EXPERIMENTS WITH INTERACTIONAL EXPERTISE.

HARRY COLLINS, ROB EVANS, RODRIGO RIBEIRO, AND MLL

DRAFT FOR SUBMISSION (FIGURES AND TABLES ARE AT THE END OF THE PAPER)

ABSTRACT

'Interactional expertise' is developed through linguistic interaction without full scale practical immersion in a culture. Interactional expertise is the medium of communication in peer review in science, in review committees, in interdisciplinary projects, and in public understanding of science. It is also the medium of specialist journalists and interpretative methods in the social sciences. We describe imitation game experiments designed to make concrete the idea of interactional expertise. The experiments show that the linguistic performance of those well-socialized in the language of a specialist group is indistinguishable from those with full blown practical socialization but distinguishable from those who are not well-socialized. The imitation game can also be used to indicate whether an individual can enter an esoteric domain and master the interactional expertise, a skill required by interpretative sociologists of science, anthropologists, ethnographers, and the like.
INTRODUCTION: THE IDEA OF INTERACTIONAL EXPERTISE

The experiments described in this paper are one approach to `studies of expertise and experience' (SEE -- Collins and Evans, 2002). Here we look at `interactional expertise.' The idea of interactional expertise is readily exemplified by the skill required by a sociologist of scientific knowledge who learns to engage respondents in technical discussions of a scientific domain without being able to publish or carry out experiments; the sociologist develops interactional expertise without developing `contributory expertise;' the ability to contribute to the science.2

1 Rodrigo Ribeiro acted as research assistant and made many valuable contributions toward the understanding of the meaning and form of the protocols. Martin Hall (Southampton University), developed the software for the experiment.

2 The contrast between interactional and contributory expertise is first discussed in Collins and Evans 2002. For a philosophical discussion of interactional expertise which relates it to existing philosophical ideas and is illustrated with examples from fiction and psychology see Collins 2004a; for a discussion of interactional versus contributory expertise in the context of sociological research on scientists, see Collins 2004b, 2004c; for types of expertise in the context of relations between scientists and the public, see Epstein, 1996, Collins and Evans 2002, and Collins and Pinch 2005. For a more complete classification of expertises see `The Periodic Table of Expertises' and the exposition at www.cf.ac.uk/sosci/expertise.
Interactional expertise should not be confused with the kind of formal knowledge that can be written down. Interactional expertise involves mastery of a language and is characterized by rules that cannot be explicated; it is a tacit knowledge-laden ability that can only be acquired through immersion in the discourse of the community. Interactional expertise stands between the idea of knowledge supported by, say, proponents of artificial intelligence, who think that all knowledge is potentially encodable, and that of phenomenological philosophers, who think that practice is the essence of all understanding. The latter would probably agree, however, that the acquisition of interactional expertise is often a stage in the process of more practical socialization. 

The uses of interactional expertise go well beyond the sociology of scientific knowledge. For example, all kinds of participant observers, including ethnographers and social anthropologists, at least those who are trying to understand the world of their subjects rather than merely to observe it, must try to acquire it. Attainment of interactional expertise is also the goal of specialist journalists. Interactional expertise is also, at least arguably, the

---

3 The locus classicus of the debate between the artificial intelligence community and the phenomenologists is Dreyfus, 1972 and 1992; see also Suchman 1987, Collins, 1990, 1998. Sometimes the acquisition of interactional expertise is socialization. For example, this would be true in language learning itself in activities, such as literary criticism, that do not involve practical activity.

4 Participant observers who are trying to acquire the culture they are studying rather than observing from a distance may be better referred to as `participant comprehenders' (Collins 1984).
basis of the assumption of false identities in internet chatrooms.\textsuperscript{5} Turning to science and technology, interactional expertise is often the medium of specialist peer review in funding agencies and in journal editing where the reviewers are only sometimes contributors to the narrow specialty being evaluated. It is the medium of interchange within large scale science projects, where again not everyone can be a contributor to everyone else's narrow specialty. It is, a fortiori, the medium of interchange in properly interdisciplinary, as opposed to multidisciplinary, research. Finally, on those occasions when activists or other concerned persons are driven to it, it can be the medium of interchange between scientists and groups of citizens (eg Epstein 1996, Collins and Pinch 2005).

The idea of interactional expertise is illustrated in the cartoon, Figure 1.

FIGURE 1 ABOUT HERE

In the cartoon the left hand group are exercising contributory expertise which involves not only hammering but also talking the language associated with hammering. The right hand group, most still identifiable by their possession of a hammer, have taken time off work and are merely talking the language. Among them is a sociologist who has never had a hammer

\textsuperscript{5} Confidence tricks of all kinds generally depend on a major contribution from the `mark' -- the person being fooled (Maurer, 1940). In the imitation game it is quite clear that the judge's job is to try to distinguish the true expert from the the expert who has mastered only the language; the judge does not make the kind of contribution made by the mark.
and could not handle it properly anyway. Nevertheless, the sociologist has mastered the language of the domain through immersion in it.

