An appeal for a methodological fusion of conversation analysis and experimental psychology

This item was submitted to Loughborough University's Institutional Repository by the/an author.


Additional Information:

- This is an Open Access Article. It is published by Taylor and Francis under the Creative Commons Attribution 4.0 Unported Licence (CC BY). Full details of this licence are available at: http://creativecommons.org/licenses/by/4.0/

Metadata Record: https://dspace.lboro.ac.uk/2134/37529

Version: Published

Publisher: © the Authors. Published by Taylor and Francis

Rights: This work is made available according to the conditions of the Creative Commons Attribution 4.0 International (CC BY 4.0) licence. Full details of this licence are available at: http://creativecommons.org/licenses/by/4.0/

Please cite the published version.
An Appeal for a Methodological Fusion of Conversation Analysis and Experimental Psychology

J. P. de Ruiter & Saul Albert

To cite this article: J. P. de Ruiter & Saul Albert (2017): An Appeal for a Methodological Fusion of Conversation Analysis and Experimental Psychology, Research on Language and Social Interaction, DOI: 10.1080/08351813.2017.1262050

To link to this article: http://dx.doi.org/10.1080/08351813.2017.1262050

Published with license by Taylor & Francis. © 2017 J. P. de Ruiter and Saul Albert.

Published online: 04 Jan 2017.
An Appeal for a Methodological Fusion of Conversation Analysis and Experimental Psychology

J. P. de Ruiter and Saul Albert

ABSTRACT

Human social interaction is studied by researchers in conversation analysis (CA) and psychology, but the dominant methodologies within these two disciplines are very different. Analyzing methodological differences in relation to major developments in the philosophy of science, we suggest that a central difference is that psychologists tend to follow Popper's falsificationism in dissociating the context of discovery and the context of justification. In CA, following Garfinkel's ethnomethodology, these two contexts are much closer to one another, if not inextricable. While this dissociation allows the psychologist a much larger theoretical freedom, because psychologists “only” need to validate their theories by generating confirmed predictions from experiments, it also carries the risk of generating theories that are less robust and pertinent to everyday interaction than the body of knowledge accumulated by CA. However, as long as key philosophical differences are well understood, it is not an inherently bad idea to generate predictions from theories and use quantitative and experimental methods to test them. It is both desirable and achievable to find a synthesis between methodologies that combines their strengths and avoids their weaknesses. We discuss a number of challenges that would need to be met and some opportunities that may arise from creating such a synthesis.

Two major scientific areas of study that focus on human social interaction are conversation analysis (henceforth CA) and experimental psychology (henceforth psychology). Both approaches ultimately aim to understand how people interact with one another by the orderly exchange of a variety of observable behaviors. But while the ultimate goals and empirical commitments of CA and interaction psychology are fairly similar, the basic definitions and methodologies they employ in studying human interaction are very different.

CA has been used in discursive psychology and ethnomethodology to redefine mind, cognition, and language—the basic phenomena that psychologists study—in terms of participants’ interactional achievements (Coulter, 1979, pp. 9–34; Potter & Te Molder, 2005). Similarly, core methods of coding and categorization often used by psychologists have been repositioned in CA as methods participants use themselves for drawing inferences and accomplishing “social action” in conversation (Sacks, 1995, p. 582; Schegloff, 2007). However, this critical respecification of psychology’s core definitions and
methods (Button, 1991) renders CA’s key findings about cognition less straightforwardly usable (or even recognizable as findings) within psychology and reduces the relevance of any specific focus on cognition for CA (Kitzinger, 2006). Suggestions as to how CA could present findings more compatibly with psychology (see, e.g., Stivers, 2015; Stokoe, 2012) have often involved discussions about the irreconcilable philosophical differences between CA and conventional approaches to coding, quantification, and statistical analysis (Couper-Kuhlen, 2010; Nishizaka, 2015; Schegloff, 1993; Steensig & Heinemann, 2015). In this article, however, we want to focus on the core aims and similarities shared by CA and psychology, to identify what they can learn from their differences. We strongly believe that these two approaches to the study of human interaction can and should find a practical synthesis that combines their strengths and avoids their weaknesses.

We will first discuss the main similarities between the two disciplines. We will then proceed to discuss a number of fundamental differences between CA and Psychology, viewed from a Philosophy of Science perspective. Not surprisingly, these differences have led to dissimilar scientific practices, which is what we discuss next. We then discuss how each discipline can benefit from adopting principles and practices from the other, followed by a list of concrete proposals and challenges for a fruitful methodological fusion.

**Important similarities between CA and psychology**

**Shared skepticism about introspection, and commitment to empirical data**

The most salient similarity between the methods used in these fields is that both CA and psychology reject introspection as a viable source of information about the adequacy of our theories. In CA, distrust of introspection is an explicit part of its methodology (see, e.g., Schegloff, 1996). Human interaction cannot be adequately analyzed and understood by attributing beliefs and desires to the participants and then applying folk-psychological reasoning about what happens in interaction. Instead, CA researchers analyze interaction strictly with reference to the observable behavior of the participants, using detailed transcripts to interrogate and describe the minutiae of naturalistic interaction. The CA research process is encapsulated by the question “why that now?” (Schegloff & Sacks, 1973), a question that researchers repeatedly ask themselves while looking at data—and which they also assume is relevant for participants at every moment during a conversation: “Why does the interactant display that particular behavior at that particular point in the sequence of mutual exchanges, and which social action is being coordinated with each exchange?” Anyone with even minimal exposure to CA will have noticed that this rigorous way of analyzing interaction is incompatible with colloquial explanations based on researchers’ own introspections or folk-psychological intuitions (Sacks, 1984, p. 25).

