied by the exclamation mark. In the third experiment, we predict that informational redundancy by itself, in absence of prosodic or discourse cues as to relevance and intentionality, triggers weaker *non-habituality* inferences, consistent with the third hypothesis (*form sensitivity*).

In the next section we review relevant literature, and in the following sections we describe the three experiments we ran.

[5] When communicating, speakers frequently use multiple cues to make an inference maximally easy to compute for the comprehender (consider, for example, the difficulty of interpreting sarcasm without supportive prosody). We therefore consider the presence of some amount of prosodic or discourse support to be the "default" case.

2 Related Work

Our work builds on two primary strains of research: interpretation and perception of informational redundancy on the one hand, and relatively new work on inferences about background world states (vs. speaker intentions) on the other. We also look at the effects of implicit prosody on pragmatic interpretation, which to date has largely been limited to semantic effects of contrastive prosody. Further, we overall look at the (systematic) interpretation of *particularized*, or *ad-hoc* pragmatic inferences, which arise only in specific contexts, and/or on the basis of reasoning about world knowledge. These have not received a lot of attention from pragmaticists, experimentally or otherwise, partially due to their idiosyncratic nature, making them difficult to study systematically; and partially due to being seen as less relevant to a theory of pragmatic vs. semantic meaning than, for example, scalar implicatures (Levinson 2000).

2.1 The problem of *overinformativeness*

First, we want to discuss a problem of terminology. In most experimental work, informational redundancy has been described as a problem of *overinformativeness*, *overspecification* or *overdescription*, and as addressed by the second part of Grice's Quantity Maxim, which states that speakers should provide no more information than is necessary to get their message across. However, *overinformativeness* in the pragmatic literature has been used to refer to both informational redundancy (Engelhardt et al. 2006, Grice 1975), as well as to the relative informativeness of terms in an implicational scale (e.g., the use of *some* when *all* is sufficient) (Horn 1984, Levinson 2000). The latter variety of *overinformativeness*, now more typically associated with the Quantity Maxim, is more a problem of unjustified vagueness where a more *precise* description is available. Informational redundancy, in contrast, is a problem of *excessive* wordiness or precision, as in the case of overinformative nominal modification (such as using *the big red cup* or *the cup on the towel* to identify the only available cup in a given context), where speakers might choose to describe objects in more detail than is strictly necessary. In this paper we concern ourselves strictly with overinformativeness in the sense of informational redundancy, as originally described by Grice (1975), and in the literature on nominal overspecification.

While informationally redundant utterances are typically regarded as infelicitous in the linguistics literature, they have been observed to be surprisingly common in natural dialog. Baker et al. (2008) observed that such utterances are frequently used in response to signs of listener non-comprehension, when responding to listener questions, or when speaking to strangers. Walker (1993) also concludes that informationally redundant utterances are specifically used to address cognitive resource limitations (e.g., memory for preceding discourse, limited inference-making capacity), as well as to serve a narrative function. In the latter case, this may for example involve drawing attention to a particularly salient or relevant fact. In other words, many or most informationally redundant utterances are not in fact redundant, as the apparent redundancy has communicative purpose. Literature on nominal overspecification has similarly found that speakers are extremely likely to attach "redundant" color descriptions to nouns, even when doing so provides no new information regarding which object is being referred to. However, in this case as well, there is evidence that most *overinformative* nominal modification is not in fact *overinformative*, as "overdescribing" an object can facilitate more rapid and efficient object identification. Here we will review some of the experimental work on informational redundancy, with a focus on interpretation of nominal overspecification.

Most experimental work on production and comprehension of informationally redundant utterances has focused on nominal modification in referent identification tasks, which typically instruct participants to look at or somehow engage with items such as: *the [red] apple*, *the [tall] boot* (Davies & Katsos 2010, 2013, Engelhardt et al. 2006, Nadig & Sedivy 2002, Pogue et al. 2016, Sedivy 2003). The aim of these studies has been to determine some combination of the following: 1) whether overinformative descriptions are perceived as infelicitous by comprehenders (i.e., whether overinformativeness apparently violates some communicative norm); 2) whether overinformativeness helps, hinders, or has no effect on object identification; 3) whether comprehenders attempt to accommodate overinformative descriptions by making inferences which increase the informational utility of the descriptions; and 4) whether comprehenders alter their beliefs about the rationality of the speaker (or the baseline informativeness of their speech) following use of overinformative descriptions.

What has been found is that in interactive, spontaneous speech, speakers frequently modify nouns with adjectives that are not strictly necessary for referent identification (e.g., referring to a cup as *the red cup*, in a context where there are no other cups of any color) (Engelhardt et al. 2006, Nadig & Sedivy 2002 : 30% and 50% of nominal descriptions were overspecified in spontaneous speech, respectively). Further, comprehenders frequently do not find such utterances infelicitous: Engelhardt et al. (2006) showed that comprehenders judge overinformative descriptions as significantly more acceptable than underinformative descriptions, and that overinformative descriptions do not trigger additional (e.g., contrastive) inferences. Davies & Katsos (2010) find that overinformative expressions are more likely to be produced, and less likely to be judged infelicitous, than underinformative expressions, although they are still judged to be suboptimal[6]. Sedivy (2003) showed that when comprehenders hear an object described with a clearly overinformative and prototypical color adjective (e.g., "yellow banana"), they make contrastive inferences (e.g., rapidly infer that a non-yellow banana must also be present).

What seems to emerge is that overinformative descriptions are easily tolerated when they describe perceptually useful or non-canonical properties, which may speed up object identification; and are more likely to be judged suboptimal, and/or trigger pragmatic inferences, when they don't. Indeed, Rubio-Fernández (2016) showed that experimentally increasing the perceptual usefulness of color adjectives causes them to be produced more frequently, as well as that color adjectives are more likely to be used for atypical than typical colors. In a related line of research, Pogue et al. (2016) found that after being exposed to a speaker repeatedly using overinformative (color or scalar) object descriptions, comprehenders are less likely to make generalization about the speaker's rationality or informativity than when they use underinformative descriptions. This suggests that comprehenders are relatively insensitive to deviations from "optimal" informativity that do not interfere with basic utterance comprehension, or else perceive them as relatively commonplace and inconsequential.

Overall, this work has shown that some types of informational redundancy may be helpful to the comprehender, and that informational redundancy in general is tolerated by comprehenders. There is, however, also evidence that informationally redundant utterances which have no apparent (e.g., perceptual) utility are unlikely to be produced, are generally judged to be relatively infelicitous, and tend to generate inferences. More generally, there is still some difficulty in distinguishing what constitutes informational redundancy, which creates difficulty in determining the precise theoretical implications of previous work (e.g., perceptually helpful "redundant" adjectives are questionably redundant in the first place, in the sense of having communicative utility). Additionally, these studies are limited by the fact that they uniformly focus on a very particular, and relatively concise variety of informational redundancy, which is further bound to a specific class of lexical items, raising the question of to what degree it's possible to generalize from the results. What this points towards is a need to look at informational redundancy in the context of utterances and constructions that are both quite costly for speakers, and have no readily apparent utility to comprehenders - either in terms of perception or comprehension, or in terms of facilitating the completion of a task. Further, we would argue that it's important to investigate constructions that are less bound to a specific set of

[6] However, Davies and Katsos purposefully use adjectives less likely to be produced spontaneously - color adjectives, by far the most likely to be used redundantly, are avoided, and the adjectives that they use are largely either inherently contrastive (e.g., "tall," "big"); or describe a default, assumed state (e.g., "unbroken egg," "fresh apple").

lexical items, and are more likely to be perceived as flouting of a conversational norm against redundancy - for example, complex and lengthy multi-word utterances such as those in Example (2).

2.2 Common ground assumptions

To date there has been relatively little work on the different strategies comprehenders might employ in making sense of an apparent violations of conversational maxims. Most work has focused on the scenario where a comprehender detects an apparent maxim violation, assumes that the speaker is in fact being cooperative, and comes up with an additional, non-literal meaning that the speaker may have intended (which repairs the apparent violation). Another strategy is simply to assume that the speaker is being plainly uncooperative, or that there is an intended meaning but that the comprehender is not privy to it, if no plausible intended meaning can be computed. A third strategy, which has received little attention, is that of modifying background assumptions about the world in which events take place, if doing so would repair the apparent violation. The lack of attention to this strategy is likely partially due to a focus on implicatures, or specifically intended meanings. To our knowledge, the only work to look at this in depth is Degen et al. (2015), which investigated comprehenders' willingness to revise their assumptions about the assumed common ground, in response to utterances whose pragmatic meaning would otherwise be inconsistent with it. They found that background assumptions about the world are surprisingly defeasible: comprehenders frequently accommodate the pragmatic meaning of utterances such as "***some** of the marbles sank*" (upon being thrown into a pool), by assuming that the utterances signify a strange scenario where physics doesn't quite work as expected. Further, a pre-utterance belief that a scenario is strange significantly increases the strength of the "*some, but not not all*" implicature that is then drawn by the comprehender.

In our case, if, as in Example (1), a speaker states that *John*, having gone shopping, *paid the cashier*, a comprehender might "repair" the redundancy by assuming that *John* does not in fact habitually pay the cashier. While this may occur parallel to an assumption that the speaker *intended* to use this utterance to communicate that *John* is not a cashier-paying individual, the strategy of modifying background assumptions can well proceed without any assumptions about speaker intent. Perhaps the comprehender is a third party not privy to the background knowledge of the speaker and intended listener, or perhaps the speaker isn't aware that the listener isn't familiar with *John*'s usual paying habits. In fact, in the case of our example, it seems relatively unlikely that a speaker would choose to communicate information about *John's* paying habits in this particular manner, making this an inference, but not an implicature. While most theoretical interest lies in implicatures, it's important to be able to model pre-utterance and changing post-utterance assumptions about the common ground, given that they have been demonstrated to have a marked effect on which inferences are drawn by comprehenders, as well as their strength (Degen et al. 2015; see also the literature on presupposition, e.g.: Stalnaker 1973). Further, an exclusive focus on intended meanings, rather than changes in background assumptions, may lead to erroneous conclusions that comprehenders are drawing no pragmatic inferences from an given utterance. In our paper, we explore a method for testing the shifting of background assumptions, collect data that can be used in the future to test formal models of pragmatic reasoning, and explore the willingness of comprehenders to shift background assumptions in different contexts.

2.3 Effect of implicit prosody on pragmatic interpretation

One of the questions we ask, relevant both to accurately detecting and modeling an effect of informational redundancy, is to what degree increased emphasis on the utterance (without changing the semantic content, or propositional value) influences the interpretation of informational redundancy. Although it is generally accepted that prosodic emphasis may influence utterance interpretation, there is very little empirical evidence that prosodic changes which contribute little by way of conventional meaning have a substantial effect on generation of pragmatic inferences[7]. One can however imagine that a redundant statement made loudly and confidently may lead a comprehender to believe that the speaker is very intentionally communicating that particular bit of information to them (Wilson & Sperber 2004 cf.), and that it should be taken seriously (signifying either that the speaker is being blatantly uncooperative by violating a communicative norm for no reason, or that there is an additional reason that the information was so purposefully transmitted). On the other hand, if a speaker vaguely mumbles an informationally redundant utterance under their breath, the comprehender might simply conclude that the speaker is reminding themselves of something, is unsure about what they really want to say, is mentally rehearsing a course of events, having some production difficulty, etc..

Along these lines, Bergen & Goodman (2015) hypothesize, on the basis of formal probabilistic models of pragmatic reasoning, that rather than focal/contrastive stress carrying conventional semantic meaning, the contrastive or exhaustive interpretation (“***BOB*** *went to the movies*” -> **only** *Bob* went to the movies) arises due to the comprehender perceiving the speaker as having made extra effort to communicate exactly that particular bit of info to them. They argue that an utterance which is increased in volume or duration is more likely to be attended to or accurately

[7] An exception is the effect of contrastive prosody (e.g., Kurumada et al. 2012), which is generally thought to be semantic – however, it has also been suggested that the effect of contrastive prosody is a pragmatic inference, as discussed in the following paragraph.

perceived by the comprehender and that, correpondingly, speakers can intentionally exploit this to signal to comprehenders that this particular utterance is important, and specifically meant to not be confused with any alternative utterances. On the basis of this and similar work, we therefore experiment with having participants interpret an informationally redundant utterance both with implicit exclamatory prosody (ending with an exclamation mark), as well as with no implicit prosody (ending simply with a full stop).

2.4 Context-dependent implicatures

To date, most formal or experimental research on pragmatic inferences has focused on the production and interpretation of scalar implicatures (Horn 1984, Levinson 2000), such as the use of *some* to implicate *not all*, or *warm* to implicate *not hot*. Non-generalized *ad-hoc* inferences, which arise only in specific contexts, have not received much attention from pragmaticists, experimentally or otherwise. Traditionally, scalar implicatures have been regarded as a separate class of *conventionalized* inferences which rely minimally on context or general reasoning about speaker intentions (Levinson 2000), and which arise from the use of specific lexical items (or classes of lexical items). In recent years this view has increasingly been challenged (Degen & Tanenhaus 2016, Grodner et al. 2010), with evidence indicating that the distinction between *conventionalized* (*generalized*) inferences, and *particularized* (*ad-hoc*) inferences is in any case not categorical, although the nature of differences between the two classes remains difficult to determine, and the latter case has traditionally been understudied.

Research on conventionalized inferences has been critical to developing formal linguistic theory, due to the role they play in disambiguating pragmatic and semantic contributions to utterance meaning. However, context-dependent (*ad-hoc*) inferences, which occur far more frequently and ubiquitously, are similarly important to developing a more general theory of human communication (as originally intended by Grice 1975). The body of experimental work teasing apart which properties of utterances trigger, alter, or modulate the strength of pragmatic inferences is still relatively small – however, having a more comprehensive model of cues which are taken into account by comprehenders, when interpreting utterances, is necessary both for building models of pragmatic reasoning, and for interpreting empirical results. In addition, there is a general need for further quantitative data on the specific conditions under which inferences are generated, in order to develop and test predictions of formal models of pragmatic reasoning (cf. Frank & Goodman 2012).

In sum, we argue that further empirical work on *ad-hoc* inferences and informational redundancy on the one hand, and contextual cues which modulate inferences drawn by comprehenders on the other, is important for developing comprehensive and predictive theories of communication, and for developing formal quantitative models of pragmatic reasoning.

3 Experimental Procedure

The three experiments presented in this paper are conceptual replications of 3 previously run studies, an account of which is available in the permanent online repository holding supplementary material for this article[8]. The current studies have an increased sample size, and were conducted concurrently on the same general population, to ensure that their results could be compared directly. The stimuli were also redesigned to read more naturally, and filler stimuli were included to ensure replicability[9].

The following experiments were run using the same interface, and on the same population of Amazon Mechanical Turk workers, in small rotating batches (of 9, or less): a batch of 9 participants completed the first experiment, after which the second experiment was scripted to go live until it was completed by 9 participants, and so forth. The only difference between the 3 experiments was the manipulation of prosody or discourse markers. Running them concurrently and on the same population therefore allows us to directly compare their results. All workers who participated in an experiment were automatically disqualified from participating in any future batches; i.e., no participant took part in more than one experiment or batch.

The number of eligible participants (n=2100) was predetermined through a simulation power analysis (adapted from Arnold et al. 2011): all predicted higher-order interactions, assuming effect sizes determined by the results of the experiments we are replicating, were detectable at $> .80$. The `R` code and a plot for the power analysis can be found in the supplementary materials[10].

4 Experiment 1: Implicit Prosodic Emphasis

We first test whether informationally redundant event descriptions trigger *non-habituality* inferences when the utterance is apparently effortful, intentional, and

[8] https://osf.io/8fz4m/?view_only=ff5859d3f33b485d95254395f95a52dc

[9] To note, in the original studies, participants saw each condition no more than once. We suspect that in crowdsourced studies, when participants can be exposed to a very small number of stimuli without repeating any experimental conditions, fillers may at times be unnecessary, provided there is sufficient difference between experimental conditions – and in fact, possibly detrimental to performance, if an increased number of items, and/or longer experiment duration, causes participants to read less closely for meaning.

[10] https://osf.io/84jdp/?view_only=ff5859d3f33b485d95254395f95a52dc

attentionally prominent – here signalled by an exclamation mark at the end of the utterance (this would disprove the "no inference" hypothesis). Intuitively, exclamatory intonation is a natural way of introducing something that may be noteworthy or unusual (Rett 2011), without otherwise altering the semantic content of the utterance. When inferences are context-dependent (and even if they are not; see Degen & Tanenhaus 2015), speakers generally provide multiple signals of their intended meaning, in order to make the inferences easier to compute for the comprehender. One would expect this to particularly be the case when the meaning of the utterance substantially violates expectations or previously held beliefs, as opposed to simply providing new but marginally expected (or at least unsurprising) information.

We present naive participants with a limited number of brief 'narratives,' which set up the common ground context, relationships between discourse participants, and some typical or atypical properties of their usual behavior (where relevant). Some of the narratives include brief dialogue between two discourse participants at the end (which may include informationally redundant or non-redundant event descriptions). After reading the narratives, participants rate how habitual they believe certain behaviors in the story to be. We expect participants who read informationally redundant event descriptions to infer that the utterance in fact signals that the event is relatively unexpected, or non-habitual (as only relatively unexpected events warrant explicit mention). In contrast, those participants who read non-redundant event descriptions should draw no such inferences.