Here we report on initial experiments intended, among other things, to make the idea of interactional expertise more concrete. The idea on which the experiments are based is more than 50 years old being the forerunner of the famous ‘Turing Test;’ it is the ‘imitation game.’ We use the imitation game to investigate the discourse of people who do not possess a certain skill but have been immersed in the language. To do this we use the imitation game to compare the domain-specific linguistic abilities of interactional experts with that of contributory experts and that of non-experts. The experiments are represented in Figure 2 below. The four circles on the left represent tests on colour blindness and perfect pitch which are used to explore the concept of the experiment. ‘Proof of concept’ having been accomplished, the rightmost circle represents a first use of the idea in an area of more direct relevance to sociology of scientific knowledge, namely as an indicative test of the ability of a participant observer to pass as a scientist.

---

6 Experiments can have a function that goes beyond their findings. Harvey (1981) shows that initial experiments on the problem of non-locality in quantum mechanics did much to develop the ideas even though the results were not treated as decisive. Collins (e.g. 2004b) argues that Weber’s failed observations nevertheless settled a long-running theoretical argument about whether gravitational waves could be detected in principle and went on to found the half-billion dollar field of gravitational wave detection.

7 For simplicity and flow of exposition, previous research using the imitation game is described in Appendix 1.
The first circle in the top row of Figure 2 represents a society like ours in which the majority of people are colour perceiving (CP); that is, they are not colour blind. A minority are colour blind (CB). The idea of interactional expertise implies that having been brought up in colour perceiving society, the colour blind will be fluent in colour perception language even though they cannot see the full range of colours — they will have acquired interactional expertise in colour perception language though they have no contributory expertise in colour discrimination. Therefore if the discourse of a colour blind person, who is trying to imitate a colour perceiver, is compared to the discourse of a colour perceiver, even if the judge is a colour perceiver, no difference should be detectable and the success rate of the judge's guesses should be no better than chance.

The second circle represents a society, also like ours, in which a small minority are `pitch perceiving' (PP), that is, they have perfect pitch. The majority are `pitch blind' (PB). The theory holds that in this case the pitch blind will not have mastered pitch perception language because they have not been brought up among pitch perceivers and have not been immersed in pitch perception language. In other words, we should not expect the pitch blind to have acquired interactional expertise in pitch perception language. Therefore, if the language of a pitch-blind person who is trying to imitate a pitch perceiver is compared with the language of a pitch perceiver, and the judge is a pitch perceiver, the difference should be detectable. We run imitation game experiments to test these hypotheses. As can be seen, in Figure 2, white signifies the group from which the imitator is drawn while the shaded areas represent the groups from which the judges are drawn.
The bottom two circles in the figure are like the top two circles except that the 'polarity' is reversed and the results should be the opposite. In the first circle of the bottom row colour perceivers try to imitate the colour blind and we should expect them to fail so that a colour blind judge can identify who is who. In the second circle pitch perceivers try to imitate the pitch blind and we would expect them to succeed so that a pitch blind judge (most of us), cannot identify them and the judge's guesses should, once more, be random.

Since we know what we should expect in these cases if the idea of interactional expertise makes sense, the experiments on colour blindness and perfect pitch can be treated as 'proof of concept.'

The rightmost circle in Figure 2 represents an application of the concept and the experiment to science studies and, in effect, participatory, anthropological, and ethnographic fieldwork in general. Here a member of the wider society enters an esoteric group in an attempt to acquire its interactional expertise. In this case the esoteric expertise belongs to gravitational wave physics. The large majority of members of our society are, as it were, 'gravitational wave physics blind' (gwB) -- they have no deep knowledge of gravitational wave physics. A small minority are, as it were, 'gravitational wave physics perceivers' (gwP).

The small white circle represents a member of gwB society who enters the grey circle of the gw perceivers hoping to learn the language without learning the practice of the physics. The final set of experiments described test whether the person represented by the small circle has succeeded in acquiring the targeted interactional expertise. The person tries to imitate the language of gw physics and the test is whether judges who are gw physicists can tell the difference between the participant's answers and those of gw physicists who do have contributory expertise. Success would be indicated by a chance outcome, failure by ready identifiability.
PROCEDURE AND RESULTS: THE PROOF OF CONCEPT EXPERIMENTS

Alan Turing's famous definition of intelligence in a computer (1950) turned on what has become known as the 'Turing Test:' a hidden computer and a hidden person would be interrogated by a judge via teletypes. If after five minutes or so of interchange the judge failed to identify the computer it would be deemed to be intelligent. Turing’s test was based on the 'imitation game,' a parlor game in which a judge asked written questions of a hidden man pretending to be a woman and compared these with the answers of a hidden woman who replied honestly. In our terms, if the hidden man succeeds in fooling the judge the man has demonstrated the interactional expertise associated with being a woman though not the contributory expertise.

