The Imitation Game and the Nature of Mixed Methods

Harry Collins, Robert Evans, Martin Weinel, Jennifer Lyttleton-Smith, Andrew Bartlett and Martin Hall.

VERSION HISTORY

This is the authors' version of a work that was submitted to the *Journal of Mixed Methods Research* on 15 October 2015 and accepted for publication on 1 November 2015. It is being made available in accordance with the publisher's policies on open access.

Please note that changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version.

For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper. For this, and other information, please go to: http://mmr.sagepub.com/

Contact details (all authors)

Cardiff School of Social Sciences
Glamorgan Building
King Edward VII Avenue
Cardiff CF10 3WT
UK

Emails:

Collins: CollinsHM@Cardiff.ac.uk (Corresponding author)
Evans: EvansRJ1@Cardiff.ac.uk

The Imitation Game and the Nature of Mixed Methods[1]

Abstract

We describe the Imitation Game, a new research method that simultaneous generates qualitative and quantitative data and which can be used in many disciplines. Drawing on two projects, one investigating gender, the other sexuality, we show that the quantitative and qualitative aspects of the Game combine in four different ways, from more quantitative to more qualitative, involving increasing cultural understanding by the researchers. Crucially, deep cultural input is initially supplied by the players of the Game, who act as 'proxy researchers,' enabling data to be gathered quickly and efficiently. The analysis, which has its roots in Sociology of Scientific Knowledge (SSK) and Studies of Expertise and Experience (SEE), emphasizes the cultural foundations of both methods and expertise more generally.

Keywords

Imitation Game, Turing Test, Proxy Researcher, Sociology of Scientific Knowledge, Studies of Expertise and Experience

Introduction

Discussions of mixed methods mostly turn on how to bring pre-existing quantitative and qualitative approaches to bear on each other (e.g. Johnson, Onwuegbuzie, Turner, 2007; Creswell and Clark, 2011; Bryman, 2006). In contrast, this paper introduces a new method – the Imitation Game – in which quantitative and qualitative analysis are integral from the outset and in which experimental and survey approaches are combined (Collins & Evans, 2014). The Imitation Game is at a pioneering stage and we provide an initial analysis of the methodology in the hope that, over time, a more complete understanding will be developed through the joint efforts of the scientific community. Because the Imitation Game is new and evolving we examine the cultural abilities and skills needed to use it and the implications of this analysis for the practice of mixed methods research in general.

We begin by describing the principles that inform the Imitation Game and the various ways in which these have been implemented. The central axis of the paper is an analysis of the way cultural competences feed into the conduct of the Games and analysis of the data. Using Table 1 as the organizing principle, we show how the Imitation Game enables quantitative results to be generated by researchers with little cultural competence but that ever more native understanding is required as analysis becomes more qualitative. We conclude by setting out the future opportunities and challenges for the Imitation Game and invite readers to join us in exploring its possibilities.

Origins of the Imitation Game

Alan Turing, mathematician, WW II code-breaker and founder of computer science, is perhaps most famous for proposing the ‘Turing Test’ as a way of determining whether a machine should be classed as ‘intelligent’. The test, in which a judge asks questions of a

hidden computer and a hidden human, drew its inspiration from a parlor game in which a 'judge' asks written questions of a hidden man and a hidden woman (Turing, 1950; Hodges, 1985). As Turing describes this 'imitation game', the woman would answer naturally while the man would pretend to be a woman; the judge's task was to devise questions that would reveal who was who. By replacing one of the human players with a machine, the Turing Test defines 'being intelligent' as the ability to demonstrate contextual understanding (Collins, 1990).

Here we return to the original parlor game and show how it can be transformed into a new form of sociological research. The method draws on insights from the sociology of scientific knowledge (SSK) and its sub-field, 'studies of expertise and experience' (SEE). From SSK, we take the idea that knowledge is a collective phenomenon that depends on tacit knowledge that can only be acquired via socialization into the relevant community. From SEE, we take the idea of interactional expertise, which allows us to see language as something that can be learned without directly experiencing the practices the language describes. Combining these ideas suggests that, where different social groups have regular interactions, then interactional expertise should be shared in sufficient depth for members of one group to describe, and hence display understanding of, the experiences of the other.