4.1 Methods

4.1.1 Participants

700 eligible participants (760 total; median age bracket 26-35 ; 50 % female), were recruited on Amazon Mechanical Turk. The task was open only to workers located in the US, and with an approval rating of $\geq 95\%$. All workers were asked to state their native childhood language (with no penalty for stating a language other than English, to encourage accurate reporting), age bracket (under 18, 18-25, 26-25, and up, in intervals of 10), and gender. Those who did not indicate English, or listed their age as outside the interval of 18-65, were excluded from all analysis (60 ; 7.89 %), with additional participants recruited to replace them.

Those who did not provide accurate or plausible responses to the trial questions, all of which had a range of 'valid' and 'invalid' responses, were unable to proceed to the main task, and their data as a result was not recorded by the platform (e.g., those who rated the likelihood of 50% heads on multiple fair coin flips as low, compared to other possible outcomes)[11]. Participants were likewise unable to proceed in the study, or submit their results, without having answered all questions.

4.1.2 Design

The primary question of interest is whether informationally redundant utterances (in this case, descriptions of highly *habitual* activities) are perceived as potentially violating conversational norms at face value, and whether they consequently trigger pragmatic inferences. These inferences should lead to the revision of common-ground beliefs about the *habituality* of said activities (and so 'repair' the violation, or dip in informational utility):

(3) "John just came back from the grocery store. **He paid the cashier!**"

The bolded utterance here, given a 'default' or *ordinary* common ground, is ***informationally redundant***. We hypothesize that readers will infer that *John* does *not* habitually pay the cashier, as such a scenario would justify overt mention of *John's* cashier-paying. The informational redundancy arises due to the high *conceptual* (or *event*) *predictability* of *paying the cashier*, and is resolved if one assumes that this activity is not as habitual, or predictable as initially assumed.

We also wanted to see whether the inference (that an activity is less habitual than would otherwise be expected) could be cancelled by manipulating the common ground.

Common ground manipulation The activity described becomes 'non-habitual' given a *wonky* common ground[12] such as in (4), where the context suggests that typical assumptions (e.g., that some given individual would *pay the cashier* when they *go to the grocery store*) may not hold. At that point, the activity description ceases to be informationally redundant, and the inference should therefore not arise. This control condition keeps the description itself constant and manipulates only the common ground. It thus ensures that any effect we measure is in fact due to the presence of informational redundancy, and verifies that comprehenders are sensitive to discourse context.

(4) COMMON GROUND CONTEXT: John habitually doesn't pay. "John just came back from the grocery store. **He paid the cashier!**"

[11] Since this data was not recorded, we cannot report on the number of participants who were unable to proceed to the main task.

[12] We borrow the term *wonky* from Degen et al. (2015), where it is similarly used to describe non-default common grounds, in which typical rules as to how things proceed are expected to not hold, and which comprehenders may assume when encountering otherwise pragmatically infelicitous utterances.

Finally, we wanted to provide a baseline for 'typical' interpretation of non-redundant event descriptions; and to confirm that similarly structured descriptions of conventionally *non-habitual* activities, as in (5), do not provoke similar inferences (which would suggest a problem with the stimulus design or response measure). In (5), the utterance is not informationally redundant, and is not expected to generate any specific inferences. We also wanted to confirm that the *wonky* common ground in the previous example does not significantly affect the interpretation of conventionally *non-habitual* event mentions (which would suggest that there is an unexpected effect of context manipulation on stimulus interpretation, in general):

(5) CONTEXT: *Ordinary* **or** John habitually doesn't pay. "John just came back from the grocery store. **He got some apples!**"

As in (4), participants should draw no non-habituality inferences here, as the event described is not (typically) overly habitual. These conditions therefore provide a secondary control measure.

4.1.3 Materials

24 stimuli were constructed as brief stories/narratives, based on distinct stereotyped scripts or events. Each story had one of 2 context types (*ordinary* vs. *wonky* common ground, relative to the *conventionally habitual* script activity). In all stories, declarative utterances, spoken by one of the discourse participants, described one of 2 types of script activities (*conventionally habitual* vs. *non-habitual*), making a total of 4 conditions[13].

Conventionally habitual activities (6) can normally be inferred simply from the 'speaker' having invoked the script, while *non-habitual* activities (7) can not be inferred automatically, as they may only occasionally occur as part of the script activity. To clarify, we are using the term *conventionally habitual* to specify that the event almost invariably occurs as part of the event script (under normal conditions, and for typical individuals). Initial common ground was either *ordinary* ([1a] below) with respect to the script, or *wonky*, in that it implied the *conventionally habitual* event was in fact unusual for the event participant ([1b] below):

(6) CONVENTIONALLY HABITUAL EVENT

[13] The complete list of stimuli can be found in our online repository: https://osf.io/h5afr/?view_only=ff5859d3f33b485d95254395f95a52dc

<table>
<tr><td>[1a] John <b>often goes to the grocery store around the corner from his apartment</b>$_{\textbf{ordinary}}$</td><td>[1b] John <b>is typically broke, and doesn’t usually pay when he goes to the grocery store</b>$_{\textbf{wonky}}$</td></tr>
<tr><td colspan="2">[2] Recently, he came home from the store with groceries. When he came in, he saw his roommate Susan in the hallway, and started talking to her about his trip to the store. As he went to the kitchen to put his groceries away, Susan went to the living room, where their roommate Peter was watching TV.</td></tr>
<tr><td colspan="2">[3] Susan said to Peter: “John just came back from the grocery store. [4] He <b>paid the cashier</b>$_{\textbf{habitual}}$!”</td></tr>
</table>

The context/common ground manipulation in [1b] was used in order to render the *conventionally habitual* event unusual, or at least not habitual. Conventionally *non-habitual* activities could not be automatically inferred from the script having been invoked:

(7) NON-HABITUAL EVENT

<table>
<tr><td>[1a] John <b>often goes to the grocery store around the corner from his apartment</b>$_{\textbf{ordinary}}$</td><td>[1b] John <b>is typically broke, and doesn’t usually pay when he goes to the grocery store</b>$_{\textbf{wonky}}$</td></tr>
<tr><td colspan="2">[2] Recently, he came home from the store with groceries. When he came in, he saw his roommate Susan in the hallway, and started talking to her about his trip to the store. As he went to the kitchen to put his groceries away, Susan went to the living room, where their roommate Peter was watching TV.</td></tr>
<tr><td colspan="2">[3] Susan said to Peter: “John just came back from the grocery store. [4] He <b>got some apples</b>$_{\textbf{non-habitual}}$!”</td></tr>
</table>

Participants saw either only the common ground *context* [1] and *discourse setup* [2] (without numbering or special formatting), which enabled us to collect estimates of how habitual activities are believed to be, based on the context alone (*pre-utterance beliefs*); or the entire text, which enabled us to collect estimates of how habitual activities are believed to be, based on both the context *and* the event description [4] (*post-utterance beliefs*).

Following each passage, participants were queried as to how habitual they believed the *conventionally habitual* and *non-habitual* activities (as well as 2 other scenario-relevant distractor activities) were, for the person who was the subject of the discourse (the individual mentioned in the context [1] and event description [4]):

1. How often do you think John usually *pays the cashier*, when going shopping? 2. How often do you think John usually *gets apples*, when going shopping? 3. How often do you think John usually goes to the grocery store? 4. How often do you think Susan and Peter usually talk to each other?

Each question could be responded to on a continuous sliding scale of 'Never' to 'Always' (see Fig. 1). The slider itself was not visible until the participant clicked on the point on the scale that they thought was most appropriate, to avoid having people default towards a particular value. After they responded to all questions, participants could submit their answers. Once they did, the next passage was displayed on a new screen.

12 of the stimuli included 3 discourse participants – one of whom engaged in the script activity (*John*), the second who learned from that participant that they engaged in it (*Susan*), and the third to whom the second communicated this fact (*Peter*). The other 12 only included two – the subject of the discourse, who engaged in the activity (*John*), and the second participant to whom they communicated this fact (*Susan*). Compared to the example above, for instance, *John* might instead be communicating directly to *Susan*: "*I just got back from the grocery store. I paid the cashier!*".

The construction of these stimuli was constrained in several ways. The scripts (e.g., *going shopping*) needed to be sufficiently complex to include multiple subactivities or subroutines, and there needed to be habitual as well as non-habitual subactivities (*paying the cashier*, *getting apples*). It needed to be possible for the script to play out without the habitual activity having taken place – otherwise, the discourse would be incoherent, or the inference would not be drawn. For example, one arguably cannot play *tennis* at all, without using a *racket*. There was also established common ground between all discourse participants, so that all were plausibly (from the point of view of the reader) aware of the typical habits of the discourse subject, particularly with regard to the activity described. Finally, the activities needed to be sufficiently stereotyped and (relatively) culturally invariant, so that participants could be expected to agree on what a script entailed, which activities were or weren't obligatory to the script sequence, etc..

All stimuli were normed on 3 qualities (in separate tasks): whether the activity fell into the *habitual* or *non-habitual* activity bin; whether the common ground manipulation was effective; and whether participants found it plausible that the script could be engaged in without the *habitual* activity. For activity predictability norming, participants were asked to rate the habituality of the activity (on a 0-100 scale), with an arbitrary cutoff of 70 between activity types. *Non-habitual* activities were on average rated 48.0 (25.1-68.1), and *habitual* activities were rated 87.8 (78.1-95.2). For common ground norming, participants rated *habitual* activities in *ordinary* (mean 83.4 [72.2-96.9]) or *wonky* common grounds (mean 39.2 [20.7-62.0]), with a

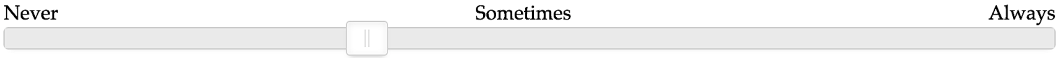

Figure 1 This is a slider, as used by experiment participants.

within-item difference between the two of at least 15 points (mean difference 44.2; [19.8-72.9]); *non-habitual* activities had to score below 70 regardless of common ground (mean 45.2; on average 10.7 points higher in the *ordinary* common ground). For plausibility norming, a statement in the form of '*John **went shopping**, but didn't pay the cashier*' was rated as either *coherent* (plausible) or *incoherent* (implausible), with criteria being a majority of participants finding the statement coherent (*habitual*: 91% [67%-100%]; *non-habitual*: 94% [80%-100%]).

4.1.4 Measures

To measure comprehender beliefs regarding activity habituality, each story we presented was followed by 4 questions presented in random order, regarding activities mentioned in the story (including both conventionally *habitual* and *non-habitual* activities associated with the stimulus item). The questions were accompanied by sliding scales which ranged from *Never* to *Always*, where participants could select any point along the scale, as seen in Fig. 1.

Prior to seeing any experimental items, participants were given several practice questions, unrelated to the experimental stimuli, which also used continuous sliding scales ranging from *Never* to *Always* (or similar). Unlike the experimental stimuli, these questions had 'correct' answers – such as *How likely is a fair coin to come up heads twice, if flipped 10 times? (very unlikely–very likely)*. If participants provided responses that could not be judged reasonably accurate, they were asked to re-read the instructions, and respond again, before they were able to proceed. This ensured that they were able to follow instructions, and were less likely to guess randomly throughout the experiment. There were no 'accurate' answers in the experiment itself. All points on the response scale were associated with a number ranging from 0 (*Never*) to 100 (*Always*).

Pre-utterance beliefs, or baseline beliefs regarding activity habituality, were estimated from responses to stimuli presented without the activity description (see the next section for a more detailed explanation). The responses, aside from setting baseline measures (*pre-utterance beliefs*) of activity habituality, also provide an additional norming measure for how likely it is that a particular activity would be engaged in, in the context of a given script. Thus, activities which are more or less habitual, within a given class, can be compared against one another.

Post-utterance beliefs regarding activity habituality were estimated from responses to stimuli which included the redundant or non-redundant utterance (activity description), or where the activity description/utterance was visible.

Belief change due to reading the activity description was determined by modeling the magnitude and direction of difference between *pre-utterance beliefs* and *post-utterance beliefs*.

4.1.5 Procedure

Participants were asked to read 6 experimental stimuli randomly selected out of the total of 24, as well as 4 filler items[14]. Each condition was only presented once, as follows. 2 of the stories were presented without the dialogue and event description (context and setting up of common ground only), and 4 stories were presented in their entirety (context, setting up of common ground, and the dialogue/event description). The 2 partial stories allowed us to collect measures of *pre-utterance beliefs* regarding activity habituality, and the 4 full stories gave us measures of *post-utterance beliefs* conditioned on the event description.

SUBJECT 1: *pre-utterance* belief	SUBJECT 2 *post-utterance* belief
<context> <setting up of common ground>	<context> <setting up of common ground> <dialogue>
#. <habituality question>	**#. <habituality question>**

The experiment thus employed a between-subject design for belief measures, where *pre-utterance* and *post-utterance* belief estimates for any given item were provided by different participants, to eliminate the possibility of participants conditioning their *post-utterance* estimates not only on inferences made from the text, but also on their own *pre-utterance* estimates[15]. The 4 filler stimuli had the same structure as above, but with the dialogue portion replaced by script-neutral utterances: “*You know, I'm really tired.*", “*Hey, do you know what time it is?*", “*So, what are you up to?*", or “*Have you heard the news today yet?*".

[14] To note, this means that each participant saw each manipulation only once, and the number of fillers was equal to the number of stimuli presented with dialogue.

[15] However, the results below largely mirror the results of a within-subjects version of the study reported in Kravtchenko & Demberg (2015).

4.2 Results

For the purposes of determining whether participants made any inferences regarding activity habituality, we modeled *belief change*, i.e. the difference between *pre-utterance* and *post-utterance* beliefs, or activity habituality estimates made with and without seeing the activity description. *Conventionally habitual* and *non-habitual* activities were modeled separately, as the conventionally *non-habitual* activity was used primarily as a control, and manipulations of common ground context did not otherwise target it. All factors were effect/sum coded.

4.2.1 *Conventionally habitual* activities ('Paid the cashier')

Pre-utterance belief ratings (obtained from participants who did not see the activity descriptions) showed that *ordinary* context activities are perceived as highly habitual (85.79 on a 0-100 scale). As predicted, *post-utterance belief* ratings (obtained from participants who saw the here, redundant, event descriptions) show lower habituality for the *ordinary context* activities (72.37) than *pre-utterance belief* ratings.

Wonky context activities (i.e., the condition where the *conventionally habitual* activity was made non-habitual by the common ground context) are perceived as relatively non-habitual a priori (48), and there was little change in participants' ratings (45.71 for *post-utterance beliefs*). The results are illustrated in Fig. 2, using violin plots.

A linear mixed effects regression analysis, the results of which are summarized in Table 1, showed that the interaction between context and belief measure is statistically reliable (β=-10.77 , p<.001). This interaction is driven by lowered activity habituality ratings when the readers see the utterance in a *ordinary* context (β=-13.21 , p<.001).

In this experiment as well as the two following experiments, we used linear mixed effects models with the maximal random effects structure that was justified by the design. This means that we included by-subject random intercepts and slopes for common ground context (*ordinary* / *wonky*) and belief measure (*pre-utterance* / *post-utterance*), as well as by-item random intercepts and slopes for both factors and their interaction (Barr et al. 2013). By-subject random slopes for the interaction were not included in the model, because we did not have any repeated measures for the interaction (each subject saw each condition only once). *P*-values were obtained using the Satterthwaite approximation for degrees of freedom, as implemented in the lmerTest package (Kuznetsova et al. 2017).

These results show that, as predicted, when a *conventionally habitual* activity is explicitly described in a *ordinary* common ground context (i.e. a context in which the activity can be automatically inferred), many readers infer that the *conventionally*

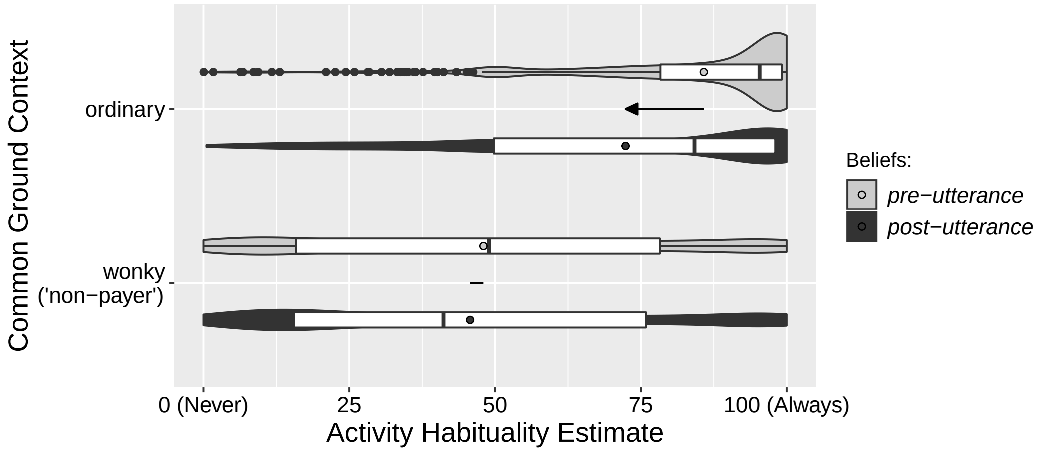

Figure 2 Experiment 1: *conventionally habitual* (*cashier-paying*) activity analysis. This plot shows changes in activity habituality estimates depending on whether the utterance is seen, as well as whether the context causes the utterance activity to be perceived as non-habitual. Violin plots, overlaid with box plots, show the distribution of estimates. A violin plot is simply a smoothed and mirrored histogram: the fatter the distribution at a given point, the more instances there are of that particular activity habituality estimate. Circles represent mean values. Arrows show statistically significant differences between *before/pre-utterance* and *after/post-utterance* ratings.