This is the protocol we apply to colour blindness and perfect pitch, in this case using purpose-built software to link three computers connected by a wireless network. Judges sit at one computer and can type any question that they think will probe for possession of the target expertise (let us say colour-perception as in the first circle of Figure 2). The question is transmitted simultaneously to both participants, one of whom will be colour blind and pretends to be a colour-perceiver and one of whom will be a colour perceiver who is instructed to answer `naturally.' When both participants have replied the answers appear simultaneously and side-by-side on the judge's screen. The judge can then make a guess and provides a `confidence level' associated with the guess. The judge is then free to ask another question. The session continues until the judge feels there is nothing further to be learned by going on. In our experiments judges usually felt that there was nothing more to be

---

8 The roles can, of course, be reversed and there is a body of philosophical literature discussing the exact protocol of the imitation game and the Turing Test (see Appendix 1).
learned after they had asked around half-a-dozen questions, though some asked fewer and
one or two asked many more.

**Judges**

We believe that to have the best chance of making the correct identification judges should
themselves possess the `target expertise' -- the expertise that the pretender is trying to
imitate. For example, in the gender imitation game if it was the hidden man pretending to be
a hidden woman then the judge should be a woman. A male judge's model of a woman, we
suggest, is likely to be similar to a male `pretender's' model of a woman and this would make
it too easy for the man to pass (Collins, 1990 Ch 13). Thus, in our tests, if the participants
are pretending to be color perceivers the judge will be a color perceiver; if the participants
are pretending to be color-blind, the judge will be color-blind.

**Chance and identify conditions and outcome**

We refer to runs in which we do not expect the judge to be able to identify the participants
as the *chance condition* and the runs in which we do expect the judge to be able to work out
who is who as the *identify condition*. To put into words what is represented in the left hand
circles of Figure 2, we have four experimental configurations and expected outcomes in the
colour blindness and perfect pitch experiments as follows:

<table>
<thead>
<tr>
<th>Pretender is</th>
<th>Target Expertise</th>
<th>Expected Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: Color-blind</td>
<td>Imitates</td>
<td>Color-perceiving</td>
</tr>
<tr>
<td>B: Color-perceiving</td>
<td>Imitates</td>
<td>Color-blind</td>
</tr>
<tr>
<td>C: Pitch-perceiving</td>
<td>Imitates</td>
<td>Pitch-blind</td>
</tr>
<tr>
<td>D: Pitch-blind</td>
<td>Imitates</td>
<td>Pitch-perceiving</td>
</tr>
</tbody>
</table>
Interactional expertise is demonstrated whenever the proportion of right guesses is greater in the identify conditions than in the corresponding chance conditions.

**Confidence and discourses**

The entire typed discourse between judges and participants was recorded and judges were asked to use a four level scale to record their confidence after each guess:

- **1st level:** 'I have little or no idea who is who'
- **2nd level:** 'I have some idea who is who ---- but I am more unsure than sure'
- **3rd level:** 'I have a good idea who is who -- and I am more sure than unsure'
- **4th level:** 'I am pretty sure I know who is who.'

In the description of results provided below, we group judges guesses in the following way: all guesses with confidence levels of 3 and 4 were scored as either right guesses or wrong guesses. All guesses with confidence level 1 and 2, along with refusals to guess, were counted as uncertain whether they were right or wrong.

Whenever judges changed their level of confidence they were prompted to explain why but in this report we discuss only the outcomes of guesses and confidence levels not the discourse or the reasons given for the guesses.

**Results**

There were two `phases' to the experiment. In `Phase 1' we conducted a total of 24 runs roughly split among the four possible conditions. Both the color-blindness series and the perfect pitch series treated in isolation support the hypothesis: in each case there were more

---

\(^9\) The results would not differ much if we used all guesses in the right-wrong analysis and ignored confidence levels.
correct guesses in the identify condition than in the chance condition. Since the numbers involved are small, however, a level of statistical significance is reached (p=0.05 in Fisher's exact test), only when we combine all 24 runs into just two groups -- chance and identify.\textsuperscript{10} Here we find that in the chance condition judges made 4 correct (high confidence) guesses out of 13 while in the identify condition they made 8 correct guesses out of 11.