In the remainder of this paper we show how this claim can be investigated using Imitation Games. We describe how Imitation Games can be played in different formats, with different numbers of players, and on topics ranging from the esoteric (e.g. gravitational wave physics) to the ubiquitous (e.g. gender or sexuality). In order to distinguish the research method from the parlor game, we capitalize 'Imitation Game' when referring to the social science usage. We also distinguish between *four* capitalized roles – Interrogator, Pretender, Non-Pretender

and Judge. Interrogators invent and ask questions while Judges gauge the plausibility of the answers. In some settings the Interrogator and Judge roles are combined (Interrogator/Judge) but they can be played by different persons when this is advantageous.

The Basic Imitation Game

The simplest Imitation Game starts with three participants drawn from two social groups and takes about an hour to play (see Figure 1 for a schematic representation). Two of these players – the Interrogator/Judge and the Non-Pretender – are from the same social group and so share the 'target culture'. The third – the Pretender – is from the other social group and pretends to share their culture. The aim is for the Interrogator/Judge to determine who is who by asking questions about the target culture. The Interrogator/Judge is allowed to ask as many questions as they like within the time limit of the test (they typically ask 6 to 8 questions) and they invent the question themselves. Given that we are interested in the experiences of the target culture, we encourage Interrogator/Judges to ask questions that relate to experiences they believe to be unique to their group, which require immersion in the practices of that group to understand, and which encourage respondents to provide examples, details or reasons. For the same reason, we also discourage them from trying to work out who is lying and to base their judgments solely on the players' substantive understanding of the target culture.

Figure 1 (Schematic Representation of the Imitation Game) about here

Even in this simple form the principal advantages of the Imitation Game are clear. Firstly, because the players devise the questions, create the answers and make the judgments about what is plausible it is the players who decide what is culturally significant, what is not, and

why. This means that the Imitation Game does not require the professional researcher to understand the culture being investigated *before* collecting data. In contrast, methods such interviews, focus groups and surveys all require the researcher to develop some initial cultural competence. Secondly, the game format provides a ready motivation for the players since Interrogator/Judges will try to ask difficult questions and Pretenders will try to provide the best possible answers. Refining and analyzing the data does require more cultural competence, as explained below, but for rapid data collection on unfamiliar topics it is hard to see how the Imitation Game can be bettered. One useful way to think about the Imitation Game is as a sociological camera: by playing the Imitation Game participants create 'sociological selfies' in which they display, capture and preserve a cross-section of social understanding.

Imitation Game Variants

The Imitation Game is usually played over the internet using standard web browsers and uses bespoke software to organize play and transfer and store data. Players can be in the same room or in different locations. Once complete, each Imitation Game generates at least five different types of data: the questions asked by Interrogators; the answers provided by the other two players; the decision of each Judge as to who is who; a measure of each Judge's confidence about this decision; and the reason given by each Judge for their decision. Judges' decisions, confidences and reasons are recorded after each question-and-answer turn. A separate final judgment, confidence level, and reason are also collected in which Judges evaluate the dialogue as a whole.

This core Imitation Game data can be supplemented with demographic information, attitude scales, psychometric tests, self-evaluations and so on by asking participants complete

questionnaires before and/or after taking part in Game. In more recent research, we have used groups of 3 or 4 players in each role: the groups discuss the questions, answers, confidence levels and reasons among themselves and these discussions can be recorded. These groups can be thought of as small, self-organizing focus groups – hereafter mini-focus groups – and each Imitation Game creates three such groups.

The creation of these variants is possible because the Imitation Game is a research method, not a research protocol. It is comparable to the survey, interview or experiment, with its uses limited only by the ingenuity of the researcher.[2] Figure 2 shows how the same underlying structure can support a variety of different research projects. At the smallest scale, represented at the top of Figure 2, the Imitation Game can be used by individual ethnographers to test the extent to which they have acquired the 'interactional expertise' (Collins and Evans 2002, 2007, 2015; Collins 2004a, 2011) of the group they are studying. For example, Collins conducted a long-term sociological study of gravitational wave physics (Collins, 2004b, 2013) and showed that, after many years immersed in the community, he could pass as a gravitational wave physicist: comparing the answers to technical questions, seven out of nine gravitational wave physicists acting as Judges were unable to distinguish between the answers provided by Collins and a real gravitational wave physicist, and two thought Collins was the physicist (Giles, 2006).

Figure 2 (Versions of the Imitation Game) about here

The Imitation Game can also be played with small numbers of participants. This is represented in the middle of Figure 2 by a study in which we showed that the blind were much more successful in passing as sighted than the sighted were at pretending to be blind

(86% vs 13%). The reason is that the blind have been immersed in the spoken discourse of the sighted all their lives whereas the sighted rarely encounter the spoken discourse of the blind (Collins and Evans, 2014). In a similar vein, we have used the Imitation Game to investigate the experiences of the colour blind and those with perfect pitch (Collins et al, 2006) and the ability of medical practitioners' to take the perspective of patients with chronic conditions (Evans and Crocker, 2013; Wehrens, 2014).