	β	SE(β)	**t**	**p**
Intercept	63.03	1.84	34.32	**<.001**
Common Ground: Ordinary	32.38	3.33	9.72	**<.001**
Belief: Post-utterance	−7.83	1.71	−4.58	**<.001**
Common Ground * Belief	−10.77	2.40	−4.50	**<.001**

Table 1 Experiment 1: *conventionally habitual* (*cashier-paying*) activity analysis. This table shows the beta coefficients associated with each main effect in the model, as well as corresponding standard errors, *t*-values, and significance levels.

habitual activity must in fact be *non-habitual*; i.e., unusual for the individual who is the subject of the story, and therefore worth mentioning explicitly.

4.2.2 Conventionally *non-habitual* activities ('Bought some apples')

There was little change in participants' ratings of conventionally *non-habitual* activities from *pre-utterance beliefs* to *post-utterance beliefs* (*ordinary*: 40.8 to 42.47 ; *wonky*: 38.49 to 39.56), see Fig. 3.

A linear mixed effects regression analysis showed that estimates of activity habituality do not vary with the common ground context, nor are they conditioned on the utterance describing the activity (see Table 2). This is also consistent with our predictions, and indicates both that the context alteration does not inherently cause a change in activity habituality estimates (regardless of how script-central the activity is), and that conventionally *non-habitual* activities, given our *ordinary* context, are not interpreted as less habitual when mentioned.

	β	SE(β)	**t**	**p**
Intercept	40.29	1.86	21.69	**<.001**
Common Ground: Ordinary	2.88	2.07	1.39	0.2
Belief: Post-utterance	1.34	1.85	0.73	0.5
Common Ground * Belief	0.01	2.14	0.01	1

Table 2 Experiment 1: conventionally *non-habitual* (*apple-buying*) activity analysis.

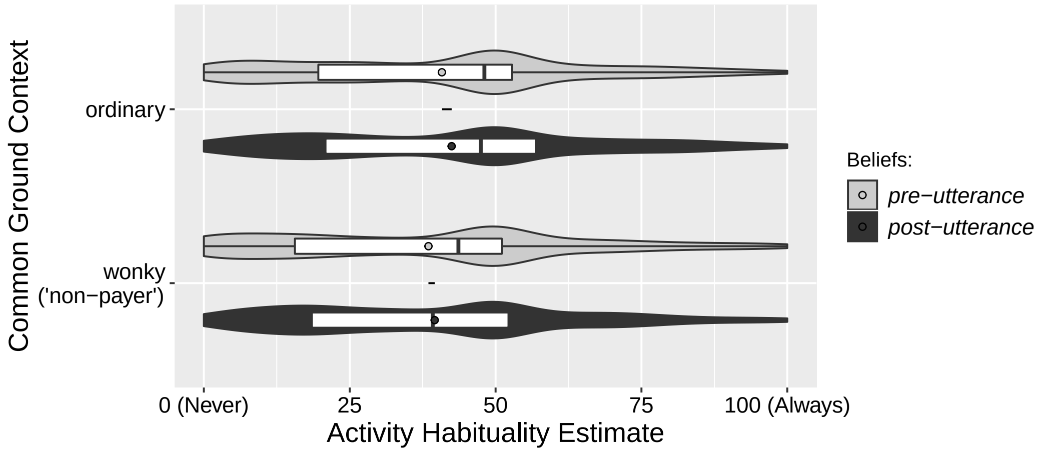

Figure 3 Experiment 1: conventionally *non-habitual* (*apple-buying*) activity analysis.

4.3 Discussion

The results of the first experiment indicate that comprehenders do in fact perceive informational redundancy, in the form of mention of overly habitual activities, as a possible violation of conversational norms, and that they resolve this violation by reinterpreting the activities described as non-habitual. On average, participants rate conceptually predictable activities as less habitual if they see them mentioned overtly, in contrast to all other activities. In other words, comprehenders react to redundancy as they typically do to other apparent maxim violations – by assuming an implied non-literal meaning, or alternate background world state, that resolves the apparent violation. This runs in some contradiction to the initial ambivalence Grice (1975) expressed about the existence of such a constraint, and equivocal evidence from studies of informationally redundant nominal modification.

These results rule out the “no inference" hypothesis outlined in Section 1.3, and raise two questions that we address in the following experiments, regarding the importance of (implicit) prosody, and that the speaker signaling intentionality of the activity description. First, an exclamation point may serve multiple purposes: it may signal surprise as to the course of described events, a speaker’s intentionality in communicating a piece of information[16], the importance and relevance of the

[16] I.e., the speaker displays clear and conscious intent to draw to the comprehender’s attention the face that a given event occurred – as opposed to stalling for time, thinking of something to say, aborting a previously planned utterance, simply being uncooperative, and so forth.

information conveyed to the general discourse and comprehender's interests, and that the information preceding the exclamation point constitutes an “encapsulated" message in its own right (rather than serving as a temporal or causal anchor[17]). Although it could be argued that the exclamation point (often a signal of surprise; Rett 2011) forces a relative ‘non-habitual activity’ interpretation independent of utterance informativity, this is not a likely explanation, as no signs of a similar effect are present in any of the other conditions.

Therefore, the first question is: how generalizable is the effect, and does the inference arise in contexts that do not implicitly signal the unexpectedness of the information conveyed (beyond the point that it is mentioned at all)? There is relatively little work on the question of which contextual cues specifically people employ in computing context-dependent inferences, as well as how these cues influence final interpretation. To test this, in Experiment 2 we use a discourse marker ("*Oh yeah, and...*") which does not clearly signal surprise – but does frame the event description as intentionally conveyed, as important/relevant to the topic at hand, and as an “encapsulated message."

The second question raised is whether informational redundancy itself is sufficient to trigger such an inference. As mentioned previously, we start from the premise that rational speakers mention only that which cannot be automatically inferred by the comprehender. A charitable comprehender may be expected to expend considerable effort on rescuing the assumption of a cooperative or rational speaker (Davidson 1974). If only activities under a certain threshold of habituality deserve mention, then comprehenders should conclude that the activity mentioned is relatively unusual, independently of any special emphasis on the utterance. In general, most types of inferences, if they occur, should occur as long as the semantic content of the utterance remain constant (cf. the “non-detachability" hypothesis).

On the other hand, pragmatic inferences must be calculable (Levinson 2000), and utterances must be attended to closely enough in the first place, before they may trigger any inferences (Wilson & Sperber 2004). That is, particularly for non-generalized (context-sensitive) inferences, the context must offer sufficient support that the reader can infer the speaker's intent, or a plausible background state, with reasonable certainty. It's not clear, in our case, if the blatant redundancy itself constitutes sufficient support. Likewise, while rational speakers may only mention activities that are not easily inferable, forcing a comprehender to expend significant effort on recovering an utterance's intended meaning or significance is not particularly rational behavior. The degree of “intentionality" on the part of the speaker (also signaled in our stimuli by the exclamation mark) may affect comprehenders' willingness and effort in guessing any implied meaning, as an

[17] For example: *He paid the cashier. Then he noticed it was his classmate.*

utterance that may be a stray thought uttered without any specific intent may not be worth much effort to attempt to decipher (cf. the “form sensitivity" hypothesis). To test whether informational redundancy itself is sufficient for triggering the inference, or whether some amount of discourse or prosodic emphasis is necessary for its generation, in Experiment 3 we present readers with the same task and stimuli, but strip the event description of prosodic or discourse cues signaling speaker intentionality.

5 Experiment 2: Implicit Discourse Support

The second experiment tests whether the effect, of informationally redundant event descriptions being interpreted by readers as signaling activity *non-habituality*, is generalizable. To do so, we can replace the exclamation point with a non-prosodic discourse marker that signals speaker intentionality and utterance relevance (but crucially, not surprise). In this experiment, we frame the informationally redundant event description as an apparent recalling of information specifically intended to be mentioned to the comprehender, and implicitly relevant to the material just discussed: “*Oh yeah, and [he paid the cashier].*"

This discourse marker does not clearly signal surprise at the activity having been engaged in, nor does it explicitly support the intended inference otherwise – and in contrast to the exclamation mark in Exp.1, is a non-prosodic manipulation of the event description. We therefore consider it a good test of whether the effect generalizes beyond the specific context used in the first experiment.

5.1 Methods

5.1.1 Participants

700 eligible participants (787 total; median age bracket 26-35 ; 51.3 % female) were recruited on Amazon Mechanical Turk. 87 participants were excluded from analysis (11.05 %), following the same exclusion criteria as applied in Experiment 1.

5.1.2 Design

The design of this experiment was motivated by the same considerations as Experiment 1 – with the exception of how the event description was framed. Instead of marking the target utterance with an exclamation mark, we framed the same utterance as a piece of information the speaker had just recalled, apparently having previously intended to mention it to the comprehender:

(8) “John just came back from the grocery store. **Oh yeah, and he paid the cashier.**"

The *oh yeah...* discourse marker does not conventionally signal surprise, and therefore does not potentially signal the specific inference that we are testing for. It does, however, imply speaker intent behind conveying precisely this message, the importance and relevance of the message to the current discourse and comprehender – as well as that the message stands alone, and is not intended to simply serve as causal or temporal scaffolding for a further message/event.

5.1.3 Materials

The same 24 stimuli were used as in Exp. 1. In this case, the critical utterance was prepended by “*Oh yeah, and...*":

(9) ORDINARY CONTEXT

<table>
<tr><td colspan="2">[1] John often goes to the grocery store around the corner from his apartment.</td></tr>
<tr><td colspan="2">[2] Recently, he came home from the store with groceries. When he came in, he saw his roommate Susan in the hallway, and started talking to her about his trip to the store. As he went to the kitchen to put his groceries away, Susan went to the living room, where their roommate Peter was watching TV.</td></tr>
<tr><td colspan="2">[3] Susan said to Peter: “John just came back from the grocery store.</td></tr>
<tr><td>[4a] **Oh yeah, and** he *paid the cashier*$_{habitual}$.”</td><td>[4b] **Oh yeah, and** he *got some apples*$_{non\text{-}habitual}$.”</td></tr>
</table>

5.1.4 Procedure

The procedure was identical to that of Exp. 1.

5.1.5 Measures

The same response measures as in Exp. 1 were used to estimate *pre-utterance beliefs* and *post-utterance beliefs*.

5.2 Results

As in Experiment 1, to determine whether participants made inferences regarding activity habituality, we modeled *belief change* - the difference between *pre-utterance* and *post-utterance* beliefs. *Conventionally habitual* and conventionally *non-habitual* activities were again modeled separately. All factors were effect/sum coded.

5.2.1 *Conventionally habitual* activities

As we predicted, *pre-utterance belief* ratings for *ordinary context* activities showed that these activities are judged to be highly habitual (84.71). As in Experiment 1, *post-utterance beliefs* about the habituality of *ordinary context* activities were significantly lower (73.84), and *wonky* common ground estimates remained stable (47.45 *pre-utterance* to 47.47 *post-utterance*).

A linear mixed effects regression analysis, the results of which are summarized in Table 3, showed an interaction between context and belief measure (β=-11.71 , p<.001), which is driven by lowered activity habituality ratings when the readers see the utterance in a ordinary context (β=-11.11 , p<.001). All model specifications are as described in Exp. 1. A plot illustrating the interaction can be seen in Fig. 4, which shows a pattern of results that is remarkably quantitatively and qualitatively similar to that of Exp. 1. Exp. 1 and 2 are compared directly, and to Exp. 3, in Section 7.

	β	SE(β)	t	p
Intercept	63.58	1.85	34.33	**<.001**
Common Ground: Ordinary	31.60	3.35	9.43	**<.001**
Belief: Post-utterance	−5.31	1.38	−3.83	**<.001**
Common Ground * Belief	−11.71	2.03	−5.76	**<.001**

Table 3 Experiment 2: *conventionally habitual* (*cashier-paying*) activity analysis.

These results support our prediction that readers perceive informationally redundant utterances as abnormal, and make pragmatic inferences (of activity *non-habituality*), regardless of whether implicit prosody or other markers conventionally associated with surprisal are present.

5.2.2 Conventionally *non-habitual* activities

In contrast to Experiment 1, there was some increase in participants' ratings of conventionally *non-habitual* activities from *pre-utterance beliefs* (*ordinary*: 40.3 to 43.22 ; *wonky*: 37.74 to 43.05), see Fig. 5.

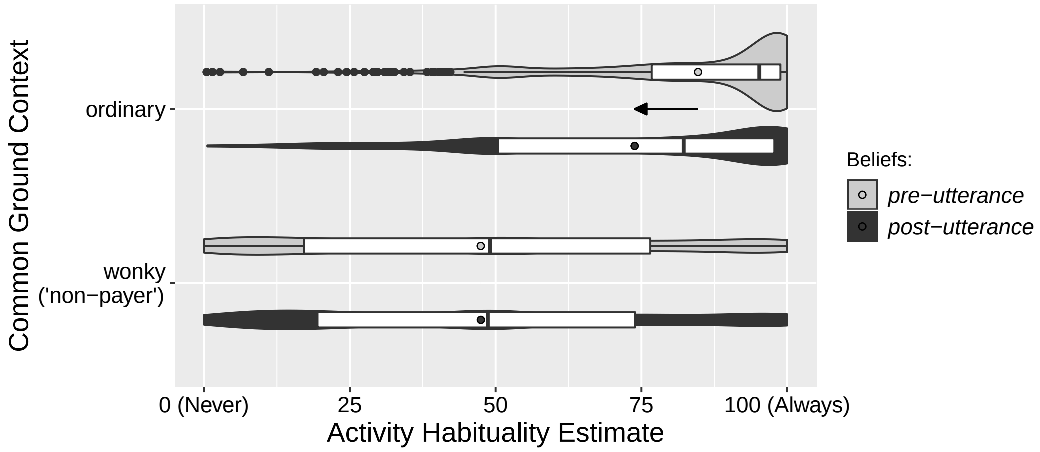

Figure 4 Experiment 2: *conventionally habitual* (*cashier-paying*) activity analysis.

A linear mixed effects regression analysis showed that estimates of activity habituality increase slightly when the utterance describing the conventionally *non-habitual* activity (see Table 4) is visible (β=5.09 , p<.01).

While not identical to the results of the first experiment (which showed a slight numerical increase in rating only), this is consistent with a peripheral prediction we made prior to running the experiments: simply mentioning a non-habitual, or non-redundant activity may increase the perception of its habituality, by providing some evidence that, e.g., *John* is at least an occasional *apple purchaser*. As the direction of this effect does not change our interpretation of the results, we leave it aside for future exploration.

	β	SE(β)	**t**	**p**
Intercept	40.99	1.85	22.14	**<.001**
Common Ground: Ordinary	0.95	1.83	0.52	0.6
Belief: Post-utterance	5.09	1.78	2.86	**<.01**
Common Ground * Belief	−1.22	1.55	−0.79	0.4

Table 4 Experiment 2: conventionally *non-habitual* (*apple-buying*) activity analysis.

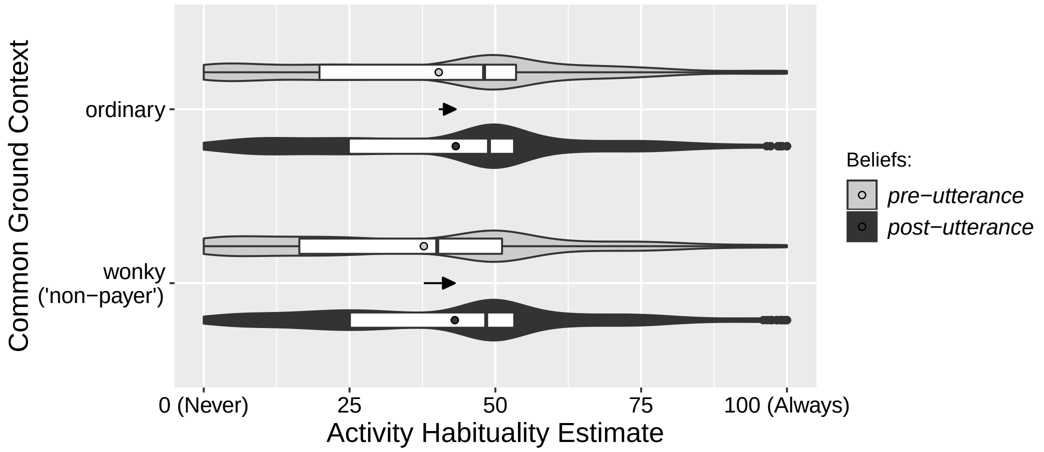

Figure 5 Experiment 2: conventionally *non-habitual* (*apple-buying*) activity analysis.