We then carried out 'Phase 2' of the experiment. This no longer involved interaction between judges and participants in real time. Instead, via ordinary email, we sent the recorded discourses to new judges possessing the target expertise to make new post hoc judgments. We may describe this in the following way: in Phase 1 of the experiment interrogator and judge were combined into one role; in Phase 2 the judge was not the interrogator. In Phase 2 the same dialogues could be sent to many judges and each judge could look at many dialogues. Again, the results of both modalities separately supported the hypothesis. In this phase the combined results were that 5 out of 57 (high confidence) guesses were correct in the chance condition while the corresponding figure in the identify condition was 8 out of 15.\textsuperscript{11} The results are represented graphically in Figure 3 which shows the proportions of wrong guess (top band), uncertain guesses (middle band), and right guesses (lower band), for the combined modalities in Phases 1 and 2; the absolute numbers are also indicated.

\textsuperscript{10} To use the Fisher test we have to construct a four cell table in which wrong guesses and uncertain guesses are treated as one category to be compared with right guesses.

\textsuperscript{11} It is much harder to run the identify condition because two members of the esoteric group are needed -- hence there are fewer identify condition runs.
When, as in Phase 1, the two modalities of Phase 2 were combined the difference between the conditions was statistically significant ($p=0.028$ in Fisher's exact test).\textsuperscript{12} When the results of Phase 1 and Phase 2 were combined the likelihood of the result being due to chance turned out to be $p=0.000$ (Fisher's exact test).

CONCLUSIONS ON COLOR BLINDNESS AND PERFECT PITCH

One striking feature of the experiments was judges' lack of success in making correct discriminations in the chance condition. In Phase 2 the randomness of the outcome is especially striking as can be seen in Table 1 which gives the complete set of results for the chance conditions with the associated confidence levels ('0' indicates 'cannot guess').

Table 1 about here

\textsuperscript{12} The results for each separate modality were also statistically significant in Phase 2.
In the table a minus sign signifies a wrong guess, a plus sign signifies a correct guess, while '00' signifies inability to guess at all -- the numbers indicate confidence levels. The rows in the table indicate the guesses made by different judges in respect of the same dialogue while the columns indicate the performance of a single judge in respect of different dialogues. Simply by looking along the rows and down the columns one can see how random were the guesses in the chance condition. This shows that the color-blind are very good at passing as color-perceivers in the imitation game while those with perfect pitch are very good at passing as those without perfect pitch. It does suggest, then, that a lifetime's immersion in the discourse of a group with a certain contributory expertise enables a person without the contributory expertise to acquire the corresponding interactional expertise, at least as tested by the imitation game. The contrast with the results of the identify condition reinforces the point and shows that without such experience it is hard to acquire interactional expertise in the absence of contributory expertise.

**DEVELOPING INTERACTIONAL EXPERTISE IN A SCIENCE**

Interpretivist social science stresses that the analyst must begin with the actor’s frame of reference and that they must, therefore, find some way to acquire it. From the beginning, empirical sociology of scientific knowledge has been concerned (or ought to have been concerned), with what this means. Is it only full-blown research scientists that can carry out interpretivist studies of their field? This was suggested by critics during the unpleasant period known as the `science wars' that played itself out during the 1990s. On the other

---

13 There are notable exceptions such as the claim made in Latour and Woolgar (1979) that it is their very scientific ignorance that enables them to see the activity of science in a new light.
hand, experience suggests that full blown research competence is not a prerequisite. This leaves us with the question of what it means to understand the framework of the actors being studied in sufficient detail to be able to do authentic social analysis of their world. The answer will differ according to the purpose of the study but it has been argued that a reasonable target for those aiming to do competent fieldwork in sociology of scientific knowledge is the acquisition of interactional expertise (Collins 2004b). We now report on the results of some informal tests along these lines represented by the right-most circle in Figure 2; we tested Collins’s abilities to pass as a gravitational wave scientist, Collins having spent considerable time trying to acquire the interactional expertise pertaining to the field. Here we make no attempt to generate enough results to provide a level of statistical significance but we also describe the reasons for their choices provided by the judges. This, then, is primarily a qualitative study though it has quantitative aspects.

These tests were carried out in a much more simplified form than the proof of concept tests described above. Sets of 7 questions about gravitational radiation physics were sent to participants by email. Participants were asked not to use any reference sources but to think out the answers rapidly from their existing stocks of knowledge. The questions, along with paired sets of answers, were then sent to judges, again via email. Judges were sent a questionnaire to frame their responses which also inquired about their academic specialism and experience. This was a Phase 2 type experiment with the generation of the questions being treated as a separate matter from the judging of the answers. Judges were asked to guess who was who without consulting any reference sources, provide the level of confidence in their guess on a four-point scale like that used in the proof of concept experiments, and explain how they made their choices, if possible on a question-by-question basis.