Finally, and as reported in this paper, the Imitation Game can be played on a large scale and used to compare the relative understanding of social groups within different nations. With this in mind, many hundreds of Games have been played on topics such as sexuality (straight men pretending to be gay men) and gender (men pretending to be women and vice versa).[3]

Practical details

Individual and small-scale Imitation Games can be played in several ways. In the gravitational-wave test undertaken by Collins and the dietitians study run by Evans and Crocker, email was used, with all questions and answers being sent via a 'postman', who anonymized all communication and kept track of participants' true identities. Where larger numbers of games and/or participants are involved logistical problems increase and we use custom-built software to run the Games and record the data.[4] As the pace of questions and answers can be slow, the program allows each player to play three games simultaneously, continually switching between the role of Pretender, Non-Pretender and Interrogator. Each player plays with six others over the course of three games but the software ensures that they are all different. An additional computer program generates a seating plan for a single computer lab that keeps all seven participants in a set of games well separated.

The 4-Step method

When playing large numbers of Imitation Games, the logistical problems can become quite daunting. In addition, sociologically interesting topics often involve easy-to-recruit mainstream populations pretending to be hard-to-recruit groups. In either situation, splitting Imitation Game into four component parts is desirable as it makes the most efficient of scarce or difficult to recruit participants and minimizes problems of organization.

Step 1: Generating Questions and Non-Pretender Answers

Step 1 consists of a number of real-time Imitation Games played simultaneously, as described above. Although Step 1 Interrogators also act as Judges, the primary purpose of Step 1 is to generate sets of questions and corresponding Non-Pretender answers that can be re-used in the next 3 Steps. The final judgments and Pretender answers generated at Step 1 are, therefore, stored by the software but not used in the subsequent quantitative analysis. Before proceeding to Step 2, all questions and Non-Pretender answers are checked to ensure that the players have followed the instructions correctly and understood the purpose of the Game. Where mistakes or other problems are detected, the individual question or, in some cases, the entire question set is discarded. This checking typically results in a few individual questions being discarded, and occasionally an entire question-set, but it is extremely rare for more than one question-set to be lost.

Step 2: Collecting Pretender Answers

The approved question-sets are administered as an online survey to around 200 mainstream respondents. Each respondent takes the role of Pretender and provides answers to one question-set, creating around 10 sets of Pretender-responses for each Step 1 question-set.

Step 3: Creating Unique Transcripts

Each set of Pretender answers from Step 2 is automatically combined by the software with the corresponding Step 1 questions and Non-Pretender answers to create approximately 200 unique dialogues.

Step 4: Judging Answers

These dialogues are distributed in batches of 8 to around 50 new Judges from the minority group in such a way that each dialogue is judged twice. As with Step 1, Judges at Step 4 are asked to provide a judgment, a confidence level and a reason. At Step 4 Judges see the whole question and answer set pertaining to each dialogue at once, rather than seeing the questions and answers build up over time in the context of two other Games. This means Step 1 and Step 4 judgments are not comparable and Step 4 judgments alone are used to calculate the summary statistic described below.

Pass Rate

For Imitation Games the basic summary statistic called the ‘pass rate’ is based on Judges’ final guesses about who is who. It is expressed as a percentage and given by the formula:

$$1 - \left(\frac{Right\ guesses - Wrong\ guesses}{Total\ guesses\ (incl.\ Don't\ Know\ responses)}\right)$$

Note that if the Judges’ guesses run out at 50/50 right and wrong, the pass rate will be 100% while it will be 0% if the Judge gets everything right; if some Judges are pusillanimous and tend toward a ‘don’t know’ choice, this will make no difference to the overall pass rate so long as equal numbers of potentially right and wrong guesses are diverted to the don’t know category.[5]

In the large scale games the pass rate can be taken as a proxy for the extent to which one group understands the other and can be used to frame and state hypotheses. A rule of thumb is that, with sample sizes of 200 dialogues, absolute differences of 10-15% in pass rate will be statistically significant at the 2-sigma level.