5.3 Discussion

Together with Experiment 1, these results show that readers find informational redundancy abnormal at face value, and make pragmatic inferences to reconcile apparent informational redundancy with their expectations of utterance utility. This further disconfirms the "no inference" hypothesis, and indicates that the effect is generalizable, and not dependent on conventional indicators of activity non-habituality, such as implicit exclamatory intonation.

The results of Experiments 1 and 2, however, do not permit us to distinguish between the 2nd and 3rd hypotheses ("non-detachability" vs. "form sensitivity"), as they leave open the question of whether the *non-habituality* effect is dependent on some degree of intentionality-signaling, or applies independently of discourse context. Experiment 2 provides some support for the "non-detachability" hypothesis, as the magnitude of the inference remains entirely stable, even as the form of intention or relevance signaling is substantially changed.

If the effect is dependent on some amount of relevance or intentionality signaling, this would support the "form sensitivity" hypothesis over the "non-detachability" hypothesis, by suggesting one of the following. Comprehenders may be relatively unwilling to expend substantial effort on decoding a plausible inference in the absence of evidence that doing so is worth it, and that the utterance has some amount of import. Similarly, they may stop short in their efforts, on the assumption that it is more likely that speakers would occasionally violate this particular conversational maxim, than that they would provide insufficient evidence that the utterance communicates something of note. Finally, they may simply be generally tolerant of informational redundancy, unless context suggests that the redundancy has a ‘point.’ Experiment 3 presents the same task and materials to participants, but removes the prosody or discourse markers that signal relevance and speaker intent.

6 Experiment 3: Removing Contextual Support

To investigate whether explicitly signaling speaker intent has an influence on the strength of the *non-habituality* effect, we designed a third experiment which differs only in the absence of special prosodic or discourse markers, or evidence for the relevance/informativity of the activity description. Our prediction is that while the effect may be attenuated somewhat, comprehenders should nevertheless make a measurable attempt to compensate for a violation in expected informational utility (i.e., while there may be some degree of "form sensitivity," the inference should nevertheless arise).

6.1 Methods

6.1.1 Participants

700 eligible participants (759 total; median age bracket 26-35 ; 51.6 % female) were recruited on Amazon Mechanical Turk. 59 participants were excluded from analysis (7.77 %), following the same exclusion criteria as applied as in previous experiments.

6.1.2 Design

The design was motivated by the same factors as Experiments 1 and 2, but all markers of relevance were removed from the activity description:

(10) "John just came back from the grocery store. **He paid the cashier.**"

In this case, there is no clear signal indicating the relevance or informativity of the utterance. One could plausibly imagine the event description, in this case, to be ‘filler material,’ only semi-intentionally uttered while the speaker is planning what to say next, or as (planned, but then possibly abandoned) temporal or causal scaffolding for a more important event to be described, such as in:

(11) "John just came back from the grocery store. He paid the cashier. *He then realized he’d forgotten his driver’s license!*"

6.1.3 Materials

The same 24 stimuli were used as in the previous experiments. The only alteration from Experiment 1 was the substitution of the exclamation point with a period:

(12) ORDINARY CONTEXT

[1] John often goes to the grocery store around the corner from his apartment.

[2] Recently, he came home from the store with groceries. When he came in, he saw his roommate Susan in the hallway, and started talking to her about his trip to the store. As he went to the kitchen to put his groceries away, Susan went to the living room, where their roommate Peter was watching TV.

[3] Susan said to Peter: “John just came back from the grocery store.

[4a] He *paid the cashier*$_{\textit{habitual}}$.” | [4b] He *got some apples*$_{\textit{non-habitual}}$.”

6.1.4 Procedure

The procedure was identical to that of previous experiments.

6.1.5 Measures

The same response measures as in the previous experiments were used to estimate *pre-utterance beliefs* and *post-utterance beliefs*.

6.2 Results

As in previous experiments, we modeled the difference between *pre-utterance* and *post-utterance* beliefs. *Conventionally habitual* and conventionally *non-habitual* activities were modeled separately. All factors were effect/sum coded.

6.2.1 *Conventionally habitual* activities

As in the previous experiments, *pre-utterance belief* ratings showed *ordinary context* activities to be highly habitual (85.59), and *wonky context* activities to be less habitual (49.5). Consistent with our predictions, *post-utterance beliefs* are significantly lower in the *ordinary context condition* (80.3), but less so than in the previous two experiments. Exp. 3 is compared directly to Exp. 1 and 2 in Section 7.

A linear mixed effects regression analysis, the results of which are summarized in Table 5, showed an interaction between context and belief measure (β=-5.4 ,

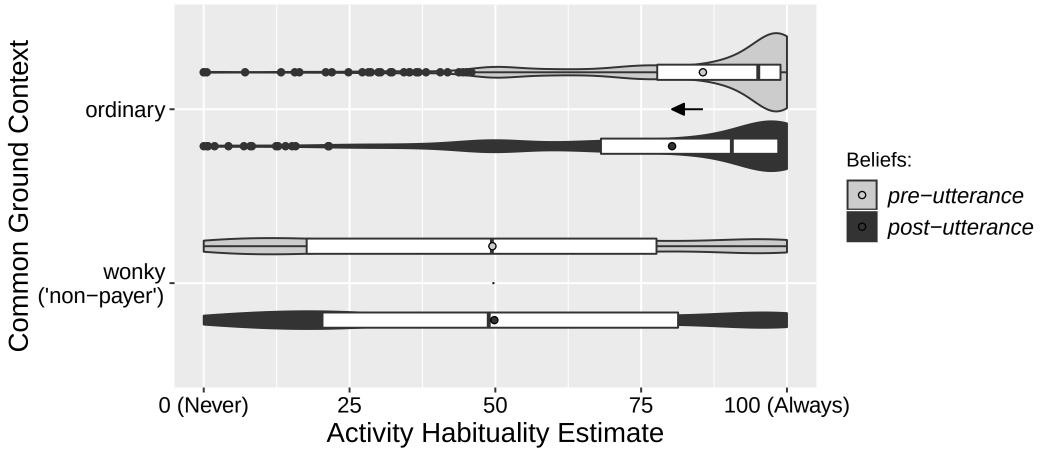

Figure 6 Experiment 3: *conventionally habitual* (*cashier-paying*) activity analysis.

p<.01), which is driven by lowered activity habituality ratings when the readers see the utterance in an ordinary context (β=-4.87 , p<.001). All model specifications are as described in Exp. 1 and 2. A plot illustrating the interaction can be seen in Fig. 6.

	β	SE(β)	**t**	**p**
Intercept	66.38	1.88	35.40	**<.001**
Common Ground: Ordinary	33.21	3.40	9.77	**<.001**
Belief: Post-utterance	−2.20	0.93	−2.36	**<.05**
Common Ground * Belief	−5.40	1.75	−3.10	**<.01**

Table 5 Experiment 3: *conventionally habitual* (*cashier-paying*) activity analysis.

These results indicate that, consistent with our predictions and the results of Exp. 1 and 2, when an easily inferable activity is overtly mentioned in a *ordinary* common ground context, comprehenders do infer some degree of activity non-habituality, even without implicit prosody or discourse markers putting additional emphasis on the utterance.

6.2.2 Conventionally *non-habitual* activities

In contrast to Experiment 1 and similar to Experiment 2, there was some increase in participants' ratings of conventionally *non-habitual* activities from *pre-utterance* to *post-utterance* beliefs (*ordinary*: 41.08 to 46.46 ; *wonky*: 37.61 to 44.42), see Fig. 7.

A linear mixed effects regression analysis showed that estimates of activity habituality do not vary with changes in the common ground context (or common ground *wonkiness*), but do increase slightly when the utterance describing the conventionally *non-habitual* activity (see Table 6) is visible (β=6.88 , p<.001). As in the case of Exp. 2, we suspect that explicitly mentioning a relatively unusual activity leads participants to believe that activity to be slightly more habitual than they would otherwise assume.

	β	SE(β)	t	p
Intercept	42.12	2.12	19.84	**<.001**
Common Ground: Ordinary	2.29	2.41	0.95	0.4
Belief: Post-utterance	6.88	1.77	3.88	**<.001**
Common Ground * Belief	−1.39	1.72	−0.81	0.4

Table 6 Experiment 3: conventionally *non-habitual* (*apple-buying*) activity analysis.

6.3 Discussion

In contrast to the results of the first two experiments, these results suggest that, when informationally redundant utterances are presented without a signal of speaker intent and utterance relevance, comprehenders are relatively unlikely to draw *non-habituality* inferences. This is consistent with the "form sensitivity" hypothesis described in Section 1.3, and the premise that, while rational speakers may typically avoid making utterances that have no literal or implied informational utility, and while such utterances may prompt pragmatic inferences on the part of comprehenders (which increase the informational utility of such utterances), such inferences are dependent on the degree to which the utterances are perceived as intentional. Further, while the results are not consistent with a strong form of the "non-detachability" hypothesis, they do broadly suggest that redundancy generates inferences regardless of form of delivery.

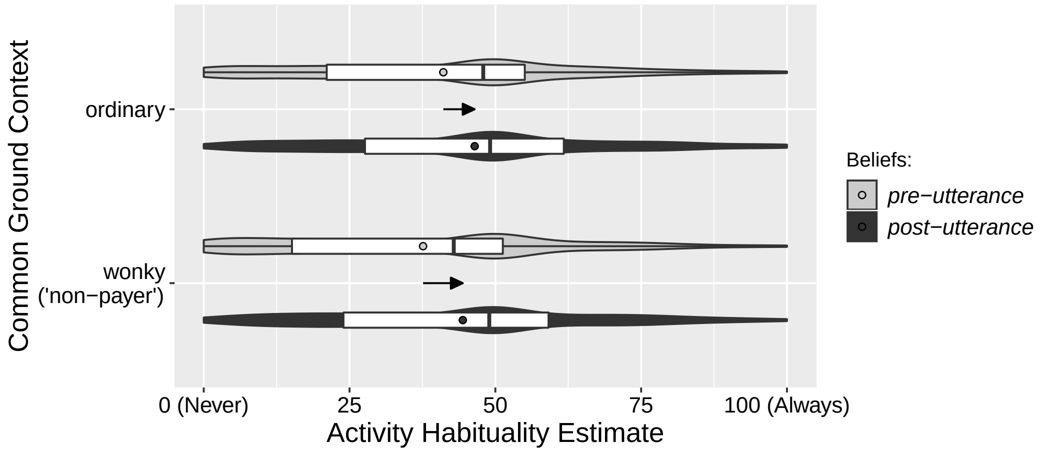

Figure 7 Experiment 3: conventionally *non-habitual* (*apple-buying*) activity analysis.

We should note, however, that this is not what we found in the experiments we are replicating - where the inference disappeared entirely without prosodic or discourse emphasis (strongly supporting the "form sensitivity" hypothesis, and at odds with the "non-detachability" hypothesis). Although the replicated experiments were not as highly powered, the difference might be due to stimulus redesign - in the supplementary materials[18], we speculate as to why this might be the case.

7 Cross-Experiment Analysis and Gradience of the Habituality Effect

In this section, we directly compare the results of the three experiments. We predict that informationally redundant utterances can trigger *non-habituality* inferences of similar magnitude independently of whether one uses an explicit marker of suprisal: in other words, that the effect is generalizable. However, we also predict that the effect is significantly attenuated in absence of a prosodic or discourse marker which signals relevance and speaker intent.

7.1 *Conventionally habitual* activities

To directly compare the three experiments, we run a 3 × 2 × 2 linear mixed effects regression analysis of *conventionally habitual* activities. We modeled *belief change*

[18] https://osf.io/8fz4m/?view_only=ff5859d3f33b485d95254395f95a52dc

(*pre-utterance* vs. *post-utterance* beliefs), as a function of common ground (*ordinary* vs. *wonky*), as well as the between-subject discourse marker manipulation (’*!*’ vs. ‘*Oh yeah, and*’ vs. ‘.’). The first two factors were effect/sum coded. We used Helmert coding for the 3-level experiment factor, as this allowed us to make the comparisons of theoretical interest: Exp. 1 vs. Exp. 2 (’*!*’ vs. ‘*Oh yeah, and*’), and then Exp. 3 vs. Exp. 1 and 2 grouped together (’.’ vs. the relevance markers).

The regression analysis showed a significant three-way interaction between relevance marker presence, common ground context, and belief measure: there was a significantly smaller *non-habituality* effect in Exp. 3 than in Experiments 1 and 2 (β=5.78 , p<.01), and no significant difference between Experiments 1 and 2 (β=-0.6 , p=.80).

We used the maximal converging model, with by-subject random intercepts and slopes for common ground context (*ordinary* / *wonky*) and belief measure (*pre-utterance* / *post-utterance*), by-item random intercepts and slopes for both factors and their interaction, and a by-item random slope for experiment. By-subject random slopes for the interaction were not included in the model due to lack of within-subject repeated measures. The random slope for the full (by-item) experiment by common ground by belief measure interaction was not included due to non-convergence. A plot illustrating the comparison can be seen in Fig. 8.

	β	SE(β)	**t**	**p**
Intercept	64.34	1.78	36.14	**<.001**
‘!’ vs. ‘Oh yeah...’	0.49	0.86	0.57	0.6
‘.’ vs. Relevance Markers	3.08	0.81	3.81	**<.001**
Common Ground: Ordinary	32.40	3.22	10.05	**<.001**
Belief: Post-utterance	−5.07	1.04	−4.86	**<.001**
‘!’ vs. ‘Oh yeah’ * Common Ground	−0.64	1.43	−0.45	0.7
‘.’ vs. Relevance Markers * Common Ground	1.25	1.24	1.01	0.3
‘!’ vs. ‘Oh yeah’ * Belief	2.50	1.30	1.92	0.1
‘.’ vs. Relevance Markers * Belief	4.52	1.13	4.01	**<.001**
Common Ground * Belief	−9.33	1.11	−8.41	**<.001**
‘!’ vs. ‘Oh yeah’ * CG * Belief	−0.60	2.35	−0.26	0.8
‘.’ vs. Relevance Markers * CG * Belief	5.78	2.03	2.84	**<.01**

Table 7 Experiments 1-3: *conventionally habitual* (*cashier-paying*) activity analysis.

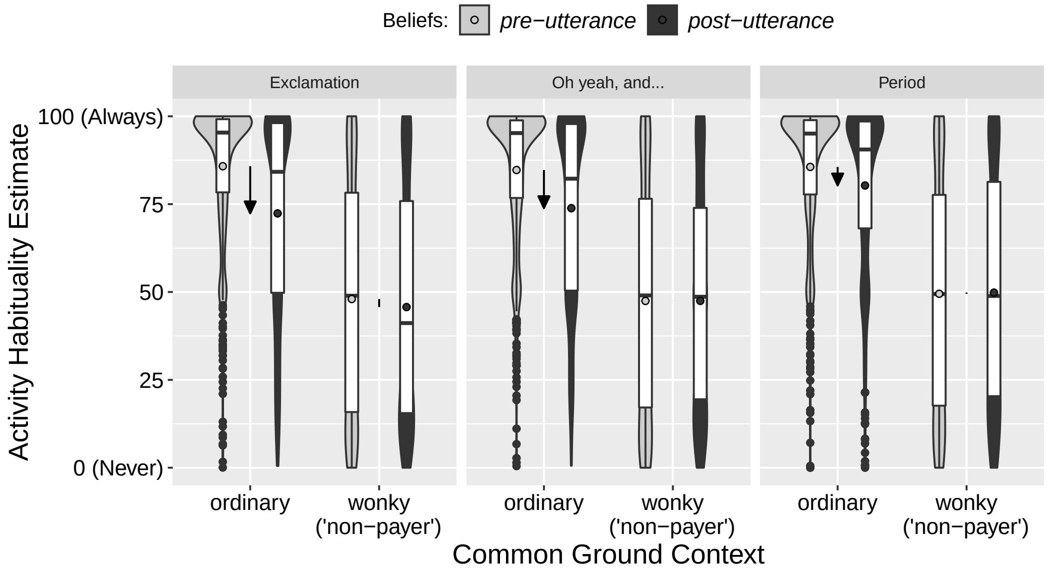

Figure 8 Experiments 1-3: *conventionally habitual* (*cashier-paying*) activity analysis.

The results are summarized in 7. As predicted, the effect holds regardless of which relevance marker is used, and in fact there is no statistically significant difference between the two markers. Further, the effect size of the common ground by belief measure interaction is significantly smaller in the absence of the markers; in other words, participants are significantly less likely to make a *non-habituality* inference in the absence of a prosodic or discourse marker signaling relevance or intentionality[19].

The effect direction is consistent across experimental items, with by-item common ground by belief measure interaction effect sizes ranging from -5.06 to -12.74 . We again note here that this set of 3 experiments is a replication of a previously run set with somewhat less naturalistic stimuli, a full description of which can be found in the supplementary materials linked to in the previous paragraph. In addition, the 'exclamation' experiment in that set is a further replication of a within-subects version (same stimuli), previously published as Kravtchenko & Demberg (2015), where participants updated their own ratings after seeing the utterance. We therefore argue that this is overall a robust and replicable effect.