While some older traditions in psychology, most notably the one founded by Wilhelm Wundt, did rely on introspection of both researchers and experimental participants (Schultz & Schultz, 2004), the basic tenet of psychologists since the beginning of the 20th century has always been that they understand their subjects better than those subjects understand themselves. This was quite explicitly the case in Freudian psychology, where hidden sexual motives were assumed to be powerful forces controlling our mental lives, while our powerless and underinformed ego is blissfully unaware of this (Winch, 1990, pp. 47–48). Later, the doctrine of behaviorism tried to get rid of all “mentalistic” concepts (Costall, 2006), which was even more incompatible with introspection. But even after the so-called cognitive revolution, when mental representations became reputable again in computational incarnations, their existence, role, and nature tended to be demonstrated using reaction times and other behavioral measures (see, e.g., the famous mental rotation experiments by Shepard & Metzler, 1971), and experimental studies continue to use and extend this approach using neurophysiological recordings and neuroimaging data. There are still pleas for the

---

reintroduction of introspective data as a useful source of information in psychology (Kingstone, Smilek, & Eastwood, 2008) in order to relate these experimentally constrained measures more directly to the natural social actions they seek to emulate and explain. However, introspection has been—and still remains—a suspect source of information in psychology. While both CA and experimental psychology have a healthy distrust toward introspective accounts, and both are committed to a strong empirical orientation, they approach the task of gathering valid and compelling empirical evidence about human interaction very differently. To explain their differences in a way that emphasizes the possibilities of a practical synthesis will require a brief discussion of the methodological foundations of the two disciplines in relation to key developments in the philosophy of science.

**Shared skepticism about metaphysics and problems regarding induction**

Since the ancient Greeks, philosophers have struggled with the so-called demarcation problem: Which theory, statement, or claim is scientific, and which one is not (Popper in Miller, 1985). This is a deceptively difficult problem, and as far as we are aware, philosophers are still far from reaching a satisfactory consensus on the issue (Laudan, 1983; Resnik, 2000). One approach to solving this problem was verificationism, suggested around the beginning of the 20th century by the logical empiricists from the Vienna Circle led by influential scientists/philosophers like Moritz Schlick, Rudolph Carnap, and Otto Neurath (Creath, 2014). Their proposed solution was to only consider scientific any statement (or “sentence,” as they often called it) that was either based on direct experience (observation), or was a “synthetic sentence,” i.e., a sentence constructed from a combination of observables and logical terms. These sentences can then be verified and/or falsified. The main goal of this attempt to solve the demarcation problem was to distinguish scientific theories from metaphysics. Logical empiricism therefore revolves around the use of unobservable, theoretical concepts and asks in what ways and under which conditions the concepts thus stated are allowed to be part of a scientific theory.

The logical empiricists’ solution to the demarcation problem turns out to be unsatisfactory for a number of reasons. First, there is a logical problem. As Popper (2005 [1959]) pointed out, logical empiricism centrally relies on induction. From many observations of white swans, one infers the generalization that all swans are white. Early Empiricist philosopher David Hume’s original “problem of induction” is that one can never be sure. The observation of only one black swan would suffice to prove the generalization wrong. Another more practical problem with verificationism is that it is very tedious to specify all aspects of a theory all the way down to the level of observables and logical rules. This requirement leads to theoretical complexity and stifles the progress of science considerably.

**Differences between CA and psychology**

**Psychology adopted Popper’s solution of falsificationism**

In an attempt to circumvent these problems, Popper (1959/2005) suggested his famous solution of falsificationism. Due to the limitations of inference by induction, we can never be certain that a theory is true, only that it is false. So instead of relying on observational and synthetic sentences to “prove” a theory directly from the empirical data, we can just propose a theory, and (collectively) try to use empirical data to disprove (falsify) it. As long as a theory hasn’t been falsified, it stands, but we can never claim that it is true, only that it has not been proven false yet. Of course, this solution only works if the theory actually generates predictions that can be empirically tested. The demarcation problem is thereby solved as well: Any theory that can (in principle, potentially) be falsified by

---

4Hume (1738–40/1888, p. 89) did not actually use the word *induction* but suggested we should be skeptical about the way “instances of which we have had no experience resemble those of which we have had experience.”
empirical data is a scientific theory. This has some unexpected and often overlooked consequences. For instance, imagine an astrologer making the claim that people with the star sign Capricorn tend to be hard-working people. This is a testable prediction. For instance, we could check whether a sample of people with star sign Capricorn work longer hours than a random control sample of people with other star signs. So as long as astrologers are willing to make empirically testable predictions, astrology is, according to Popper’s demarcation criterion, a scientific theory. In Popper’s falsificationism any theory that can, at least in principle, be proven wrong by data is a scientific theory. This assumption liberated science because not every theory had to be meticulously reduced to observables and logical connectives. So when Einstein conceived of his general theory of relativity on the basis of a number of highly imaginative thought experiments, or when other theoretical physicists postulated the existence of new elementary particles that nobody had ever observed (see, e.g., Higgs, 1964), these were bona fide scientific theories because they generated empirically testable predictions (CMS Collaboration, 2012).

Here, we want to focus on an important assumption that Popper made for his meta-theory of science—namely, the separation of the context of discovery from the context of justification. The context of discovery is about the origin of our theory, how we arrived at it; whereas the context of justification is how we establish its empirical adequacy. In verificationism, these two are very similar, if not the same. But in falsificationism, they are completely independent. That is, it is irrelevant for the scientific-ness of a theory how someone comes up with it. It doesn’t matter whether a theory is discovered through thought experiments, divine revelation, hard work, induction, drug use, or having an apple falling on one’s head, as long as the predictions of the resulting theory can be empirically tested. It is in this respect that the methodological commitments of CA and psychology, despite their shared suspicions about introspection and reliance on empirical data, perhaps differ most crucially. Methods within psychology are heavily influenced by the doctrine of falsificationism; whereas CA inherently relies on inductive methods (Levinson, 1983).