The main result is that judges who compared answers provided by Collins and answers provided by gw physicists were unable to identify the participants. If, as in the
proof of concept tests, we count high confidence guesses (levels 3 and 4) as right or wrong and low confidence guesses (levels 1 and 2) as indicating uncertainty, we find that out of 9 judges, 2 chose Collins as the gw physicist and 7 were unsure. This outcome was only possible because Collins did not make any technical mistakes in his answers. Collins, then, demonstrated interactional expertise according to the standards of the test.14

Reasons that judges gave for their choice can be divided up into two classes. First there are reasons based on the technical content of the answers. Some of Collin's answers were quite different in terms of technical content to the answers of the gw physicists. Second, there were reasons based on the style of the answers. (Table 2 shows examples of the questions and answers.)

| Table 2 about here |

Where technical content differed it was thought that Collins's answers showed more evidence of being thought through in practical terms whereas the gw physicist's answers tended to be more theoretical or as though drawn from a textbook. This was likely to be the case since in one or two cases, such as Question 7 in Table 2, Collins had no choice but to think the answer through since he had not encountered anything like it before. Therefore he did not know (though he could understand) the 'ready made,' and somewhat more theory-

---

14 If we neglect confidence levels, judges 7 chose Collins as the gw physicist, 1 chose correctly, and 1 could not decide.
based answer that the gravitational wave scientist offered; Collins's answer was correct for the current generation of detectors. Crucially, in such a technically demanding (and nerve-wracking) test, Collins's answer passed muster and, given this, they had an air of authenticity in respect of contemporary experimental practice in gravitational wave physics.\textsuperscript{15}

Some of the judges, and all of the judges in respect of some of the answers, did not feel that the technical differences in the answers allowed them to make a judgement. In these cases they fell back on style and, in the main, Collins's answers were preferred because they were shorter and thus bore the hallmarks of someone who was answering impatiently; this suggested a scientist to other scientists.\textsuperscript{16}

As a “control” we then tried using persons who were not gravitational wave physicists as judges, comprising one scientist in another field and 7 haphazardly chosen academics from the social sciences and philosophy. Again, counting guesses as right or wrong only if confidence scores were 3 and 4, 2 correctly identified the gw scientist, 1 chose Collins, and 5 could not decide. If we ignore confidence levels, however, these judges' guesses were more successful, 5 being right, 2 wrong 2, and 1 unable to guess. Examination of the reasoning shows, unsurprisingly, that it was all based on style and in this case the judges tended to think that the more technical “text-book” answers given by the scientist were the more authentic. This is perhaps best illustrated by the reasons offered by the two judges who made the incorrect identifications, preferring Collins:

\textbf{J11} I have no idea about the detail of any sets of answers, not knowing this field. I thought [Collins] was more persuasive as he/she seemed not to feel the need to elaborate on

\textsuperscript{15} The two participants are identified at the end of the bibliography.

\textsuperscript{16} See note 14XX
answers quite so much or set them in some wider didactic context. As such, [Collins] did not strike me as someone trying to persuade anyone else of their own credentials, presumably because they are not in question.

My guess was based on accumulating evidence from the series of questions, rather than any particular one. It seemed to me that the responses [of the scientist] were going out of their way to appear knowledgeable and 'scientific/specialist'. I suspect that the specialists actually talk to one another in more natural terms [as in Collins's answers], being able to assume shared background knowledge. I'm also aware, though, of how I'm interpreting responses to individual questions to fit in with my overall decision. As a possible get-out, of course individuals vary in manner - and a very senior scientist might give different kinds of answers to a junior one than to another senior colleague.

Taken together with the tendency in the colour blindness and perfect pitch experiments for judges sometimes to take long answers as indicating lying and sometimes short answers as indicating lying (see Appendix 2 below) we seem here to be exploring the literally understood, 'ethnomethodology' of decision-making in the imitation game.

To explore 'the terrain' of the experiments further we ran the experiment again in other configurations. Evans, who has considerable knowledge of gravitational physics as a result of talking to Collins over the years and reading his papers, tried to take Collins's role in the experiment but failed. It was easy to establish that Evans was not able to use his knowledge in the creative ways needed in these tests. What he did know about gravitational wave physics was better seen as 'information' rather than interactional expertise.  

17 Evans knowledge of gw physics fits into one or other of the leftmost categories of expertise find in the third row of the 'Periodic Table of Expertises'
We also asked physicists from other specialisms to try to pass themselves off as gravitational wave physicists but they were easily detected because they made technical mistakes or showed glaring gaps in their technical knowledge. Collins also tried taking the role of both interrogator and judge and found he could use both questions he generated himself and those set by other gravitational wave scientists to distinguish gw physicists from other physical scientists who were not gw specialists and from Evans; he could do this without fail at confidence level 4 based on the technical mistakes or lacunae evident in the non-specialists answers. These experiments establish that it is possible for a participant observer (comprehender) to acquire sufficient interactional expertise to pass as a specialist in these tests but that it is not a trivial accomplishment. Those with mere 'information' about the field cannot pass and this includes physical scientists in cognate fields (astrophysicists, astronomers, and relativists). The last finding is especially important as in public life (www.cf.ac.uk/socsi/expertise) -- these depending on only ubiquitous tacit knowledge; Collins's interactional expertise is in the right hand group which depends on specialist tacit knowledge.