This result clearly favors the "form sensitivity" hypothesis described in Section 1.3 over a strong version of the "non-detachability" hypothesis (which might predict an effect of the same magnitude for all experiments). We conclude that in the absence of a clear signal of utterance relevance or speaker intentionality, comprehenders are either less likely to attempt to resolve the violation, resolve it in a manner that is not reflected in our response measures, or do not detect the violation the first place. The first possibility is supported by observations that comprehenders approach speaker utterances *charitably*, and may expend significant effort on interpreting them in a manner that is consistent with the speaker making cooperative conversational choices (Davidson 1974). However, it is also possible that comprehenders are less 'charitable' in general when presented with oddly phrased psycholinguistic stimuli in an artificial setting – as well as less motivated on expending cognitive effort on calculating a non-obvious inference in a non-interactive environment, on the basis of an utterance that their attention is not otherwise drawn to.

Less charitable comprehenders, who may detect the redundancy but fail to in some way resolve it, may assume that the speaker is odd or not a particularly cooperative speaker, or perhaps that they are having production difficulties. Another possibility is that they assume the speaker is in the process of planning a more informative utterance (where, for example, the description might serve as a temporal/causal anchor; see Example (11)). Determining which strategies comprehenders do in fact resort to, and in which contexts, is left to future work. Finally, there is the possibility that, given the non-interactive experimental setting, comprehenders are

[19] A similar cross-experiment analysis of the conventionally *non-habitual* activities can be found in the supplementary materials: https://osf.io/8fz4m/?view_only=ff5859d3f33b485d95254395f95a52dc

processing the utterances at a relatively shallow level, and absent some (prosodic, discourse) indication that an utterance is somehow important, they do not expend effort on it (Sanford et al. 2006). To note, it has frequently been observed that comprehenders often do not make expected inferences in behavioral studies, for reasons that are not yet fully known (cf., Noveck & Posada 2003). Determining whether this plays a role in our studies is left to future work, as is the question of whether similar or stronger effects can be observed in less artificial, and/or more interactive settings.

7.2 Is the effect of habituality on pragmatic inferences gradient?

Fig. 9 plots the measured average activity habituality, with and without seeing the target utterance, for each item in each condition, for all three experiments. The diagonal dashed line demonstrates what the "no inference" hypothesis would predict: i.e., no effect of the utterance on belief change (*pre-utterance* ratings mapping straightforwardly onto *post-utterance* ratings). Points found above the line indicate that for those items, participants were more certain, for example, that *John usually buys apples* when the story mentioned that "he got some apples." Points below the line indicate a *non-habituality inference*: e.g., mentioning that "he paid the cashier" causes people to believe that *John does not usually pay the cashier*.

In Experiment 1 (exclamation mark), we see that for *ordinary* common grounds, and *conventionally-habitual* activities (e.g., *paying the cashier* given an ordinary common ground), most data points fall below the line, indicating a non-habituality inference. Interestingly, we also see a gradual 'trend' towards *non-habituality* in the other three (non-redundant) conditions: items that are similar to *ordinary habitual* items, in terms of pre-utterance habituality estimates, are more likely to trigger non-habituality inferences. In contrast, items with low pre-utterance habituality estimates show the opposite effect: i.e., if it's mentioned that an individual engaged in a particularly non-habitual activity, it leads comprehenders to believe that the individual is more likely to engage in that activity habitually. The same observations also hold for Experiment 2.

In Experiment 3 (period), we again see a gradual effect of *pre-utterance* beliefs regarding activity habituality on the likelihood of a *non-habituality inference*, but this time the slope of the regression line is shifted upwards (Exp. 1: β=0.64 ; Exp. 2: β=0.64 ; Exp. 3: β=0.76). We still see, however, that there is a gradient difference between highly expected vs. relatively expected events, in terms of likelihood of a non-habituality inference occurring.

Taken together, we can see in these figures that the exclamation mark and the '*oh yeah...*' discourse marker, as signals of speaker effort and intentionality, make it more likely that non-habituality inferences will arise for *ordinary* common ground,

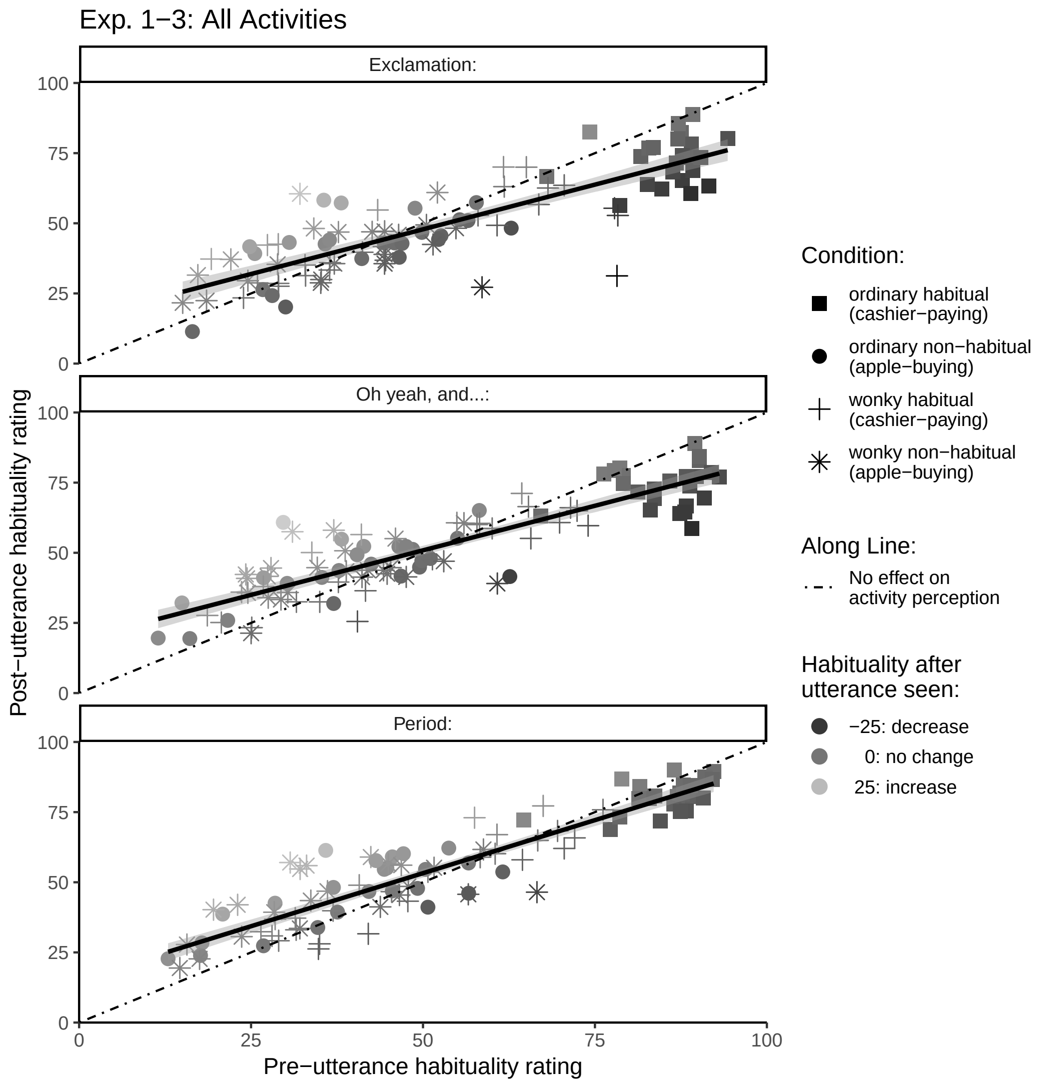

Figure 9 These plots show by-item belief change for all conditions of our three experiments. The dotted diagonal line represents the “no inference” hypothesis; i.e., what we would expect the data to look like if the critical utterance had no effect on habituality beliefs. The solid black line is a regression line with 95% CIs across all conditions. The shading of the points represents the degree and direction of *belief change*: negative/black indicates a *non-habituality* inference; positive/light gray indicates a perception of increased habituality.

habitual activity activity mentions. Furthermore, we can see that the effect of pre-utterance beliefs on non-habituality inferences is clearly gradient rather than binary: relatively more habitual activities, in all conditions, generally elicit larger *non-habituality* inferences.

8 General Discussion

Taken together, this series of experiments shows that comprehenders react to informationally redundant utterances by shifting their beliefs about the common ground, such that the utterances are more "informative" in context, thus increasing their utility. In other words, comprehenders expect for speaker utterances to have a certain level of informational utility, and they adjust their beliefs about the world and/or utterance meaning when such expectations are violated. In fact, this occurs even though informational redundancy, or overinformativity, in itself has no obvious negative impact on basic message comprehension. This is consistent with theoretical accounts of what constitutes "cooperative" communicative behavior (Grice, 1975), as well as of comprehenders' attempts to resolve speaker behavior that at face value does not appear particularly rational. However, as the third experiment shows, the effect is significantly modulated by how the utterance is framed in the discourse, supporting the hypothesis that inference strength is sensitive to utterance form. Overall, we provide robust evidence that informational redundancy is perceived as anomalous, and that comprehenders alter their situation models to accommodate it, particularly when there's evidence that there was specific intent behind the utterance.

As discussed in Section 2, while redundancy may not obviously impair comprehension (unlike underinformativeness), comprehenders may nevertheless prefer or expect that speakers be relatively concise, as it allows them to receive more information in a shorter span of time. Excessive redundancy may make it more difficult to follow the point of a conversation, or to reliably distinguish important from unimportant information. Provided that speakers are aware of comprehender preferences, and are likely independently motivated to conserve articulatory energy, it would be expected that comprehenders should perceive highly and unnecessarily redundant utterances (e.g., "*yellow banana*," "*he went shopping and paid the cashier!*") as such. Correspondingly, they should infer that the speaker is either not behaving rationally[20], or is conveying a message or background world state that is unusual from the comprehender's perspective. "Moderately" redundant utterances, such as when a speaker points out "*the **long** fork*" in the absence of another fork to compare, are not particularly useful in most tasks, but at least provide information that can't otherwise be inferred from the rest of the utterance, and do not require

[20] See the discussion of Gricean vs. Bayesian rationality in Section 8.1.1 below.

much additional articulatory effort. We believe that the clearest results on how redundancy is perceived should come from relatively costly utterances like the ones investigated here, or what we term *highly redundant utterances*.

Another area of contribution is that we illustrate a case in which comprehenders are willing to revise the assumed common ground of the discourse, in order to accommodate a perceived violation in the informational utility of an utterance. The redundant utterance violates conversational norms, or comprehender expectations, under the default assumed world state, but not the alternate "*wonky*" world state (e.g., one in which *John* is a habitual non-payer). Hearing such an utterance therefore biases comprehenders towards assuming that the alternate, or *wonky*, world state holds. Unlike shifting assumptions about intended utterance meaning, this is a strategy of accommodating potential violations of conversational norms that has not received much attention to date, with the notable exception of Degen et al. (2015). The shifting of common ground assumptions appears to be an important, and surprisingly understudied strategy for interpreting utterances that, at face value, violate conversational norms, and neglecting it as a possibility likely results in misinterpretation of online effects and under-detection of pragmatic inferences in experimental work.

Finally, we show that semantically "vacuous" utterance features (those that do not alter the propositional content of an utterance), in the form of implicit prosody or discourse markers, significantly influence the extent to which comprehenders are willing to draw an inference predicted by pragmatic theories of rational speaker behavior. Aside from the case of contrastive prosody (Bergen & Goodman 2015, Kurumada et al. 2012, Ward & Hirschberg 1985), this has not to date been systematically investigated in formal or experimental literature, and most likely also extends to other pragmatic phenomena. In our case, we argue that comprehenders are carefully weighing and evaluating multiple cues of how likely it is that a speaker intended to communicate a particular meaning, or that a deviation from expected utterance form and/or meaning signifies a common ground or background state that is substantially different from what was initially assumed.

8.1 Processing difficulty and surprisal

This subsection is primarily aimed at those who are interested in the processing cost of computing pragmatic inferences, or computational models of language processing. A question that we raise for future research is whether encountering informationally redundant utterances results in measurable processing difficulty on the part of comprehenders. We further argue that this has significant implications for current models of language processing.

As Walker (1993) points out, informationally redundant utterances are common in natural dialog - they therefore cannot be regarded as edge cases, and must be integrated into our models of language. We believe, however, that they pose several unique challenges for formal models of language processing. There are several potential sources of processing difficulty associated with such utterances, resulting on the one hand from processing the *surface form* (the particular string of words that comprises the utterance), and on the other hand from computing the pragmatic inference itself[21]. First, there could be processing difficulty associated with the (un)predictability of the *surface form* of the utterance: *John paid the cashier!* is an utterance we would not expect to hear in a ordinary context, as paying the cashier is normal, and reading unpredictable utterances such as this should cause some difficulty (Smith & Levy 2013)[22]. Second, we work on the assumption that context-dependent implicatures incur processing cost (Levinson 2000, Sedivy 2007), although there is evidence that processing may be relatively rapid, provided the context adequately supports the inference (Degen et al. 2015, Grodner et al. 2010).

8.1.1 Speaker rationality

First, we want to briefly talk about the link between Gricean notions of rationality, and an information-theoretic or Bayesian approach to rational speaker behavior. Rationality in the Gricean sense concerns whether speakers are constructing their utterances in a manner that is consistent with their goals, which is accurately communicating their message to a comprehender (Grice 1975). To this end, *underinformativeness* (saying less than needed), for example, is clearly inconsistent with this goal. Saying *more* than needed, however, does not clearly impair one's ability to accurately communicate a message - hence, the general uncertainty over whether overinformativeness violates Gricean norms. In the information-theoretic tradition, the speaker's goal is to expend no more energy than needed to accurately transmit a message (Jaeger 2010). Expending more effort than required to accurately transmit a message is inefficient from the speaker's perspective, and therefore not particularly rational, even while it is worse from a communication standpoint to not expend *enough* effort. The two traditions therefore make roughly similar predictions –

[21] Although there is debate currently over just how rapidly or efficiently comprehenders are able to make pragmatic inferences, much of the evidence converges on there frequently being some cost, even for relatively conventionalized inferences (Degen & Tanenhaus 2016).

[22] It should however be noted that while pragmatic processing must, on some level, incur cost, it may be sufficiently small and poorly localized that one would have difficulty detecting it using traditional online measures (eye-tracking, self-paced reading). Further, it is possible that the ease of semantic integration, in the case of conventionally habitual activities, would eclipse any difficulty due to the unpredictablity of the utterance itself – although the two exactly cancelling each other out, either way, is relatively unlikely.

weakly in the Gricean case, and more strongly in the information-theoretic: that redundancy should be avoided.

What the Gricean tradition adds to this mix is an idea of how comprehenders might interpret deviations from the communicative norm; traditional information-theoretic accounts make no predictions about how perceived utterance meaning might be altered when there's a mismatch between the expected and perceived utterance, beyond the possibility that intended utterance form or structure may be assumed to be something different from what was perceived, or that perception itself may be altered. Jaeger & Buz (2017) do note that if the speaker's aim is to accurately communicate a message, then they must take into account the signal, or *surface form*, that comprehenders expect to hear for that particular message. If they produce something that deviates from the expected signal, then even if the utterance is perceived accurately, and believed to have been perceived accurately, comprehenders may be led to assume a different intended message, which is more compatible with the signal that was in fact produced. It may be possible to fully reconcile the various traditions, but we leave this to future work (see, however, Frank & Goodman 2012, for a formal model of how speakers' and comprehenders' reasoning about each others' intentions might account for and predict utterance choice and pragmatic interpretation, while incorporating cost-based pressures on production).

While the above covers rationality from the point of view of the speaker, information-theoretic models of language processing propose that comprehenders also have strong expectations for how things will be said, and encounter processing difficulty (reflected by a variety of online measures) when these expectations are violated. In this tradition, what comprehenders specifically have expectations about is the *form* of utterances. After hearing something like *John went to the store*, they do not expect for *He paid the cashier!* to follow, as it is redundant: they are therefore surprised by the *form* the discourse has taken. The Gricean tradition similarly suggests that comprehenders have a base expectation that speakers will behave rationally, and interpret utterances literally or non-literally in a manner that will, generally, help to match this expectation (Grice 1975, Levinson 2000). We explicitly propose also that comprehenders have expectations as to the *state of the world* conveyed by the speaker. In the above example, the state of the world is precisely the one that is expected (i.e., one in which *John* has paid the cashier). The two forms of predictability - *form-based* and *meaning-based* - are often treated as essentially identical, but as we discuss in the following section, need to be disentangled to make accurate predictions about language processing.

8.1.2 Surprisal

An area where our work might have particular implications is in formal modeling of language processing. The mathematical concept of *surprisal* (Hale 2001, Levy 2008), traditionally, represents how (un)predictable a word or a string of words is in context. Specifically, it is the negative log of the probability of encountering a specific word or utterance. As hinted in the name, words or utterances that one might expect to see in a given context have *low* surprisal values, and those that one would *not* expect to see in a given context have *high* surprisal values. Smith & Levy (2013) show that difficulty in processing a word (or string of words), as reflected in online measures like reading times, is proportional to the word's unpredictability in context, or *surprisal*. In other words, comprehenders read or process words or utterances that are predictable (low *surprisal*) quickly, and those that are unpredictable (high *surprisal*) slowly. An utterance you don't expect to see in ordinary contexts (*John paid the cashier!*) should incur some processing difficulty for comprehenders. However, a problem with this account is that it treats all forms of *predictability* similarly. Consider, for example, two utterances that one might be (hypothetically) equally likely to hear: *John paid the cashier!* and *John punched the cashier!*. Processing theories which take into account only the predictability of an utterance would predict similar processing difficulty or processing times for both.