CA adopted Hume’s skeptical solution of reflexive normativity

Popper’s falsificationism was a solution to Hume’s “problem of induction” with the aim of banishing metaphysics from causal claims in the natural sciences. While remaining skeptical about induction, Hume suggested it is inevitable that we interpret the world by drawing causal inferences based on our “customary” or normative perceptual experiences (Hume, 1748/1921). In other words, Hume assumed that induction must be inherent to human cognition (Kripke, 1982, p. 67). Hume developed this “skeptical solution” (Dauer, 1980) to problems of induction specifically in relation to human social action by suggesting how individuals within societies must determine their own moral codes. Hume (1748/1921) used an allegorical figure, which he called the “sensible knave,” to explain how people make autonomous moral choices even in the absence of an omnipotent God or a ubiquitous system of laws or social rules. They do so by internalizing social norms and “surveying ourselves as it were, in reflection” through the eyes of others. He called the knave “sensible” because when he makes knavish choices, and transgresses norms and laws, he feels his transgression as a moral “sense” and also because he is sensible enough not to behave so badly that he brings down the system of laws on which his exploits depend: “[a] sensible knave, in particular incidents, may think that an act of iniquity or infidelity will make a considerable addition to his fortune, without causing any considerable breach in the social union and confederacy” (Hume, 1777/2010, pp. 282–283). In human behavior, then, Hume outlined a self-interested, autonomous basis for morality in reflexive normativity (Korsgaard, 1994, pp. 55–61). The knave upholds social norms by bending, not “breaching,” them and thereby demonstrates his reliance on—and the inherent relevance of—the norms themselves. Like verificationism in the natural sciences, Hume’s reflexive normativity has informed sociological theories of “voluntaristic” social action (Parsons, 1937, p. 76) that appeal to normative conventions in order to avoid metaphysical explanations in the human sciences. A similar principle of reflexive normativity is used in speech act theory to explain how performative speech acts,
through repeated use, reflexively engender normative states of affairs (Austin, 1979, pp. 233–235), or how Gricean maxims function under the normative assumption of cooperativity (Grice, 1975).

**Garfinkel’s move from reflexive normativity to reflexive accountability**

However, Parsons’s student Garfinkel—like Popper—was still very skeptical about induction as a method and doubted that reflexive normativity could account for specific instances of social interaction. Garfinkel saw that this solution simply deferred problems of induction and turned them into questions of *interpretative relevance* (Schütz, 1962, p. 113), leaving open the problem of precisely which convention might be relevant in each specific situation and in what ways. This is especially problematic for scientific theories that separate the context of discovery from the context of justification and use “bridging assumptions” to link measures of reflexively normative tendencies with causal claims about human action. Garfinkel (1967, pp. 70–71) illustrated this criticism by describing a counterpart allegorical figure for Hume’s “sensible knave,” which he called the “judgmental dope.” The dope exhibits behaviors during experimental tasks that are set out for him in advance by the scientist’s own normative expectations about how such things *should* work. He is “judgmental” because when given a choice, he inevitably chooses between the alternatives decided for him in advance. He is a “dope”—meaning gullible or easily fooled—because he only ever makes these fixed choices. Like the tourist in a bazaar, when given the option to haggle, he always just pays the seller’s asking price (Lynch, 2012).

Garfinkel thought that if Hume’s sensible knave exhibits reflexively normative choices by bending the rules without a “breach in the social union and confederacy,” there are two problems. First, which norms and ostensible rules are relevant in specific situations? And second, where and at which point do they breach? To investigate the latter question, Garfinkel conducted a series of “breaching experiments,” which would horrify today’s ethics committees, that attempted to find the limits of the sensible knave’s norms. He asked his students to deliberately start behaving in transgressive ways and to observe the reactions of the public. They found that in practice, no reliable normative limits exist. Instead, people tend to reinterpret transgressions as intelligible—or “accountable”—within the situation at hand. Some transgressions even turned into new norms; when students were asked to haggle for fixed-price goods and were often—though not always—successful, many then adopted haggling as a new standard practice. Following Wittgenstein (1953), Garfinkel’s solution to the problem of induction was to suggest that there are no reflexively normative, relevant “rules” for human action but rather ongoing processes of reflexive mutual accountability. He showed that these problems of interpretative relevance are insoluble in theory and emphasized that scientists must pay close attention to each situation and how participants themselves solve these problems in practice. Instead of studying norms, Ethnomethodology studies the *methods* that participants develop and use to make sense of situated social action. CA, like most social sciences, still relies on reflexive normativity to explain, for example, how participants in conversation only optionally conform to normative patterns of turn taking, and how they exploit these patterns interactionally by deviating from them in an orderly manner (Sacks, Schegloff, & Jefferson, 1974, pp. 696–697, note 1). The key difference is that—as a branch of ethnomethodology focused on conversation (Lynch, 2000)—CA avoids relying on these norms to formulate causal explanations or theoretical claims about human action to be tested “outside” a situated domain of discovery. This means that CA is not inherently incompatible with psychological theories that postulate the relevance of mental events, as long as participants themselves treat such events as relevant, situated, interactional issues.