Numbers were small for these last tests -- only two runs in each configuration -- but the technical errors and gaps were so obvious that it was pointless to go on.

Academics attending the 2005 meeting of the Society for Social Studies of Science (many of whom have science backgrounds), were also asked to try to judge between answers given by gravitational wave specialists and non-specialist scientists. Of the 20 who replied, 8 guessed right (high confidence) and 2 wrong with 10 uncertain. (Unweighted results were 12 right 7 wrong, 1 could not guess). We infer that it is easier to act purely as a judge than as an interrogator and judge because, once more, there are cues in the style of the answers.
`scientists' are often asked to fill the role of generalised expert in matters remote from their specialism.

OVERALL CONCLUSIONS

The first experiments described here, concerning colour blindness and perfect pitch, are a 'proof of concept' of the idea of interactional expertise and of the imitation game as an empirical approach to the idea. The second set of experiments illustrate one way in which both the idea and the investigative technique can be used in a way which directly relevant to social studies of science and its wider concerns.

The entire set of results begs the question of why all those who claim to practice a research method that involves taking on the frame of reference of the actors should not base their thinking on the idea of interactional expertise and should not look to the possibility of using the imitation game to explore the extent to which they have acquired an understanding of their respondents' frame of meaning.

The importance of the second set of results for the role of scientists in public has been indicated. It may also be that a detail of the design of the experiments points to something else. Those composing the questions for the gravitational wave imitation games were asked to avoid mathematics. Were calculation or algebraic manipulation needed to answer the questions, Collins would have had no chance of passing the test. Does this make the test worthless? We would argue that it does not for the following reasons. First, a subset of the most creative experimental scientists are known for their weakness at mathematics, their creativity coming primarily from their ability to visualise the way the world fits together and/or their abilities at the bench (Collins, 2004b, pp 561,774). Second, and much more important, the kind of discourse actually used in the imitation games described here is the level that is used on peer review committees and the like (Collins,
2004b, Ch 24). This has to be the case because peer review committees, such as those that make largish funding decisions, are composed of scientists who do not know the field they are judging to the level of its algebraic or calculational niceties. For example, to know the exact pattern of the directional sensitivity of a gravitational wave interferometer is a matter of mathematical analysis and calculation such as could be done only by someone with contributory expertise, but to see and understand a diagram representing the results of those calculations, and to understand their significance for the astronomical value of the instrument under review, does not require that the calculation be repeated, only that it is known to have been broadly agreed to be correct. Like every other scientist, peer reviewers have to take nearly all the exact work of science in anything other than their narrow specialisms on trust, so they need only interactional expertise to understand the implications of that detailed work. Since science studies’ interest in the topic of interactional expertise is mostly likely to turn on the role of review committees, public understanding of science, and the like, we have all that we need in the experiments as conducted here. We might add, that this is a good thing or the complaint made by the science warriors would be upheld: it would be impossible to do the sociology of scientific knowledge without the full range of abilities possessed by the contributory expert!

The still wider significance for science studies of this approach lies in a move from studying truth to studying expertise as the touchstone of reliable knowledge and as a basis for forming prescriptive principles for the utilisation of scientific and technical knowledge.
APPENDIX 1: PREVIOUS RESEARCH ON THE IMITATION GAME

There is a literature about the Turing Test. Its roots in the imitation game are interestingly explored in Richard Hodges's (1985) biography of Alan Turing. A collection of philosophical essays about the Turing Test can be found in a special issue of the journal *Mind and Machine* published in 2000. Particularly informative is the opening review essay by Saygin, Cicekli, and Akman. For a detailed consideration of the Turing Test and, by implication, the imitation game, as a real experiment see chapter 13 of Collins 1990.

We know of several previous attempts to use the imitation game to explore the nature of gender. Collins tried informal but informative gender-based experiments at the University of Bath in the early 1990s using pencil and paper as the medium of interchange. At the University of Southampton in 1995 the gender game was tried using computers by Collins and Hall. Subsequently, at the University of Southampton's Psychology Department, Remington supervised an MSc thesis on the gender game which did not produce anything considered worth publishing (private communication). Simon at Queen's University, Ontario, has been running gender imitation game experiments since 1999 using pencil and paper; at first he found he was obtaining gender correlated results but these were no sustained (private communication). The gender imitation game was also made available by Collins and Hall for public participation during the celebration of the painter van Dyke's quadricentennial as described in Collins, 2001.