Imitation Game Data, Cultural Understanding and Mixed Methods

Large-scale Imitation Games give rise to a blizzard of data that can be analyzed in many ways. The main approaches are illustrated in Table 1, which will be used to organize the discussion of how the Imitation Game bears upon mixed methods and, in particular, the increasing cultural competence required as the researcher moves from column 1 to column 4.[6]

Table 1 (Types of Imitation Game Analysis) about here

In setting out our analysis we draw on two main Imitation Game projects:

- Project 1: Imitation Games on gender conducted in Granada, Spain, and using the 4-Step method. The purpose of the research was to compare the effect of group versus individual play at Step 1 on the final pass rate. The research participants were students at the University of Granada and, in the first set of Imitation Games, we ran two Step 1 sessions, one with individuals and one with groups. Steps 2 and 4 were all played by individuals. In the second study, we aimed to replicate the results of the group games and so ran a single Step 1 session with groups, followed by a new round of Steps 2 and 4. Gender was chosen as the topic because the combination of group games with a subsequent replication required approximately 800 participants. A breakdown of

sample sizes at each stage is provided in Table 2, along with a formal statement of the research hypothesis

- Project 2: Imitation Games on male sexuality conducted in Cardiff, UK and Wroclaw, Poland and using the 4-Step method. The purpose of the research was to compare the ability of straight men to pretend to be gay in two countries with very different attitudes to homosexuality. The research participants were students at the Polish University of Humanities and Social Science Faculty in Wroclaw and Cardiff University in the UK. In each location we ran a single Step 1 session followed by Steps 2 and 4. Sample sizes for each Step and a statement of the research hypothesis are listed in Table 2

Table 2: (Characteristics of Imitation Games) about here

Quantitative Analysis

We start with the quantitative analysis of entire games, as shown in column 1 of Table 1. Here pass rates are compared and differences analyzed for statistical significance. Although this appears to be a purely quantitative procedure that requires no cultural knowledge, we nevertheless ask the question that we are going to repeat for every element of the Table: who supplies what in the way of cultural understanding?

We, the authors of this paper, are mostly native English speakers with one native German speaker (who is also fluent in English) among us. How do we conduct Imitation Games in, say, Poland? The answer is that, for a month or so, we employ English-speaking Polish researchers, ideally based in the social sciences, as ‘local organizers’. We specify the number of players required and the kinds of facilities needed, and ask the local organizers to find the

computer laboratories, recruit the participants and organize the games. To do this, the local organizers must find their way around Polish universities, solve all manner of administrative and technical problems and ensure the right numbers of participants turn up to the right place at the right time, all of which requires a good cultural knowledge of their university and its students. It is also the local organizers who translate the instructions and check the Step 1 questions, both of which require competence in the local language. A subset of our team will then travel to Poland for about a week to supervise Step 1 and Step 2 data collection, leaving Step 4 for the local organizer to complete. These duties are not fixed, however. More experienced local organizers can run Step 2 independently and, as an international body of expertise builds up, even Step 1 can be run independently.

Note also that the Games described in Table 2 cover several topics. Neither we nor the local organizers can be knowledgeable about so many cultures and practices. What makes the Imitation Game possible is that the players, as 'proxy-researchers', supply the domain-specific knowledge needed to explicate the local culture and define its boundaries.[7] Thus, it is the players who know the local practices, work out what kinds of questions to ask as they fulfill the role of Interrogator, what kind of answers to give when they are acting as Non-Pretenders, what kind of Pretender answers might succeed, and make the all-important judgments about which answers are to count as plausible. The Interrogators, then, fill the role of questionnaire designers, with the mini-focus groups 'piloting' the questions and reflecting on what 'people like them' know and do. This input from the players applies across the table and it is the foundation for all the data generated.

Examining the consistency of these judgments (second row of column 1) is, as far as the researchers are concerned, a purely quantitative procedure that can be carried out without

understanding the target culture or local language. Although we have not done this, using slightly different computer set-ups it would be possible to try consistency experiments for paired sets of Non-Pretenders at Step 1 or split Step 2 Pretenders into random groups and see how they compare. All this contributes to our analysis of the systematic and random errors within Imitation Games.

Other possibilities for future research include using non-expert Judges at Step 4, to gauge how much expertise is required to judge the dialogues accurately, or using non-expert Interrogators at Step 1 for generating questions. We assume that both would lead to higher pass rates. These experiments and their analysis are mentioned in the third row of column 1.