---

5 A bridging assumption suggests that normatively appropriate contextual factors can be reliably inferred to uphold the causal connections between an observable fact and its ostensible meanings (Matsui, 2000, pp. 93–94). Garfinkel warned that—if used inappropriately in studies of human action—these assumptions could allow problems of induction to creep back into the causal explanations supported by an otherwise rigorous experiment.
**Difference in scientific practices between CA and psychology**

The distinct solutions to problems of induction represented by falsificationism and ethnomethodology point to clear differences in how research is conducted in practice within psychology and CA, and how each has advantages and disadvantages. These differences have major implications (and also suggest incentives) for a methodological fusion. It is therefore useful to outline the major points of difference between their scientific practices.

**Psychology’s reliance on null hypothesis significance testing**

Research psychologists mostly try to follow the role model of the natural sciences, which is still heavily influenced by falsificationism. When psychologists formulate a new theory or hypothesis, they usually support it by deriving a prediction, for instance about a difference between a number of aggregated measurements, called an *effect*. Then, the psychologist will collect data (by performing an experiment or by analyzing existing data sources, e.g., corpora), and show that the predicted effect indeed exists. Formally, however, the psychologist posits a *null hypothesis*, corresponding with there being no such difference or effect. After collecting data, she uses inferential statistics to show that the predicted effect is indeed present and is probably not caused by random variation. This then allows the psychologist to *reject* the null hypothesis, meaning that the alternative hypothesis, which is actually what she wanted to prove, is accepted. Oddly, while this procedure superficially looks like the application of Popper’s falsificationism, and it may have been motivated by it, really it isn’t. The aim of the typical successful (i.e., publishable) psychological study is to show that the prediction derived from the proposed theory is true, which is claimed as support for the theory. So it is not the case, as strict falsificationism would require, that the aim is to show that one has tried hard but so far failed to falsify one’s own theory. This is clearly revealed by the fact that if the null hypothesis cannot be rejected (because the predicted effect was not significant), the results are usually not even published because null effects are deemed uninformative. So the result that would actually count as a falsification (not being able to reject the null hypothesis) is usually not published and seen as an uninteresting “failure.” It is not deemed particularly interesting news that someone has come up with a theory that produced a prediction that hasn’t been empirically confirmed. When null effects representing unconfirmed predictions are published in the peer-reviewed literature, it is usually in the context of an explicit attempt at replicating a previously reported finding (e.g., Doyen, Klein, Pichon, & Cleeremans, 2012; Open Science Collaboration, 2015). In fact, null hypothesis significance testing (NHST—the standard inferential-statistical paradigm) does not even *allow* for accepting a null hypothesis. All it allows for is stating that the null hypothesis has not been rejected, but NHST can never tell us whether this was due to an insufficient amount of data (too much noise to be able to tell) or to the null hypothesis probably being true. So, intriguingly, neither the publication practices nor the standard statistical paradigm in psychology allow for direct falsification of a proposed theory.

While the logic of supporting theories by rejecting null hypotheses in psychology is not strictly falsificationist, the separation of context of discovery and context of justification is, and this distinction has been enthusiastically embraced in psychology. In principle, as long as it predicts an effect that is subsequently shown to be statistically significant, a psychologist can come up with literally any hypothesis. In practice, the research psychologist is often inspired by theories in other fields (see, e.g., Pickering & Garrod, 2013) or lacunae in the existing literature. The social psychologist Daryl Bem went a step further and explicitly recommended searching in already collected data for significant differences that suggest new hypotheses (Bem, 2003). But this practice is controversial, as it uses the same data for the generation as well as the testing of hypotheses (a practice Kriegeskorte, Simmons, Bellgowan, & Baker, 2009, call “double dipping”) and fails to make the methodologically important distinction between exploratory and confirmatory research (De Groot et al., 2014; Wagenmakers, Wetzels, Borsboom, & Van Der Maas, 2011).
The statistical testing of predictions from hypotheses is connected to another way that methods in psychology differ from CA: the abundant use of summary statistics, such as means and medians. Using aggregate values (e.g., frequency counts or proportions) in the study of human interaction has been sharply criticized by Schegloff (1993). In Schegloff’s own compact formulation: “It seems quite clear to me that parties to interaction do not laugh per minute” (p. 104). Another problem with using aggregate values is that they may inadvertently conceal the role of hidden but important variables, which, when taken into account, may even reverse the direction of a relation between two variables: a problem known as Simpson’s paradox (Simpson, 1951).

**CA’s reliance on naturalistic data**

This approach is fundamentally different from methods in CA, which explicitly avoid invoking entities or concepts that are not firmly grounded in natural observation. In the words of Heritage (1984, p. 236):

[CA’s] insistence on the use of data collected from naturally occurring occasions of everyday interaction is paralleled by a corresponding avoidance of a range of other research methodologies as unsatisfactory sources of data. These include: (1) the use of interviewing techniques in which the verbal formulations of subjects are treated as an appropriate substitute for the observation of actual behavior; (2) the use of observational methods in which data are recorded through field notes or with pre-coded schedules; (3) the use of native intuitions as a means of inventing examples of interactional behavior; and (4) the use of experimental methodologies involving the direction or manipulation of behavior.

Here, Heritage comprehensively sums up the types of data that linguists, psycholinguists, and cognitive psychologists tend to use to confirm the predictions generated from their theories. This is a major difference in approach between psychology and CA, and we want to suggest here that the reason psychologists are not worried about using these types of data is partially due to the historical influence of falsificationism, which tries to “decouple” the context of discovery from the context of justification.