However, this is not only conceptually problematic, but would likely make the wrong predictions. Considering only *surface-level* or *form-based* predictability (the predictability of the string of words, in context) doesn't take into account the fact that the utterances are unpredictable for entirely different reasons: dispreference for redundancy, vs. event unpredictability. Further, the first (*cashier-paying*) utterance may contribute additional processing cost due to encountering pragmatic abnormality, or due to the need to make a pragmatic inference to resolve the apparent redundancy. In this case, despite identical surface-form predictability, we would expect that conceptually redundant utterances would be associated with greater processing difficulty. Of course, it may also be the case that conceptually redundant utterances are relatively easy to process, due to the relative ease of semantic integration (there are no unusual or unexpected facts to integrate into one's world model[23]). Either case, however, poses problems for the link between *surprisal* and processing difficulty, as utterances matched on predictability (and, consequently, *surprisal*) would still not end up with identical processing difficulty or reading times.

[23] Of course, our experiments make clear that many comprehenders do end up integrating an unusual common ground belief (*John is a habitual non-payer*) when trying to resolve the apparent pragmatic violation. For those comprehenders, we would predict that the processing cost would, in fact, be greater than the cost of simply integrating an unusual event into one's world model.

Several other interesting implications remain for surprisal theory, or the claimed link between *surprisal*, and reading times, or processing cost. First, it is commonly assumed that processing difficulty, in the context of this theory, is caused by encountering a particularly unexpected *form*. However, in our redundant *cashier-paying* example, the *form* of the utterance is unexpected *precisely because* the predictability of the utterance meaning is so high. In other words, in order for comprehenders to have expectations about the *form* of the utterance, they must also have expectations about the global *meaning* conveyed by the utterance, as it is precisely the meaning that renders the form surprising to comprehenders. We therefore consider it a significant shortcoming of these theories that they frequently either do not consider the predictability of meaning (what can also be termed *conceptual* predictability), beyond the truth-conditional meaning of an utterance, or treat the two probabilities, that of *form* and *meaning*, as essentially identical – whereas we have argued that they not only can influence each other, but in fact can diverge systematically at their extreme values. In the following subsection, intended for readers interested in language processing and formal language models, we talk about this relationship in more detail, as well as the implications it has for what *types* of language models could in principle address the issues we've outlined.

8.1.3 Formal models of language processing

The predictability of informationally redundant utterances, as we've mentioned, should be fairly low at the *surface* level, and reading times have been argued to reflect the predictability of *surface-level* linguistic events, rather than the conceptual predictability of the scenarios they describe, i.e., their broader *meaning* (Smith & Levy 2013). There is evidence, however, that comprehenders predict at multiple levels: for example, the *event* (in our case, *meaning*) level, as well as at the level of *surface form* (Kuperberg & Jaeger 2016), although it remains unclear how these levels interface (e.g., if comprehenders expect something predictable at the *event* level to go unmentioned at the *surface* level). In the case of surprisal theory (Hale 2001, Levy 2008), this may have interesting implications, given that the surprisal values that have been linked to processing times have largely been obtained using formal (computational) language models. If informationally redundant utterances result in longer reading times, it's unclear how formal models could accurately generate the high surprisal values one would expect for those utterances[24].

[24] For that matter, if they result in shorter reading times, as speculated in the previous section, there would similarly be a problem given that the processing difficulty should not rely simply on *surface-level* probability, but also on *event* or *meaning* probability, which, as we explain, current models cannot adequately integrate.

For example, in the case of our *conventionally habitual* event utterance, in the *ordinary* vs. *wonky* common ground, the event description (*John paid the cashier*) consists of exactly the same string of words, with the preceding context identical stretching over multiple preceding sentences. The utterance is informationally redundant in the *ordinary* context, and non-redundant in the *wonky* context. Simple or even complex n-gram models, which can't represent long-distance dependencies, would not show any difference in predictability, and therefore would predict no differences in processing difficulty. Relatively sophisticated models which incorporate syntax or semantics, similarly, would not predict a difference, as there are no meaningful differences in syntactic structure, and semantic models would not have access to the relevant event-based information which distinguishes the utterances.

Models of event sequences, which estimate *event* (vs. string) probability, may be able to estimate differences in predictability, and, consequently, processing difficulty, between utterances describing script-congruent and script-incongruent events (e.g., events likely and unlikely to be a part of *grocery shopping*). However, the general prediction such models would make is that the more congruent an event is with an invoked script (i.e., the more predictable the event is given the script), the more predictable (and easy to process) the utterances which describe that event should be. There is no principled way, within this framework, to divide activities up into different "grades" of predictability, such that utterances describing *moderately habitual* activities are easier to process than those describing *not-so-habitual* activities, yet those describing *very habitual* activities incur difficulty. In light of this, we suggest that to predict any processing difference between informationally redundant and non-redundant utterances, formal models of language comprehension would need to incorporate some form of pragmatic reasoning.

Although attempts to build formal or computational language models may appear to have limited relevance to how humans process language - which is typically thought of as a seamless integration of information from the surrounding context - it should be recognized that humans do not make predictions about upcoming material based simply on the preceding string of words, as formally assumed by simple models of language processing and prediction. The vast majority of word/utterance sequences have never been previously encountered by a comprehender, and predictions concerning upcoming material cannot be based on them alone. Regardless of the modeling approach one takes, it must be concluded that humans also make predictions by keeping track of certain cues - semantic, syntactic, lexical, and pragmatic (e.g., whether a speaker is generally adhering to conversational norms). Thus, determining the specific cues that are necessary to accurately model language processing is also relevant to understanding how humans accomplish the same task, and what information they must keep track of in order to do so. There are two ways of elucidating which linguistic and contextual cues influence language comprehension:

one may manipulate relevant cues in tightly controlled stimuli, and observe their influence on interpretation, or online measures such as reading times; or one may build formal language models which make specific, testable predictions regarding the influences of these cues on processing and comprehension. We believe that a combination of the two is likely to be the most fruitful approach.

To summarize, we think that it would be informative to investigate the processing of informationally redundant utterances, using online measures such as eye-tracking of self-paced reading. On the one hand, there are many claims, but still relatively little data on the online processing of pragmatic inferences, and little is known about the cost (or efficiency) of pragmatic reasoning (Degen & Tanenhaus 2016). The data that does exist is for the most part limited to scalar implicatures, which are often argued to be computed relatively automatically (but see Huang & Snedeker 2009). On the other hand, determining whether informational redundancy contributes to the processing cost of utterances, above and beyond the surface predictability of those utterances, is critical to determining whether formal language models need to integrate pragmatic reasoning to correctly predict processing cost. The main challenge to using online measures is that to compare (for example) the reading times of utterances, they must be matched on all factors which may affect reading times, but are irrelevant to the experimental manipulation (in a case such as ours: length, word frequencies, etc.). This makes it very difficult to compare reading times for utterances that are not identical in their surface form. One possibility is to compare reading times for otherwise identical phrases that are informationally redundant in one context, but not the other, as with our *cashier-paying* examples in the *ordinary* and *wonky* common grounds. We leave this to future work.

8.2 Perspectives for future work and conclusion

There are several avenues for further research. First, the range of inferences that comprehenders might draw from informationally redundant utterances may extend well beyond what we tested in this series of experiments. For instance, in the absence of a possible pragmatically felicitous interpretation, as the one suggested by our response measure, comprehenders may simply assume that a speaker is being uncooperative, having some production difficulty, or has unconventional speaking patterns (cf. Grodner & Sedivy 2011, Pogue et al. 2016). There is also the possibility that informationally redundant event descriptions, especially as seen in Experiment 3, are initially interpreted as likely, and possibly aborted, temporal or causal anchors for more “interesting” information. For example, in the context of a *grocery trip*, an “informationally redundant” description such as *John paid the cashier*, when followed by *with euros instead of dollars*, would likely not be considered anomalous. In this case, the description would not be redundant in its broader context, as it’s part of a more extended description that overall contributes previously unknown, or not easily inferable information. These hypotheses might be investigated using rating studies, sentence or passage completion studies, or more naturalistic tasks where participants' behavior provides a clue as to their interpretation of these utterances.

Overall, our results strongly suggest that, at least at face value, informational redundancy is perceived as anomalous. However, comprehenders are able to accommodate the provision of "unnecessary" information by altering their pre-utterance beliefs about individuals' behavior, or, more broadly, the common ground between speaker and listener. The results also complement work in the dialogue literature (Walker 1993), which illustrates that informationally redundant utterances are frequently used to convey "informationally useful" non-literal content. They raise presently important questions regarding which cues are systematically tracked by comprehenders, as well as how those cues are integrated during pragmatic interpretation. Finally, they address the pragmatic interpretation of complex utterances, not bound to specific classes of lexical items, which to date have largely been treated as either too complex or too idiosyncratic to study systematically.

References

Arnold, Benjamin F., Daniel R. Hogan, John M. Colford & Alan E. Hubbard. 2011. Simulation methods to estimate design power: An overview for applied research. *BMC Medical Research Methodology* 11(94).

Aylett, Matthew & Alice Turk. 2004. The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47. 31–56.

Baker, Rachel, Alastair Gill & Justine Cassell. 2008. Reactive redundancy and listener comprehension in direction-giving. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, 37–45. Columbus, Ohio.

Barr, Dale J, Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3). 255–278.

Bergen, Leon & Noah D. Goodman. 2015. The strategic use of noise in pragmatic reasoning. *Topics in Cognitive Science* 7(2). 336–350. http://dx.doi.org/10.1111/tops.12144.

Bower, Gordon H., John B. Black & Terrence J. Turner. 1979. Scripts in memory for text. *Cognitive Psychology* 11. 177–220.

Cohen, Philip R. 1978. *On knowing what to say: planning speech acts*. Doctoral Dissertation, University of Toronto.

Davidson, Donald. 1974. Belief and the basis of meaning. *Synthese* 27. 309–323.

Davies, Catherine & Napoleon Katsos. 2010. Over-informative children: production/comprehension asymmetry or tolerance to pragmatic violations? *Lingua* 120. 1956–1972.

Davies, Catherine & Napoleon Katsos. 2013. Are speakers and listeners 'only moderately gricean'? an empirical response to engelhardt et al. (2006). *Journal of Pragmatics* 49(1). 78–106.

Degen, Judith & Michael K. Tanenhaus. 2015. Processing scalar implicature: A constraint-based approach. *Cognitive Science* 39(4). 667–710.

Degen, Judith & Michael K. Tanenhaus. 2016. Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science* 40(1). 172–201.

Degen, Judith, Michael H. Tessler & Noah D. Goodman. 2015. Wonky worlds: Listeners revise world knowledge when utterances are odd. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci2015)*, 548–553.

Engelhardt, Paul E., Karl G. D. Bailey & Fernanda Ferreira. 2006. Do speakers and listeners observe the gricean maxim of quantity? *Journal of Memory and Language* 54. 554–573.

Fillmore, C. J. 2006. Frame semantics. In *Encyclopedia of language and linguistics*, 613–620. http://dx.doi.org/10.1016/B0-08-044854-2/00424-7.

Frank, Michael C. & Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science* 336(6084). 998–998.

Grice, H. Paul. 1975. Logic and conversation. In Peter Cole & Jerry L. Morgan (eds.), *Syntax and semantics: Vol. 3: Speech acts*, 41–58. New York: Academic Press.

Grodner, Daniel & Julie C. Sedivy. 2011. *The processing and acquisition of reference* chap. The effects of speaker-specific information on pragmatic inferences, 239–272. MIT Press: Cambridge, MA.

Grodner, Daniel J., Natalie M. Klein, Kathleen M. Carbary & Michael K. Tanenhaus. 2010. "some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition* 116(1). 42–55.

Hale, John. 2001. A probabilistic earley parser as a psycholinguistic model. *Association for Computational Linguistics* 1–8.

Horn, Larry. 1984. Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (ed.), *Meaning, form and use in context*, 11–42. Washington: Georgetown University Press.

Huang, Yi Ting & Jesse Snedeker. 2009. Online interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology* 58(3). 376–415.

Jaeger, T. Florian. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61(1). 23–62.

Jaeger, T. Florian & Esteban Buz. 2017. *The handbook of psycholinguistics* chap. Signal reduction and linguistic encoding, 38–81. Wiley-Blackwell. To appear.

Kravtchenko, Ekaterina & Vera Demberg. 2015. Semantically underinformative utterances trigger pragmatic inferences. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci2015)*, 1207–1212.

Kuperberg, Gina R. & T. Florian Jaeger. 2016. What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience* 31(1). 32–59.

Kurumada, Chigusa, Meredith Brown & Michael K. Tanenhaus. 2012. Pragmatic interpretation of contrastive prosody: It looks like speech adaptation. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society (CogSci2012)*, 647–652.

Kuznetsova, Alexandra, Per B. Brockhoff & Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13). 1–26.

Levinson, Stephen C. 2000. *Presumptive meanings - the theory of generalized conversational implicature*. The MIT Press.

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106. 1126–1177.

Mahowald, Kyle, Evelina Fedorenko, Steven T. Piantadosi & Edward Gibson. 2013. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition* 126(2). 313–318.

Minsky, Marvin. 1975. A framework for representing knowledge. In P. H. Winston (ed.), *The psychology of computer vision*, New York: McGraw-Hill.

Nadig, Aparna S. & Julie C. Sedivy. 2002. Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science* 13. 329–336.

Noveck, Ira A. & Andres Posada. 2003. Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language* 85(2). 203–210.

Pogue, Amanda, Chigusa Kurumada & Michael K. Tanenhaus. 2016. Talker-specific generalization of pragmatic inferences based on under- and over-informative prenominal adjective use. *Frontiers in Psychology* 6(2035).

Regneri, Michaela, Alexander Koller & Manfred Pinkal. 2010. Learning script knowledge with web experiments. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 979–988.

Rett, Jessica. 2011. Exclamatives, degrees and speech acts. *Linguistics and Philosophy* 34(5). 411–442.

Rubio-Fernández, Paula. 2016. How redundant are redundant color adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology* 7. 153.

Sanford, Alison J. S., Anthony J. Sanford, Jo Molle & Catherine Emmott. 2006. Shallow processing and attention capture in written and spoken discourse. *Dis-*

course Processes 42(2). 109–130.

Schank, Roger C. & Robert P. Abelson. 1977. *Scripts, plans, goals and understanding*. Hillsdale, NJ: Lawrence Erlbaum.

Sedivy, Julie C. 2003. Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research* 32(1). 3–23.

Sedivy, Julie C. 2007. Implicature during real time conversation: A view from language processing research. *Philosophy Compass* 2(3). 475–496.

Smith, Nathaniel J. & Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition* 128(3). 302–319.

Stalnaker, Robert. 1973. Presuppositions. *Journal of Philosophical Logic* 2(4). 447–457.

Walker, Marilyn A. 1993. *Informational redundancy and resource bounds in dialogue*. Doctoral Dissertation, University of Pennsylvania, Philadelphia, PA.

Ward, Gregory & Julia Hirschberg. 1985. Implicating uncertainty: The pragmatics of fall-rise intonation. *Language* 61. 747–776.

Wilson, Deirdre & Dan Sperber. 2004. Relevance Theory. In Laurence R. Horn & Gregory Ward (eds.), *The handbook of pragmatics*, vol. 1, 606–632. Oxford, UK: Blackwell Publishing. http://dx.doi.org/10.1002/9780470756959.ch27. http://www.dan.sperber.fr/wp-content/uploads/2004{_}wilson{_}relevance-theory.pdfhttp://doi.wiley.com/10.1002/9780470756959.ch27.

Zwaan, Rolf A., Joseph P. Magliano & Arthur C. Graesser. 1995. Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21(2). 386–397.

A Replicated Work

A.1 Replicated experiments

Here we present a previous iteration of this series of experiments, using the same design as that reported in the main body of the paper, but run on separate populations (as opposed to concurrently), and with a slightly different set of stimuli. We include these results here as evidence that the effects we report are robust, replicating closely despite being run on a different population, substantial revision of the stimuli to improve naturalness, addition of filler stimuli, and a larger amount of data being collected to improve power for all relevant comparisons.

A.1.1 Methods

Participants 1200 eligible participants (1242 total), 400 per experiment, were recruited on Amazon Mechanical Turk, with the task only open to workers located in the US, and with an approval rating of $\geq 95\%$. Participants who did not report their native language, or reported their native language as other than English, were excluded (42; 3.38%), with additional participants recruited to replace them.

Materials The design was identical to that reported in the paper, aside from the inclusion of fillers, as each participant saw only 6 stimuli and no condition more than once, with all stimuli differing across multiple non-critical dimensions. We therefore reasoned that there was little likelihood of learning the purpose of the experiment in the course of the task, and there was risk of increased task length/tedium decreasing the likelihood of participants reading passages closely enough to pick up on relatively subtle effects.