So far as we know the only published analysis of the detailed results of imitation game runs emerge out of the web-based, chatroom style experiment set up by Berman and Bruckman at Georgia Tech in the late 1990s. The initial publication (Berger and Bruckman

---

19 Turing Tests with machines are also run every year in pursuit of the `Loebner Prize'

(www.loebner.net/Prizef/loebner-prize.html)
2001) describes the tests as indecisive although giving rise to many examples of gender stereotyping and an inventory of useful questions to ask in gender imitation games. Nyboe, (2004) uses the Georgia Tech results not to look for the discriminatory power of the imitation game but for the permeability and negotiability of gender categories.

Herring and Martinson (2004) usefully describe a number of other analyses of the discourse found in gender chat-rooms. They go on to do a post hoc statistical analysis of the Georgia Tech discourses. They report that though the judges failed to be able to distinguish genders, their analysis of a large sample of the discourse show that those who identified themselves as truly males were more verbose. Males' conversational turns have a mean of 14.8 words while females' turns average 13 words. They note, however, that judges involved in the games failed to use this cue as an aid to identification and in so far as they did refer to word length they expected females to be more verbose. (In any case, it is not clear that judges could have detected such small differences in the length of conversational turns even if they were aware of the way it correlates with gender. Small effects that are only distinguishable as a result of statistical analysis of large samples, as in the Herring and Martinson example, however statistically significant are not likely to be useful discriminators in everyday life.) The other problem with the Georgia Tech results is that they depend on participants being honest in their close-out self-identifications.

APPENDIX 2: CONFOUNDING FACTORS

It is important in these experiments to minimise the effect of certain potential confounding factors. Target expertise, or lack of it, can be correlated with some other factor and it is the presence or absence of the other factor that is recognized by the judge rather than the target expertise itself. For example, if the judge knows the participants personally it may be possible to identify them as individuals -- and hence identify their capacities -- by recognizing
features of their typed responses that have to do with them as persons and nothing to do with the target expertise. In the proof of concept experiments described here the participants came from a limited number of university departments and were likely to know each other. This problem can be largely eliminated by making sure that those taking part in any one game do not know who the other players are. This can be managed if the three locations needed are remote from each other.

It is also possible that the pretender could be detected because they have to lie and it is discourse associated with lying that is detected rather than anything to do with the target expertise per se. The problem was illustrated when a judge set a logical trap for the participants. But even where logical traps are not used, we found that judges tend toward folk-theories of lying; for example, some thought that a very short answer indicated a liar while others thought a very long answer indicated a liar and this does confound the experiment to some extent. We tried to ameliorate this effect by asking judges to direct their questions to the expertise rather than trying to trap the participants in lies.

A third possibility is that possession/non-possession of a target expertise is correlated in the population with some other factor that is easier to spot. For example, people with perfect pitch usually know a lot about music. Care needs to be taken with this problem, the effects of which can, once more, be ameliorated by adjusting the instructions given to judge and participants.

The effect of the confounding factors described above should be to make it easier for the judge to identify the participants. This would lead to an anomalous lack of support for the hypothesis under the chance condition and too much support for it under the identify condition. With luck, then, the effect of these confounding factors, even if they cannot be entirely eliminated, will balance out. The experiment, as we carried it out, also offered one strong indication that the confounding factors did not play a significant role: the
guesses made within the chance condition were notably random and this indicated that no anomalous preponderance of identifying features was present.

There are, of course, many other potentially confounding factors: variation in the ability of the judge to think of good questions and make good distinctions; the number of questions asked; variations in the ability of the participants to take the role of the other; and variations in the extent to which those with the ubiquitous expertise have been exposed to the specialist expertise. Some of these could be ameliorated by more intricate experimental design or could themselves become a topic for experimental investigation. In the report presented here they are treated as unavoidable `noise' which will result in a less than perfect outcome.

Finally, color-blindness and perfect pitch have been discussed as though they were binary qualities -- one either has them or one does not. But they each come in different types and each type lies on a continuum. To get a strong effect in the experiment participants need to be located toward the extreme end of the spectrum. For example, the well-known Isihara test for color perception requires people to read numbers that are distinguished from the background only by a slight variation in color. A person can fail the Isihara test but still be able to distinguish the much brighter colors of, say, traffic lights. The answers given by such a minimally color-blind person are likely to be indistinguishable from that of a color-perceiver under any test; likewise, a judge who is only marginally color-blind in this way does not really possess color-blindness as a `target expertise.' It is sensible, therefore, to test for color-blindness, not with something as sensitive as the Isihara test, but by discussing more brightly and obviously colored objects with a potential participant or judge.

Perfect pitch is even more confounding. Many people have `relative pitch,' which means they can tell the exact interval between notes without being able to place the notes on an absolute scale; this is not perfect pitch. Others are able to name the notes produced by
musical instruments but not those that make up more complex sounds, such as the bleeps made by road-crossings or the notes in a human voice. The experience of the world is quite different for these two groups and to avoid getting the two kinds of experience confused we picked participants who could identify mixed sounds, not just pure sounds, but such people are rare.