However, while CA has maintained its methodological focus, discursive psychology (Potter & Te Molder, 2005; Tileagă & Stokoe, 2015) and ethnomethodological studies of work and “institutional talk” (Drew & Heritage, 1992; Heath & Luff, 2000) have also used CA successfully in more formal and constrained interactional settings. Recordings of interaction can be informative materials for CA as long as they are treated as evidence of how participants themselves coproduce each setting as interactionally relevant (Sacks, 1995, pp. 12–13). As CA’s contexts of study have become increasingly diverse, multimodal (Nevile, 2015), and mobile, it has relied on its core methods to remain resolutely “qualitative, inductive, and strictly empirical” (Haddington, Mondada, & Nevile, 2013, p. 7) and essentially tried to minimize the separation of the context of discovery from the context of justification as much as possible. Analysts begin by noticing recurrent patterns or skewed distributions of some candidate phenomena while reviewing and transcribing recordings of interaction, then interrogating their possible interactional uses during presentation and discussion at regular CA “data sessions” (Harris, Theobald, Danby, Reynolds, & Rintel, 2012). Many iterations of this process produce detailed “single-case analyses” until sufficiently large collections of standard cases and various “deviant cases” (that explicate participants’ own orderly management of nonstandard cases) can provide generic criteria for a clear characterization of the phenomenon (Schegloff, 1996). All this careful, qualitative work is necessary before researchers can even begin the process of developing a formal coding schema for a quantitative study (Stivers & Enfield, 2010)—and this last step is still far from being universally accepted within CA (Stivers, 2015). This process has the reassuring advantage that the phenomena described are guaranteed to have actually occurred in reality, not only in our theoretical imagination, and—ideally—to have been subjected to various

---

6Schegloff (1968) famously describes collecting up to 500 cases of telephone call openings, and his key finding about the sequential order of ringing and greeting exchanges hinges on the detailed explication of a single “deviant case.”
forms of critical “peer review” early and often. CA’s ethnomethodological research tradition does not attempt to solve problems of induction but goes to great lengths to mitigate them. This procedure probably underpins the fact that the body of knowledge accumulated by half a century of CA research is impressively robust: Many of the central phenomena that have been discovered and reported by CA researchers have later been confirmed using a variety of quantitative and/or experimental methods (see, e.g., Bögels, Kendrick, & Levinson, 2015; De Ruiter, Mitterer, & Enfield, 2006; Enfield et al., 2013; Kendrick & Torreira, 2015; Stivers et al., 2009).

**How CA and psychology can improve each other**

**CA can help with psychology’s replication crisis**

Psychology, on the other hand, especially social psychology, is going through a serious “replication crisis” (Pashler & Harris, 2012; Pashler & Wagenmakers, 2012): It has recently been discovered that many of the published findings of experimental studies turn out not to be reproducible. This not only holds for a few selected, central, and widely cited findings, but also for the majority of 100 randomly selected effects published in three authoritative psychology journals. This has stirred a heated debate about the role, necessity, and desirability of replication in psychology. It has led Nobel laureate Daniel Kahneman, in an open letter published in the journal *Nature*, to urge social psychologists to “do something about this mess.”

We suggest that one largely underestimated cause of the replication crisis in psychology is the wide conceptual gap between the context of discovery and the context of justification. The often large number of reasoning steps between a psychological theory or hypothesis and the prediction that is derived from it means that many “bridging assumptions” have to be made to arrive at a testable prediction. And the more assumptions that have to be made, the more alternative explanations (that are not forcibly required by the tested theory) for the findings are possible. Wagenmakers (2007) suggests that another problem contributing to the replication crisis is that NHST, the statistical paradigm that is overwhelmingly used by psychologists, tends to overestimate the evidence against the null hypothesis and also that the conventionally accepted significance threshold of 5% is probably putting the bar for rejecting our null hypotheses too low (Johnson, 2013; Sellke, Bayarri, & Berger, 2001; Wetzels et al., 2011). In other words, with our standard statistical conventions, it is “too easy” to find and publish significant effects that are then accepted by the scientific community as reliable evidence for theories. Alternative paradigms such as Bayesian statistics (see, e.g., Dienes, 2011; Wagenmakers, 2007) in combination with a higher threshold for the minimal amount of required evidence (e.g., Etz & Vandekerckhove, 2016) would largely remedy this problem but are still controversial because of the reluctance of many researchers to specify their preexperimental uncertainty in the form of so-called prior distributions, which is mathematically required to estimate the probability that our hypotheses are true, given our collected data (Jaynes, 2003; Wagenmakers, 2007).

The robustness of the knowledge acquired by CA mentioned previously could inspire psychologists studying interaction (and even those that do not—but that is an interesting topic that is beyond the present scope) to stay conceptually closer to our actual social behavior “in the wild.” To illustrate how important it can be not to rely solely on results from laboratory experiments, Branigan, Pickering, and Cleland (2000) found that in a controlled experiment, people tend to copy the syntactic structure of their interlocutor’s previous utterances more often than pure chance would predict. This is one of the central findings often cited as providing support for Pickering and Garrod’s (2004) theory that our dialogue behavior is primarily driven by resource-cheap automatic priming processes. However, extensive corpus studies by Healey et al. (Healey, Howes, & Purver, 2010; Healey, Purver, & Howes, 2014) demonstrated that in natural dialogues, people copy the

---

1Kahneman’s open letter can be found at [http://www.nature.com/polopoly_fs/7.6716.1349271308!/suppinfoFile/Kahneman%20Letter.pdf](http://www.nature.com/polopoly_fs/7.6716.1349271308!/suppinfoFile/Kahneman%20Letter.pdf).