The stimuli in the replicated experiments were constructed to minimize variation in syntactic and information structure, as well as length, between stimuli. However, this came at the cost of naturalness. Here we present a stimulus example:

(13) Original Stimulus

[1a] John often *goes to his local supermarket, as it's close by*$_{ordinary}$.	[1b] John often *doesn't pay at the supermarket, as he's typically broke*$_{wonky}$.
[2] Today he entered the apartment with his shopping bags flowing over. He ran into Susan, his best friend, and talked to her about his trip. Susan then wandered over to Peter, their roommate, who was in a different room.	

<table>
<tr><td colspan="2">[3] Recently, he came home from the store with groceries. When he came in, he saw his roommate Susan in the hallway, and started talking to her about his trip to the store. As he went to the kitchen to put his groceries away, Susan went to the living room, where their roommate Peter was watching TV.</td></tr>
<tr><td colspan="2">[4] She commented: “John went shopping.</td></tr>
<tr><td>[5a] He <b>paid the cashier</b>$_{\textbf{habitual}}$!</td><td>[5b] He <b>got some apples</b>$_{\textbf{non-habitual}}$!</td></tr>
<tr><td colspan="2">[6] I just saw him in the living room.”</td></tr>
</table>

Procedure The procedure was identical to that of the other experiments.

Measures The same response measures as in the other experiments were used to estimate *pre-utterance beliefs* and *post-utterance beliefs*.

A.1.2 Results

As in the experiments reported in the main body of the paper, we modeled the difference between *pre-utterance* and *post-utterance* beliefs. *Conventionally habitual* and conventionally *non-habitual* activities were modeled separately. All binary factors were effect/sum coded, and the experiment factor was Helmert coded.

***Conventionally habitual* activities** (‘Paid the cashier’)

The regression analysis showed a significant three-way interaction between discourse marker presence, common ground context, and belief measure: there was a significantly smaller *atypicality* effect in Exp. 3 than in Experiments 1 and 2 (β =6.16, p <0.05), and no significant difference between Experiments 1 and 2 (β =0.42, p =0.89).

We used the maximal converging model, with by-subject random intercepts and slopes for common ground context (*ordinary* / *wonky*) and belief measure (*pre-utterance* / *post-utterance*), as well as by-item random intercepts and slopes for all factors. By-subject random slopes for the interaction were not included in the model, because we did not have any repeated measures for subjects for the interaction. By-item random slopes for the interactions were not included in the model due to nonconvergence. A plot illustrating the higher-order experiment by common ground by belief measure interaction can be seen in Figure 10.

A similar analysis of the conventionally *non-habitual* activities can be found in the next section of these materials.

	β	SE(β)	**t**	**p**
Intercept	61.22	2.11	29.01	**<.001**
'!' vs. 'Oh yeah...'	1.30	1.02	1.28	0.2
'.' vs. Relevance Markers	4.34	0.88	4.92	**<.001**
Common Ground: Ordinary	38.04	3.72	10.22	**<.001**
Belief: Post-utterance	−0.46	1.42	−0.32	0.8
'!' vs. 'Oh yeah' * Common Ground	−0.87	1.82	−0.48	0.6
'.' vs. Relevance Markers * Common Ground	2.44	1.57	1.55	0.1
'!' vs. 'Oh yeah' * Belief	0.61	1.76	0.35	0.7
'.' vs. Relevance Markers * Belief	6.68	1.52	4.39	**<.001**
Common Ground * Belief	−12.66	1.27	−9.97	**<.001**
'!' vs. 'Oh yeah' * CG * Belief	0.42	3.11	0.14	0.9
'.' vs. Relevance Markers * CG * Belief	6.16	2.69	2.29	**<.05**

Table 8 Replicated Experiment 1-3: *conventionally habitual* (*cashier-paying*) activities analysis.

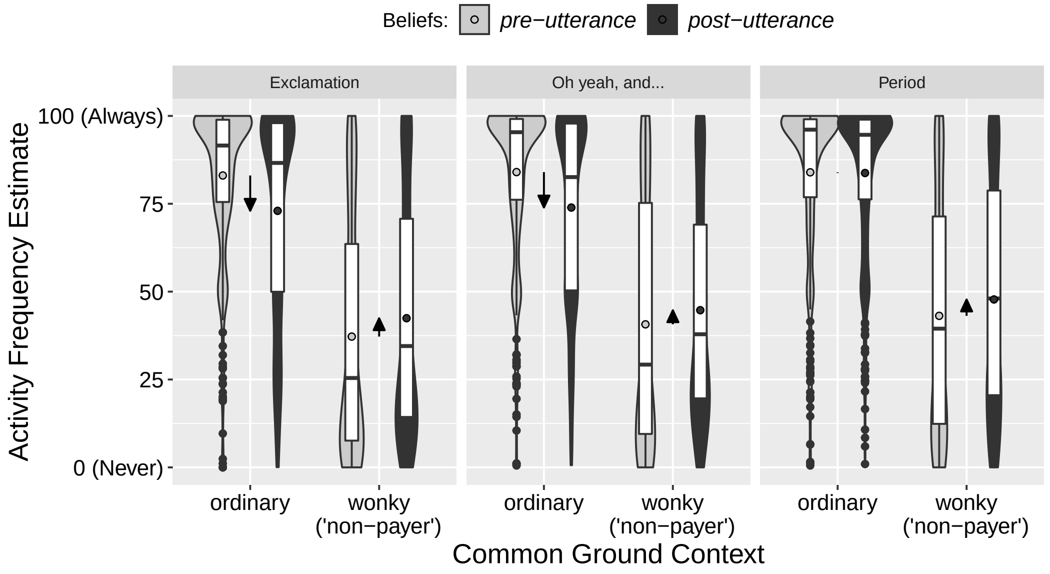

Figure 10 Replicated Experiments 1-3: *conventionally habitual* (*cashier-paying*) activities analysis.

A.1.3 Discussion

Overall, the results of these experiments were broadly replicated by those reported in the main body of the paper. The only salient difference is that in the original iteration of Exp. 3, there was no measurable effect of informational redundancy on perceptions of activity typicality, while in the 'new' Exp. 3, there was a significant, but diminished effect, as we had originally predicted. The absence of a significant effect in the first iteration surprised us, and we attribute it to either chance (possibly due to fewer subjects run) or to increased prominence of the utterance in the revised stimuli. To compare:

(14) REVISED: "John just came back from the grocery store. **He paid the cashier**."

(15) ORIGINAL: "John went shopping. **He paid the cashier**. I just saw him in the living room."

The utterance in question appears more discourse-prominent in the revised version of the stimuli, as it is utterance-final (i.e., we removed the last sentence), and in general competes with fewer adjacent utterances for attention. We leave it to future work to definitively answer whether the minor change in utterance prominence does indeed eliminate the effect entirely.

	β	SE(β)	t	p
Intercept	41.18	1.89	21.84	**<.001**
'!' vs. 'Oh yeah...'	0.84	0.72	1.17	0.2
'.' vs. Relevance Markers	1.50	0.62	2.41	**<.05**
Common Ground: Ordinary	2.00	1.97	1.02	0.3
Belief: Post-utterance	4.51	1.70	2.65	**<.05**
'!' vs. 'Oh yeah' * Common Ground	−2.11	1.05	−2.02	**<.05**
'.' vs. Relevance Markers * Common Ground	0.34	0.91	0.38	0.7
'!' vs. 'Oh yeah' * Belief	3.71	1.11	3.34	**<.001**
'.' vs. Relevance Markers * Belief	3.76	0.96	3.90	**<.001**
Common Ground * Belief	−0.73	1.43	−0.51	0.6
'!' vs. 'Oh yeah' * CG * Belief	−1.34	2.07	−0.65	0.5
'.' vs. Relevance Markers * CG * Belief	−0.64	1.79	−0.35	0.7

Table 9 Experiment 1-3: *conventionally non-habitual* (*apple-buying*) activities analysis.

A.2 Conventionally *non-habitual* activities ('Bought some apples')

Here, we present the results of our cross-experiment analyses of *non-habitual* activities.

A.2.1 Experiments reported in paper

We used the maximal converging model, with by-subject random intercepts and slopes for common ground context (*ordinary* / *wonky*) and belief measure (*pre-utterance* / *post-utterance*), by-item random intercepts and slopes for both factors and their interaction, and a by-item random slope for experiment. By-subject random slopes for the interaction were not included in the model due to lack of within-subject repeated measures in our data for the interaction. The random slope for the full by-item experiment by common ground by belief measure interaction was not included due to non-convergence.

The results are shown in Table 9 and Figure 11.

A.2.2 Replicated experiments

We used the maximal converging model, with by-subject random intercepts and slopes for common ground context (*ordinary* / *wonky*) and belief measure (*pre-*

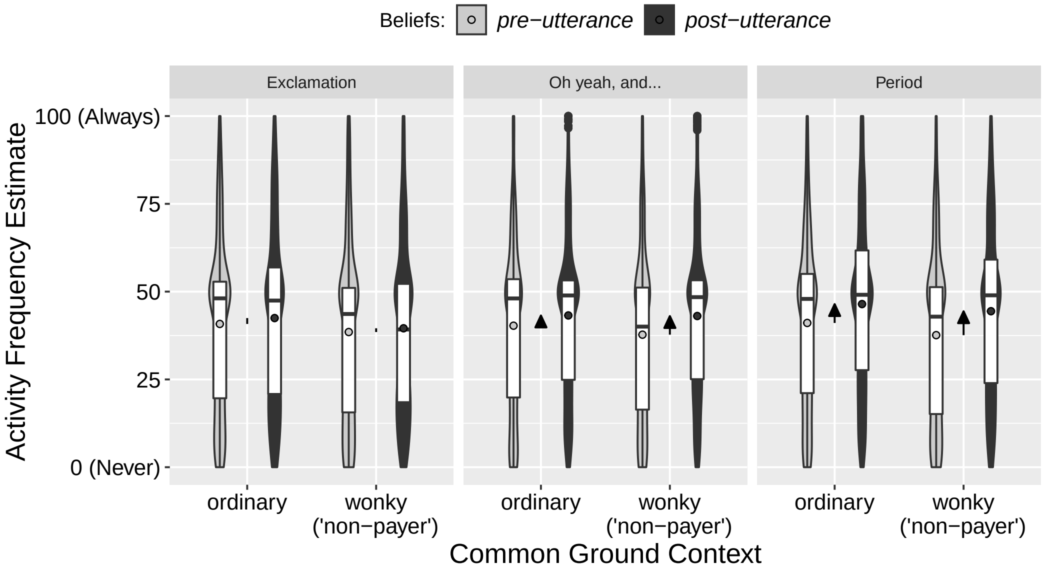

Figure 11 Experiment 1-3: *conventionally non-habitual* (*apple-buying*) activities analysis.

utterance / *post-utterance*), and by-item random intercepts and slopes for both factors and their interaction. By-subject random slopes for the interaction were not included in the model due to lack of within-subject repeated measures in our data for the interaction. The by-item random slope experiment was not included due to non-convergence.

The results are shown in Table 10 and Figure 12.

	β	SE(β)	t	p
Intercept	39.80	2.47	16.09	**<.001**
'!' vs. 'Oh yeah...'	2.25	1.01	2.22	**<.05**
'.' vs. Relevance Markers	2.33	1.02	2.28	**<.05**
Common Ground: Ordinary	2.98	2.01	1.49	0.2
Belief: Post-utterance	6.20	1.98	3.14	**<.01**
'!' vs. 'Oh yeah' * Common Ground	−0.02	1.40	−0.02	1
'.' vs. Relevance Markers * Common Ground	1.05	1.22	0.87	0.4
'!' vs. 'Oh yeah' * Belief	4.37	1.53	2.85	**<.01**
'.' vs. Relevance Markers * Belief	5.52	1.33	4.15	**<.001**
Common Ground * Belief	−4.65	1.14	−4.08	**<.001**
'!' vs. 'Oh yeah' * CG * Belief	−3.38	2.80	−1.21	0.2
'.' vs. Relevance Markers * CG * Belief	−4.28	2.43	−1.76	0.1

Table 10 Replicated Experiments 1-3: conventionally *non-habitual* (*apple-buying*) activities analysis.

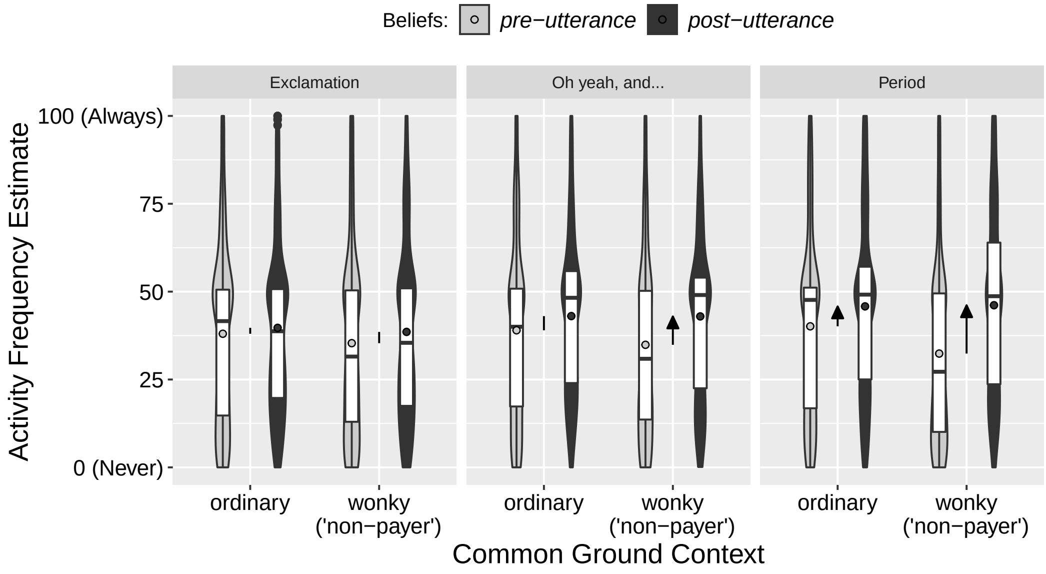

Figure 12 Replicated Experiments 1-3: *conventionally non-habitual* (*apple-buying*) activities analysis.

B Experimental Stimuli

	COMMON GROUND			UTTERANCE
1	John often goes to the grocery store around the corner from his apartment. ^{ordinary}	Recently, he came home from the store with groceries. When he came in, he saw his roommate Susan in the hallway, and started talking to her about his trip to the store. As he went to the kitchen to put his groceries away, Susan went to the living room, where their roommate Peter was watching TV.	Susan said to Peter:	“John just came back from the grocery store. He paid the cashier.” ^{habitual}
	John is typically broke, and doesn’t usually pay when he goes to the grocery store. ^{wonky}			“John just came back from the grocery store. He got some apples.” $^{non-habitual}$
2	Mary is a journalist who often goes to restaurants after her interviews. ^{ordinary}	Yesterday, she went to a popular Chinese place. As she was leaving, she ran into her friend David, and they started talking about the restaurant. After they parted, David continued on his way when he suddenly ran into Sally, a mutual friend of him and Mary.	David said to Sally:	“I ran into Mary leaving that Chinese place. She ate there.” ^{habitual}
	Mary is a journalist who often interviews restaurant waiters, but doesn’t like eating out. ^{wonky}			“I ran into Mary leaving that Chinese place. She got to see their kitchen.” $^{non-habitual}$
3	Jim lives in a shared apartment, where it’s his job to feed the dog in the evenings. ^{ordinary}	The other day he was feeding the dog some canned food, as his roommate Lucy came into the kitchen, and made herself a snack while chatting with him. Later in the evening, she settled down to watch TV alone with their roommate Carl.	Lucy said to Carl:	“Jim was feeding the dog earlier. He threw the can away.” ^{habitual}
	Jim lives in a shared apartment with his dog, but can’t seem to keep the common areas clean. ^{wonky}			“Jim was feeding the dog earlier. He had to add some medication to the food.” $^{non-habitual}$
4	Jane takes the subway all the time to get around the city. ^{ordinary}	Today she was entering a subway station when she ran into her friend Don, and they took the train together as they were heading in the same direction. Later that day, Don ran into Beth, Jane’s sister, on the street.	Don said to Beth:	“I took a train with Jane today. She bought a subway ticket.” ^{habitual}
	Jane takes the subway all the time to get around the city, but rarely pays for her rides. ^{wonky}			“I took a train with Jane today. She almost fell off the platform.” $^{non-habitual}$