The effect of this fourth kind of confounding factor would be to prevent the experiment 'working' because there would be too little difference between the participants to make it possible for the judge to distinguish them. This kind of factor, then, is not a worry in the case that the experiment produces a positive result.

**Phase 3 of the experiments on colour blindness and perfect pitch**

There remains a possible alternative interpretation of the proof of concept results: that the difference lies in the judges not the participants. All judges possess the target expertise and this means that the different conditions utilised different judges. It could be that being color-blind or having perfect pitch encourages one to think about the issues more deeply than those who have 'normal' abilities. It could be, then, that identify-condition judges would be better discriminators in the chance condition too.

To test this we had dialogues generated in the chance condition examined by the 4 most successful identify-condition judges, generating 18 new guesses. It turns out that their discriminatory ability was no better than chance. The difference in guessing ability under these circumstances and their ability when judging questions and answers generated under the identify condition is statistically significant (p= .028 in Fisher's exact test). This strongly suggests that the differences discovered in Phase 2 of the experiment were not primarily due to the differential skills of the judge. The numbers are small, however, and to pick out any residual contribution of a 'judge effect' with certainty would need a new and bigger experiment.
We also examined the possibility that identify condition dialogues are specially revealing because 'specialists' have generated them. We gave identify-condition dialogues to 6 chance-condition judges giving rise to 31 new guesses. The results show that either the dialogues were in fact not specially revealing or the chance-condition judges did not have the skills to recognize what was revealing about them -- as per the major hypothesis.
REFERENCES


Collins, H. M., (2004a) 'Interactional Expertise as a Third Kind of Knowledge.', *Phenomenology and the Cognitive Sciences*, 3, 2, 125-143.


Collins, Harry, (2004c) 'How Do You Know You've Alternated?' *Social Studies of Science* 34, 1, 103-106


Nyboe, Lotte, (2004) "You said I was not a man": Performing Gender and Sexuality on the Internet.', *Convergence*, 10, 2.


[In Table 2 the gw physicist is A, Collins is B] **EDITOR, PLEASE PRINT THIS LINE IN 7 POINT OR LESS**
Figure 1: Interactional expertise
Figure 2: Experimental configurations
Figure 3: Results of proof of concept experiment
Table 1: Phase 2 Chance condition results

<table>
<thead>
<tr>
<th>JUDGES</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1</td>
<td>+1</td>
<td>+2</td>
<td>-2</td>
<td>+1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-2</td>
<td>0</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>+1</td>
<td>-2</td>
<td>-2</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>+2</td>
<td>+2</td>
<td>0</td>
<td>+2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0</td>
<td>+1</td>
<td></td>
<td>+2</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>+2</td>
<td>0</td>
<td>+2</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td>-2</td>
<td>-1</td>
<td>+1</td>
<td>+2</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td>-2</td>
<td>+2</td>
<td>-3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>JUDGES</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-3</td>
</tr>
<tr>
<td>2</td>
<td>-2</td>
<td>+3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-3</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>+1</td>
<td>+4</td>
<td>0</td>
<td>-4</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>-1</td>
<td>+3</td>
<td>+4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>+3</td>
<td>-3</td>
<td>+1</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Phase 2 Chance condition results
### THE IMITATION GAME FOR GRAVITATIONAL WAVES

<table>
<thead>
<tr>
<th>Q2) Is a spherical resonant mass detector equally sensitive to radiation from all over the sky?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2) Yes, unlike cylindrical bar detectors which are most sensitive to gravitational radiation coming from a direction perpendicular to the long axis.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Q3) State if after a burst of gravitational waves pass by, a bar antenna continues to ring and mirrors of an interferometer continue to oscillate from their mean positions? (only motion in the relevant frequency range is important).</th>
</tr>
</thead>
<tbody>
<tr>
<td>A3) Bars will continue to ring, but the mirrors in the interferometer will not continue to oscillate.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Q5) A theorist tells you that she has come up with a theory in which a circular ring of particles are displaced by GW so that the circular shape remains the same but the size oscillates about a mean size. Would it be possible to measure this effect using a laser interferometer?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A5) Yes, but you should analyse the sum of the strains in the two arms, rather than the difference. In fact, you don't even need two arms of an interferometer to detect GWs, provided you can measure the round-trip light travel time along a single arm accurately enough to detect small changes in its length.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Q6) Imagine that the end mirrors of an interferometer are equally but oppositely (electrically) charged. Could the result of a radio-wave incident on the interferometer be the same as that of a gravitational wave?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A6) In principle you could detect the passage of an electromagnetic (EM) wave, but the effect is different than for a GW. The reason is that unlike EM waves, GWs produce quadrupolar deformations. A typical EM wave would change the distance in only one arm while a typical GW wave would change the distances (in opposite ways) in both, so the differential signal for the EM wave would be half that for a GW.</td>
</tr>
</tbody>
</table>