2This is a problem narrowly related to the famous “Quine-Duhem Thesis” (see Harding, 2012).
syntactic structure of utterances in their interlocutor’s previous contribution less often than chance would predict. Also, an analysis by Schegloff (1997) revealed that when repetitions occur in conversation, they are not produced automatically but instead perform a range of specific social actions, mostly related to conversational repair. So it is overly optimistic to assume that effects found under controlled laboratory conditions provide sufficient support for theories that explain behavior outside of the lab in our real lives as social agents.

Specifically, it is important for psychologists to be aware of the limitations inherent in using the types of data mentioned in the earlier quote by Heritage, especially data that are either generated by or filtered through the preempirical introspection and intuitions of the researcher. This is especially important for the study of interaction because interactive behaviors are not (only) symptoms of some cognitive process but above all actions designed for a recipient (Melden, 1961). If we have evidence that average reaction times or pupil dilation measurements are systematically related to mental processing load, we can at least infer something about the average processing load during that particular mental process. However, if we average laughter rates or mutual gaze, these aggregated values may have little relevance for the individual interaction events on which they are based because the interpretation of these signals is primarily done by the interactants themselves and not by the measurements of the “objective” researcher. In other words, we also need to know what the behaviors under study are doing. A common defense of the use of aggregate values is that even when not every individual data point is “valid” in the sense outlined previously, they still give reliable information about tendencies. But it still heightens the risk of hidden but relevant variables changing or even reversing findings, as is most convincingly demonstrated by Simpson’s paradox (Blyth, 1972; Simpson, 1951).

Also, in order to achieve experimental control, many interaction studies rely on tasks that are invented by researchers, which causes the participants to communicate in unusual ways with strangers or preinstructed confederates. These types of task may in fact be closer to an exercise in role-playing or amateur theater than to a socially relevant real-life interaction (see also De Ruiter, 2013; Stokoe, 2014).

**Psychology can help clarify CA’s epistemological claims**

Following Garfinkel’s (1967) and Schegloff’s (1993) critiques of experimentation and quantification in the social sciences, CA researchers have studiously avoided using these methods. However, although CA’s “theoretical asceticism” (Levinson, 1983) has allowed it to avoid many of the pitfalls that psychology seems to risk stumbling into, abstinence from falsificationism has come at a price. Firstly, there is a lack of clarity about the role of falsificationism within the CA literature. For example, Heritage notes the risk that CA’s focus on the reflexive accountability of all action within social interaction may “produce generalizations that become unfalsifiable and, hence, nonempirical” (Heritage, 2012a). Indeed, CA research into epistemics (which deals with ostensible states of mutual knowledge as ubiquitously relevant interactional issues) has been criticized for straying too far from CA’s ethnomethodological roots in this direction (Lindwall, Lymer, & Ivarsson, 2016). However, Heritage’s foundational descriptions of “epistemics in action” (2012b, 2012c) do not in fact formulate hypothetical claims or attempt to falsify them. CA has, nonetheless, provided grist for the mill of experimental interaction studies. For example, Bögels, Kendrick, and Levinson (2015), who do formulate and test hypotheses, operationalize the phenomenon of preference organization in conversation. A clear understanding of the epistemological distinctions between falsificationism and ethnomethodology could allow researchers to use these approaches in concert rather than “throwing the baby out with the bathwater,” as Steensig and Heinemann (2016) put it. There are also many other research methods and sources of data in psychology that could open new avenues of research within CA without requiring researchers to abandon their ethnomethodological principles. The fact that preempirical conceptualizations can and often do lead to “shallow” theories that are not firmly grounded in social reality doesn’t mean that deriving and testing predictions from theories is in itself
a bad idea. It would, in our view, be very fruitful if CA researchers took some of the patterns of action and social practices they have discovered and generate predictions that could subsequently be tested cumulatively, in ways that contribute to falsifying theories in other fields including, of course, psychology.

More broadly, the issue of generalization—or as psychologists would say, “external validity”—is a divisive epistemological issue within CA. It is of great value to know to what extent a practice or phenomenon found in one language community or on one situated occasion of use might or might not be the same in another. CA’s generalizations have tended to focus on the systematic conventionalization and grammaticalization of recurrent patterns of talk. However, video-based studies of embodied interaction tend instead to emphasize the evident importance of specific, situated occasions of interaction (Mondada, 2016). The “embodied turn” in interaction research (Nevile, 2015) toward exploring visible, copresent social action is productively challenging many of CA’s foundational generalizations, which were mostly derived from analyses of talk on the phone (Couper-Kuhlen, 2010). We argue that experimental methods for predicting and testing generalizations could offer a similarly productive experimental counterpart to the challenges of analyzing embodiment, which would benefit CA by highlighting findings that have been, for whatever reason, inaccurate or inadequately specified. One of the most convincing ways to argue that a CA finding is not reliable is by taking the additional (if epistemologically distinct) step of formulating it into a falsifiable hypothesis, testing it, then trying to replicate the analysis as closely as possible with new data. A good example of such an enterprise is the study by Stivers et al. (2009), which provided evidence that the uncanny ability of conversationalists to anticipate the end of their interlocutor’s turns (De Ruiter et al., 2006; Sacks et al., 1974; Wilson & Zimmerman, 1986) is a universal property of human communication.