	COMMON GROUND			UTTERANCE
5	Rick prefers to go to the local gas station to fuel up, though they overcharge him. *ordinary*	Yesterday he was fueling up there when he saw his friend Annie. They talked until he was finished, then said goodbye. After he was gone Annie went inside to buy some cigarettes, and ran into Sean, a mutual friend of theirs.	Annie said to Sean:	“Rick just fueled up here. He closed the fuel cap.” *habitual*
	Rick often doesn’t pay attention to what he’s doing when he fuels up at the local gas station. *wonky*			“Rick just fueled up here. He got some discounted gas.” *non−habitual*
6	Lisa likes to go swimming at a nearby pool after work. *ordinary*	A couple days ago she was at the pool when she saw Harvey, another regular member, and they stopped to chat. After Harvey changed and went out into the pool area, he ran into Jen, another swimmer and a friend of Lisa’s.	Harvey said to Jen:	“Lisa’s here to swim, too. She brought her swimsuit.” *habitual*
	Lisa likes to go swimming regularly, but often forgets to bring the things she needs. *wonky*			“Lisa’s here to swim, too. She came with her children.” *non−habitual*
7	Brian takes the train most mornings, though the commute takes a long time. *ordinary*	Last week when he was getting on the train, he ran into his old colleague Rachel, and they chatted until Brian got off. When Rachel got to work, she saw Oliver, who also used to work with Brian.	Rachel said to Oliver:	“I saw Brian on the train this morning. He got off at his stop.” *habitual*
	Brian takes the train most mornings, though he has a terrible habit of getting distracted and missing his stop. *wonky*			“I saw Brian on the train this morning. He was running late to work.” *non−habitual*
8	Laura works as a software engineer at a large company. *ordinary*	A couple of days ago she was getting ready to leave for work together with her husband Dustin. After they both left the house, he ran to catch his bus, and met up with Courtney, an acquaintance who took the same bus with him every day.	Dustin said to Courtney:	“Laura was just getting ready for work with me. She grabbed her house keys.” *habitual*
	Laura works as a software engineer at a large company, but is scatter-brained in everyday life. *wonky*			“Laura was just getting ready for work with me. She put on several layers of clothing.” *non−habitual*

	COMMON GROUND			UTTERANCE
9	Bruce goes to his local medical practice every few years. *ordinary*	Yesterday after leaving the practice he ran into his friend Sarah on the street, and they stopped to catch up. After they parted, Sarah walked on and soon saw Bruce's brother Drake on the street. She stopped to say Hi.	Sarah said to Drake:	"Bruce was just leaving the medical practice. He got examined by the doctor." *habitual*
	Bruce goes to his local medical practice every few years, but usually only sees the nurse. *wonky*			"Bruce was just leaving the medical practice. He was wearing a heart rate monitor." *non−habitual*
10	Olivia has beautiful hair, and pays a lot of attention to it. *ordinary*	Today, when she was leaving the bathroom after showering, she ran into her roommate and best friend Thomas. She talked to him briefly about her hair, as she tends to do. Later that day, when their housemate Jill came home, she and Thomas started talking about Olivia.	Thomas said to Jill:	"Olivia was talking to me about washing her hair. She used shampoo." *habitual*
	Olivia has beautiful hair, although she uses a cleansing conditioner only. *wonky*			"Olivia was talking to me about washing her hair. She found some split ends." *non−habitual*
11	Jared takes skydiving courses at the local airfield, when he has free time. *ordinary*	Last week he was at the skydiving center, with his friend Stella in the same group as him. They spent the day together, and when Stella went home in the evening, she texted Jared's brother Don, who was also a good friend of hers.	Stella said to Don:	"Jared was in the skydiving course today. He jumped out of the plane." *habitual*
	Jared takes skydiving courses at the local airfield, although he is still terrified of heights. *wonky*			"Jared was in the skydiving course today. He was the first to jump." *non−habitual*
12	Amy enjoys writing letters to people she is close to, especially around holidays. *ordinary*	About two days ago, she wrote a letter to her cousin Michelle, and today she talked about it with her brother Steve. In the evening, Steve got a call from Michelle, and they started talking about family.	Steve said to Michelle:	"Amy wrote you a letter. She mailed it." *habitual*
	Amy enjoys writing letters to people she is close to, but prefers to keep them to herself rather than mailing them. *wonky*			"Amy wrote you a letter. She used really expensive stationery." *non−habitual*

	COMMON GROUND			UTTERANCE
13	Adam usually takes the bus to work, as the stop is a few blocks from his house. *ordinary*	Last week, after he got off the bus, he ran into Virginia, his ex-girlfriend. They stopped for a little while to catch up.	Adam said to Virginia:	“I took the bus this morning. I walked to the bus stop.” *habitual*
	Adam usually takes the bus to work, but bikes to the stop although it’s only a few blocks from his house. *wonky*			“I took the bus this morning. I barely had room to stand.” *non−habitual*
14	Esther often goes along with her friends when they go clothes shopping, as it’s something she also enjoys. *ordinary*	Today, when she was walking out of a mall after spending time with her friends, she ran into George, another old friend of hers. They decided to catch up while walking to the bus stop.	Esther said to George:	“I was out clothes shopping. I tried something on.” *habitual*
	Esther often goes along with her friends when they go clothes shopping, although it bores her, and she just reads as they browse. *wonky*			“I was out clothes shopping. I came across a big sale.” *non−habitual*
15	Nick enjoys making pasta dishes for his roommates, as it’s an easy way to contribute to the household. *ordinary*	Yesterday he was preparing pasta in the kitchen, to sit in the fridge until a party tomorrow. When he was done and cleaning up, his roommate Clara came into the kitchen, and they started talking about his dish.	Nick said to Clara:	“I made some pasta for the meal. I boiled it in water.” *habitual*
	Nick enjoys making pasta dishes for his roommates, but prefers to bake fresh pasta that doesn’t need to be pre-boiled. *wonky*			“I made some pasta for the meal. I added some vegetables.” *non−habitual*
16	Grace enjoys baking, as it’s a great way to make new friends. *ordinary*	A few days ago she was baking a cake in her kitchen. After she had put it in the oven, her roommate Kyle came into the kitchen to make a salad for himself. They started chatting about food.	Grace said to Kyle:	“I’m baking a cake right now. I preheated the oven.” *habitual*
	Grace enjoys baking, although she’s terrible at following basic directions in recipes. *wonky*			“I’m baking a cake right now. I added chocolate chips to the recipe.” *non−habitual*

	COMMON GROUND			UTTERANCE
17	Greg frequently travels by air, to see family and attend conferences. *ordinary*	Last week he flew to a conference, and met up with Helen, an old colleague he occasionally traveled with. They went to breakfast together, and started talking about their travel.	Greg said to Helen:	"I flew here. I took my cell phone on board with me." *habitual*
	Greg frequently travels by air, but hates carrying things around with him, and checks absolutely everything into the hold. *wonky*			"I flew here. I got into business class." *non−habitual*
18	Sandy usually cuts her own hair, although she has no training. *ordinary*	Two days ago, after she gave herself another haircut, she went for a walk along her street. She quickly ran into her ex, Patrick, and they stopped to catch up for a few minutes.	Sandy said to Patrick:	"I just cut my hair. I used scissors." *habitual*
	Sandy usually cuts her own hair, simply by taking a buzzer to it. *wonky*			"I just cut my hair. I cut it a bit shorter than intended." *non−habitual*
19	Henry often goes to art exhibitions, as there's an art museum a short walk from his place. *ordinary*	Last week, after going to a new photography exhibition, he encountered his friend Max on his way home. They paused on the street and chatted for a while.	Henry said to Max:	"I just went to the new photo exhibit. I looked at the photographs." *habitual*
	Henry often goes to art exhibitions, but only because his girlfriend drags him. *wonky*			"I just went to the new photo exhibit. I decided to buy a photograph." *non−habitual*
20	Helen works hard at her job, and enjoys the challenges she's given at work. *ordinary*	Today, after driving her car to work as usual, she ran into her office-mate Peter while walking into the building. They stopped briefly to say hello.	Helen said to Peter:	"I just parked my car. I locked it." *habitual*
	Helen works hard at her job, although she is incredibly scatter-brained. *wonky*			"I just parked my car. One of my tail lights has gone out." *non−habitual*

	Common Ground			Utterance
21	Gary often orders pizza at work, from a famous pizzeria nearby. ^{ordinary}	A few days ago, after he placed an order, his colleague Stephanie walked over to his cubicle to chat.	Gary said to Stephanie:	"I just ordered pizza. I picked the toppings." ^{habitual}
	Gary often orders pizza at work, but doesn't usually get to choose which type of pizza to get. ^{wonky}			"I just ordered pizza. I used a gift certificate." $^{non-habitual}$
22	Julia always tries to wash the dishes after eating, to avoid annoying her roommates. ^{ordinary}	A few days ago, she was getting ready to go out after doing the dishes. She ran into her roommate Justin on her way out, and started talking to him.	Julia said to Justin:	"I just did the dishes. I rinsed them." ^{habitual}
	Julia always tries to wash the dishes after eating, but doesn't always bother to rinse them. ^{wonky}			"I just did the dishes. I polished them." $^{non-habitual}$
23	Emma often borrows books from the library, as she doesn't have much spare cash to spend. ^{ordinary}	Last week, after going to the library, she was heading home with several books, and ran into her best friend Tim on the street. They stopped to quickly say hello.	Emma said to Tim:	"I just got some books at the library. I checked them out." ^{habitual}
	Emma often steals books from the library, as she doesn't have money to buy her own copies. ^{wonky}			"I just got some books at the library. I looked at the library's exhibit." $^{non-habitual}$
24	Logan recently started doing his own laundry, after moving out of his parents' house. ^{ordinary}	Yesterday, after doing a load, he went to the living room to watch some TV. Soon his roommate Sophia came home, and asked about his day while taking off her coat.	Logan said to Sophia:	"I just did the laundry. I used detergent." ^{habitual}
	Logan recently started doing his own laundry, but can't get a handle even on the basics. ^{wonky}			"I just did the laundry. I added some softener to the wash." $^{non-habitual}$

C Power Analysis

```r
# Power by simulation for a normally distributed
# continuous outcome with subjects, items, and residual
# variability

# Population parameters mu: underlying mean of the
# outcome in the reference group betaN: effect size of
# predictor or interaction sdItem: sd of random effect
# at the item level sdSubject: sd of random effect at
# the subject level sdResid: sd of residual error

# Design parameters nSubjects: number of subjects in
# simulation nIterations: number of iterations in
# simulation

fnPower <- function(mu, beta1, beta2, beta3, beta4, beta5,
    beta6, beta7, beta8, beta9, beta10, beta11, sdItem,
    sdSubject, sdResid, nSubjects, nIterations, dots = TRUE) {
    start.time <- Sys.time()
    progress <- ") \n----|--- 1 ---|--- 2 ---|--- 3 ---|--- 4 ---| --- 5 \n"
    if (dots)
        cat("Simulations (", nIterations, progress, sep = "")
    # objects to store pvalue, beta, and standard error from
    # each iteration of simulation
    pVals <- betaVals <- seVals <- matrix(NA, nrow = nIterations,
        ncol = 11)
    # build design matrices
    m <- matrix(NA, nrow = nSubjects * 4, ncol = 7)
    colnames(m) <- c("worker", "exp.alike", "exp.diff",
        "story", "condition", "context", "slider")
    m[, 1] <- rep(1:nSubjects, each = 4)
    m[, 2] <- rep(c(-0.5, 0.5, 0), each = 4, length.out = length(m[,
        2]))
    m[, 3] <- rep(c(-0.3333333, -0.3333333, 0.6666667),
        each = 4, length.out = length(m[, 3]))
    i <- 1
    while (i < (length(m[, 4]))) {
        m[i:(i + 3), 4] <- sample(1:24, size = 4, replace = FALSE)
        i <- i + 4
    }
    m[, 5] <- rep(c(-0.5, 0.5))
```

```r
m[, 6] <- rep(c(-0.5, 0.5), each = 2)
# i <- 1 # (when testing for-loop)
for (i in 1:nIterations) {

    # draw random effects
    itemRE <- rnorm(24, 0, sdItem)
    subjRE <- rnorm(nSubjects, 0, sdSubject)
    residRE <- rnorm(nrow(m), 0, sdResid)

    # create outcome
    y <- mu + beta1 * m[, 2] + beta2 * m[, 3] + beta3 *
        m[, 6] + beta4 * m[, 5] + beta5 * m[, 2] * m[,
        6] + beta6 * m[, 3] * m[, 6] + beta7 * m[, 2] *
        m[, 5] + beta8 * m[, 3] * m[, 5] + beta9 * m[,
        6] * m[, 5] + beta10 * m[, 2] * m[, 6] * m[,
        5] + beta11 * m[, 3] * m[, 6] * m[, 5] + itemRE[m[,
        4]] + subjRE[m[, 1]] + residRE
    m[, 7] <- y
    dm <- as.data.frame(m)
    dm$worker <- as.factor(dm$worker)
    dm$story <- as.factor(dm$story)
    # fit model, store p-value, beta and standard error
    o <- lmer(slider ~ exp.alike + exp.diff + context +
        condition + exp.alike:context + exp.diff:context +
        exp.alike:condition + exp.diff:condition + context:condition +
        exp.alike:context:condition + exp.diff:context:condition +
        (1 | story) + (1 | worker), dm)
    pVals[i, ] <- coef(summary(o))[2:12, 5]
    betaVals[i, ] <- coef(summary(o))[2:12, 1]
    seVals[i, ] <- coef(summary(o))[2:12, 2]

    if (dots)
        cat(".", sep = "")
    if (dots && i%%50 == 0)
        cat(i, "\n")

}

if (dots)
    cat("\nSimulation Run Time:", round(difftime(Sys.time(),
        start.time, units = "hours"), 3), " Hours \n")
    # calculate power
    powerOut <- apply(pVals, 2, function(x) length(x[x <
        0.05])/length(x))
    return(list(power = powerOut, p = pVals, beta = betaVals,
        se = seVals))
}


# calibrate by setting betas = 0; histograms should all
# look level, with about 5% of results significant by
# chance
outCalibrate <- fnPower(mu = 61.0444, beta1 = 0, beta2 = 0,
    beta3 = 0, beta4 = 0, beta5 = 0, beta6 = 0, beta7 = 0,
    beta8 = 0, beta9 = 0, beta10 = 0, beta11 = 0, sdItem = 10.138,
    sdSubject = 9.285, sdResid = 21.839, nSubjects = 1200,
    nIterations = 10000, dots = TRUE)
outCalibrate$power
hist(outCalibrate$p[, 1], main = "beta1", xlab = "p value")
hist(outCalibrate$p[, 2], main = "beta2", xlab = "p value")
hist(outCalibrate$p[, 3], main = "beta3", xlab = "p value")
hist(outCalibrate$p[, 4], main = "beta4", xlab = "p value")
hist(outCalibrate$p[, 5], main = "beta5", xlab = "p value")
hist(outCalibrate$p[, 6], main = "beta6", xlab = "p value")
hist(outCalibrate$p[, 7], main = "beta7", xlab = "p value")
hist(outCalibrate$p[, 8], main = "beta8", xlab = "p value")
hist(outCalibrate$p[, 9], main = "beta9", xlab = "p value")
hist(outCalibrate$p[, 10], main = "beta10", xlab = "p value")
hist(outCalibrate$p[, 11], main = "beta11", xlab = "p value")

nSubjects <- seq(1200, 2400, 100)

st <- c()
for (i in 1:length(nSubjects)) {
    st[[paste("subj", nSubjects[i])]] <- fnPower(mu = 61.2216,
        beta1 = 1.3018, beta2 = 4.3355, beta3 = 38.0383,
        beta4 = -0.4559, beta5 = -0.8669, beta6 = 2.4416,
        beta7 = 0.6141, beta8 = 6.6776, beta9 = -12.6572,
        beta10 = 0.4207, beta11 = 6.1601, sdItem = 10.138,
        sdSubject = 9.285, sdResid = 21.839, nSubjects = nSubjects[i],
        nIterations = 1000, dots = TRUE)
}
```

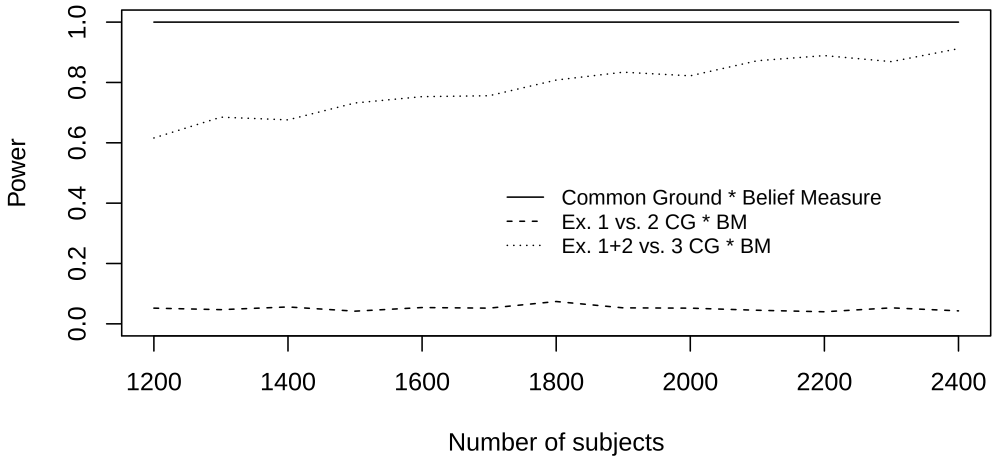

Figure 13 Power curves for effects of interest

C.1 Plot

Figure 13 shows a plot of the power curves for the critical common ground by belief measure interaction, a comparison of that effect across Experiments 1 and 2, and another comparison across Experiment 3 and Experiments 1/2 (as a group). We expect a robustly replicable common ground by belief measure interaction (power $= 1.00$ at all sample sizes), no significant difference between Experiments 1 and 2 (power ≤ 0.1 at all sample sizes), and a significant difference between Experiment 3 and Experiments 1/2 (power ≥ 0.85 at a sample size of 2100 or more).