We appeal to CA researchers to engage with more conventional experimental methods in the human sciences and to share CA’s critical awareness of the distinctions between coding and quantifying more reflexively normative practices on the one hand and the pitfalls of attempting to code more reflexively accountable practices on the other. For example, Stivers (2015) points out that coding turn-initial utterances as more or less “response-relevant” social actions (Stivers & Rossano, 2010) is potentially an inappropriate use of coding and quantification because when analysts code an action as either initial or not, the analyst’s binary decision eclipses the subtle ways participants themselves may maintain a degree of uncertainty about whether an action requires a response (Couper-Kuhlen, 2010). Similarly, the ascription of analyst-oriented “action-types” (Levinson, 2012) to particular utterances may be relatively secure when coding highly routinized patterns of social action such as “requests for information” designed as reflexively normative “first order actions” in talk on the phone (Heritage, 2012c). However, coding may be less appropriate in copresent settings, where participants may manage the potential consequences of doing unequivocal assessments, noticings, or announcements using embodied actions or by producing them as “defeasibly” equivocal social actions (Sidnell, 2012). By combining CA and experimental approaches, we suggest that researchers can better understand and take into account the limits of what kinds of phenomena and what kinds of social settings are formally analyzable.

**Proposals and challenges for a methodological fusion**

For a methodologically coherent fusion of psychology and CA in the study of human interaction, a number of challenges will have to be addressed. We outline five proposals and challenges, without claiming to be comprehensive. We do not aim to supplant one method with another or to cast either as primary. A fusion of CA and psychology should combine their practical methods and respect their distinct epistemological constraints.
CA should be the starting point for experiments in human interaction

In order to be able to do controlled experiments, which is seen as the gold standard of providing empirical support for causal explanations (see De Ruiter, 2013, for a more detailed discussion) we have to find acceptable compromises between internal validity (e.g., control of independent and potentially confounding variables, proper operationalizations, and accurate measurement, etc.) on the one hand and ecological validity on the other. Many interaction psychologists, one of the present authors included (see Newman-Norlund et al., 2009; Willems et al., 2010), have been dangerously optimistic about the consequences of letting experimental participants perform artificial interactive tasks, of having them interact with confederates whose behavior we try to control, or even having them believe interlocutors are interacting with them while the behavior of these presumed interlocutors is in fact generated by automated scripts. Even if such studies result in statistically significant (and therefore potentially publishable) findings, that does not automatically mean that the results are valid. The findings could be significant and irrelevant or misleading at the same time. We therefore propose that experimental hypotheses are formulated based on CA studies using maximally detailed recordings of natural interaction.

Experimental situations should also be subjected to interactional analysis

A phenomenon observed in natural interaction may be impossible to recreate as such in a laboratory. Also, in terms of what is interactionally relevant in any given situation, there is no intrinsic distinction between “natural” and “contrived” data (Speer, 2002). We therefore propose combining validation studies that explore the limits and consequences of artificial scenarios, role-playing, and the use of confederates (see, e.g., Kuhlen & Brennan, 2013) with CA studies of experimental scenarios as interactional occasions. For example, CA studies of experiments based on introspective self-report have provided valuable insight into the pragmatics and accountability of introspection (Wooffitt & Holt, 2011). There is also a rich tradition of CA research into the interactional dynamics and constraints of standardized survey procedures (Antaki, 2000; Drew, Toerien, Irvine, & Sainsbury, 2014; Maynard, Houtkoop-Steenstra, Schaeffer, & Van Der Zouwen, 2002; Maynard & Schaeffer, 2006). Doing a CA study of an experimental setting in conjunction with a replication could simultaneously pursue complementary if epistemologically distinct research questions. Researchers could then analyze and compare target phenomena in both natural and experimental settings. The basic idea of combining naturalistic observation and experimental measures is already well established within psychology. For example, Doob and Gross’s (1968) famous car-honking study demonstrated that using questionnaires to ask people how they think they would behave in a certain situation does not necessarily yield the same results as observing people’s actual behavior in that situation. Similarly, Schober and Conrad (2012) demonstrate that measuring gaze aversion and disfluencies in survey interviews provides valuable information about the reliability of participants’ answers. Surveys are used almost ubiquitously in psychology to make bridging assumptions between behavioral or neurocognitive measures, analysts’ theories and subjects’ interpretations. Meanwhile, there is already a wealth of CA research into survey taking. We therefore propose a combined approach to explore whether—and under which conditions—experiments, surveys, and CA studies of these formally routinized interactional settings suggest convergent or divergent causal explanations.9

Quantification of CA-derived phenomena should be reported systematically

A third synthesis that would be required is a set of widely accepted guidelines regarding proper quantification. Quantification is very useful to communicate the degree of “pervasiveness” of certain phenomena within a corpus to other scientists and can be used to quantify and report uncertainty

9Potter and Hepburn (2012) outline a similar approach to using CA methods to study interview settings as a way of interactionally grounding the research questions of survey studies common in qualitative sociology and social psychology.
about specific phenomena or explanations. We cannot always expect 100% certainty in our theories, as no theory is ever perfect. But if 98% of all swans are white, assuming that swans are white is a much better heuristic than if only 60% of all swans are white, which makes reporting the estimated percentages very useful in guiding further research in the relevant branch of ornithology. The classic CA papers that Stivers (2015, p. 6) cites as exemplary of how CA studies tend to use vague distributional descriptors such as “massively,” “overwhelmingly,” and “regularly” could easily replace those terms with accurate but clearly informal “counts.” Gail Jefferson (1988) sets a precedent when she describes searching through masses of transcribed conversational data in the early stages of identifying a candidate CA phenomenon before starting to work on a more qualified analytic characterization of it. Today this practice is not unusual in CA studies of particular linguistic phenomena (see, e.g., Mandelbaum, 2014; Thompson, Fox, & Couper-Kuhlen, 2015). This kind of informal counting does nothing to diminish the value of CA findings. If anything, exposing this process would clarify and distinguish CA’s use of inferential heuristics from the more formal quantifications required by preexperimental CA-based coding schemes (Dingemanse, Kendrick, Enfield, & Heritage, 2016; Enfield, Stivers, & Levinson, 2010). As argued in De Ruiter (2013), the problems that Schegloff (1993) identifies with quantification are nontrivial and worrying, but Schegloff’s analysis itself suggests a possible solution for some of these problems. Indeed, as Schegloff argues, if we count occurrences of laughter or continuers, we would have to count them as a proportion of the number of opportunities in which laughter or continuers could have occurred. But knowing that, we can proceed to actually identify and count those opportunities and by doing so allow our quantifications to become much more informative. So instead of adopting either CA’s tendency to simply avoid doing or even discussing formal quantification, or interaction psychology’s tendencies to count whatever phenomena we imagine could be informative and then perform statistical tests on them as a kind of scientific ritual (Gigerenzer, 2004), it would be much better to work toward an approach to quantification that addresses the valid concerns of both psychologists and conversation analysts.

Our questions should be alert to, but not limited by, our methodological choices

We have to be careful not to adapt our research questions too much to the limitations of one methodology. For example, Schegloff (1987) finds two general types of problems that lead to conversational repair: One is problematic reference, which Schegloff calls “relatively straightforward,” and the other has to do with sequential implicativeness, which he qualifies as “interesting” (p. 204). However, most, if not all, psycholinguistic studies on conversational repair using dialogue tasks are about problematic reference (E. A. Schegloff, personal communication, August 3, 2010). This is because it is the only type of repair situation that we can easily invoke in a laboratory with experimental participants performing a dialogue task (e.g., Anderson et al., 1991). It would not be a good idea to only study repair with problematic reference and then assume that our findings generalize to all types of repair. We therefore propose to systematize some ways to discover more “interesting” and hard to operationalize interactional phenomena for psycholinguistic study using CA. For example, CA’s long-term engagement with detailed qualitative analysis of interaction in medical care has revealed the highly routinized structure of doctor-patient consultations (Heath, 1990; Heritage & Maynard, 2006). The normative constraints on interaction in many institutional settings (Drew & Sorjonen, 1997) include obvious issues such as that the consultation lasts a limited time, and both doctors and patients therefore have to manage these constraints interactionally. CA researchers’ familiarity with normative patterns of interaction in these settings can help to formulate situated research questions, to define appropriate behavioral measures, and to choose experimental variables that would be unimaginable through introspection—even by experts. In this vein, a subtle but powerful experiment by Heritage, Robinson, Eliott, Beckett, and Wilkes (2007) demonstrates how exploiting minor differences in the preference structure associated with the single word any or some when the doctor asks “Do you have any/some other concerns?” at the end of a consultation can
vastly reduce the number of unmet concerns patients report afterwards. The experimental data, statistical analyses, and CA findings from the combined approach proposed here will not triangulate on the same empirical questions, the same kinds of causal explanations, or build on the same epistemological foundations. Nonetheless, CA researchers would clearly benefit from being able to exploit the opportunities opened up by this proposed fusion for developing—as Steensig and Heinemann (2015) suggest—“new angles and new qualitative studies.” Similarly, psychologists could find unique opportunities within these studies to develop new and elegant ways to operationalize Schegloff’s more “interesting” interactional phenomena.

**Conversation analysts and cognitive psychologists need to establish theoretical interfaces**

It is urgently necessary that we develop “theoretical interfaces” between CA and cognitive psychology, enabling us to theorize about the interlocking of cognitive faculties with social behavior. On the one hand, theories of cognition depend on interaction as the most evidently and fundamentally cognitive activity. It is not a coincidence that the high bar represented by the Turing Test for Artificial Intelligence is based on having a convincing informal conversation with a machine. On the other hand, interaction itself is not possible without the vast neurocognitive architecture supporting it. Given that we are a highly social species, it is plausible that our cognitive faculties have been strongly shaped by adaptations that meet the complex demands of interaction (Levinson, 1995). It makes sense, therefore, that we should combine approaches designed to investigate the human faculties shaped by evolution with approaches that ask how we use those faculties as resources for social action. For example, De Ruiter and Cummins (2012) propose a computational model that combines interactional resources (utterances, gestures, etc.) and information about their conventional uses to generate testable predictions about mental processes of action recognition (see Cummins & De Ruiter, 2014 for a discussion of other computational approaches to the same problem). By extension, CA studies of mechanisms such as repair can provide insights into the practical procedures of action recognition that work as naturalistic constraints on psychological theorizing (Schegloff, 2006).

**Concluding remarks**

Luckily, there are many signs that multidisciplinary fusions of CA and psychology are on the rise. The recent interest in the cognitive mechanisms underlying predictive turn taking (De Ruiter et al., 2006; Keitel, Prinz, Friederici, Von Hofsten, & Daum, 2013; Levinson, 2016) and the use of quantitative methods to find universal norms of human interaction (Dingemanse, Torreira, & Enfield, 2013; Stivers et al., 2009) are all examples of a fruitful synthesis of the empirical rigor of CA and the methodological flexibility of psychology.

Looking critically at results from other fields through one’s own methodological glasses, as we tend to do instinctively, is hardly going to facilitate collaboration between different disciplines. For developing effective multidisciplinary collaborations, it is much more fruitful to look critically at one’s own results from the methodological perspective of others.

**Funding**

Saul Albert’s work was funded by the Engineering and Physical Sciences Research Council (EPSRC) through the Media and Arts Technology Programme, a Research Councils UK Centre for Doctoral Training (EP/G03723X/1).

**ORCID**

J. P. de Ruiter http://orcid.org/0000-0003-2897-8804
Saul Albert http://orcid.org/0000-0003-1043-1237
References