The next row of column 1 describes experiments and measurements that use 'dialogue-sets' as the unit of analysis. Each question set carried forward from Step 1 generates around ten completed dialogues and, by treating each of these as a unit, with its own pass rate, comparisons between dialogue-sets are possible. That is how we discovered that dialogue-sets using groups at Step 1 gave rise to much lower pass rates than Games using individual Interrogators. This seems to be because the discussions within the mini-focus groups increase reflexivity, allows players to pool their expertise and weed out atypical or idiosyncratic views.[8]

Figure 3 shows, for the first time, the results of Project 1, which was designed to test the hypothesis that using groups at Step 1 would give rise to a lower pass rate at Step 4 (see Table 2 for details). To make the comparison, the real-time Step 1 element was played twice, once with 40 individuals acting as Interrogators/Judges and once with 20 groups of 4 players acting as Interrogators/Judges. This generated a total of 54 usable sets of questions, 27 for female Judges (18 individual Step 1, 9 group Step 1) and 27 for male Judges (19 individual

Step 1, 8 group Step 1). In Step 2, a new sample of 537 participants provided new answers to the questions generated at Step 1. Each participant answered one set of questions, with the allocation being done on a 'round-robin' basis to ensure roughly equal numbers of Pretenders answered each of the 54 sets of questions.

These Pretender answers were then combined with Step 1 questions and Non-Pretender answers to create the dialogues for Step 4 Judges. In total, 130 Judges each evaluated 8-10 dialogues, each containing a mixture of individual and group question-sets. Dialogues were allocated in such a way that each unique dialogue was judged twice, with the data reported as 'Individual' and 'Group (1) in Table 3 being the average of the two sets of evaluations. Differences between male and female judges are not significant for individual judges ($t(368) = 1.764$, $p = 0.079$) and barely significant for group games ($t(160.6) = 1.982$, $p = 0.049$). In contrast, the difference between individual and group games is clearly significant for both male judges ($t(286) = -2.572$, $p = 0.011$) and female judges ($t(185.1)$, $p = 0.000$).

Table 3: Distribution of Results in Granada Imitation Games

To check that the difference between group and individual play at Step 1 could be reproduced, we returned to Granada six months later and ran another set of Imitation Games (Granada (2) in Table 2) in which we played one round of Step 1 with group Interrogators/Judges, followed by Step 2 and Step 4 as shown in Table 2. The results were broadly similar, with the pass rates for Games with group Interrogators being similar to each other and obviously lower than those produced by individual Interrogators. The comparison between pass rates, in this case for female judges, is represented graphically, and far more strikingly, in Figure 2, where the dialogue-sets are ordered by pass rate. As can be seen,

where Step 1 was conducted using groups (the grey columns) rather than individuals (the black columns), pass rates were halved. One important implication of this is that sociologically meaningful comparisons require the Games to use similar protocols.

Figure 2 (Bar chart of Pass Rates) about here

In the fifth row of column 1 are found analyses of individual judging performances. Using only arithmetical procedures we can look to see how often Step 1 Judges change their guess about who is who through a single game or use the confidence levels to identify especially revealing question-answer combinations. Of course, to understand why a question-answer turn enables a judge to correctly identify the Pretender with a high level of confidence requires some level of linguistic and cultural competence; once more, this makes the point about the inter-dependence of quantitative and qualitative approaches.

Analysis of Textual Data

Columns 2-4 of Table 1 refer to analyses in which deeper, more local, levels of cultural expertise are brought into play by the researchers. To understand these different kinds of expertise we draw on the Studies of Expertise and Experience (SEE) to provide a language for talking about expertises.

SEE treats all expertise as cultural and vice versa. As far as SEE is concerned, being a native English speaker, being a member of the gay community, and being a gravitational wave physicist are all matters of acquiring tacit knowledge through socialization into the relevant community. The only difference is that gaining some of these expertises involves interaction with the entire society, some with large groups within that society and some with small

groups of specialists. Under this approach, expertise is no longer scarce by definition and the difference between widespread and specialist expertises is sociological not epistemological.

Although SEE identifies a dozen-or-so different types of expertise that are classified in a Periodic Table of Expertises (Collins & Evans, 2007), three are particularly relevant for understanding the Imitation Game. The first, 'Contributory Expertise,' is defined as full practical expertise in some domain and is close to the common sense meaning of 'expert'. The second is 'Interactional Expertise,' which is defined as fluency in the language used by a group of contributory experts to describe their practices. Interactional expertise is acquired through prolonged socialization but may be obtained without any practical experience. The third category of expertise needed to understand the Imitation Game is 'Ubiquitous Expertise.' Ubiquitous expertise is the general cultural knowledge needed to live as an ordinary member of a society. It includes, inter alia, the ability to speak the native language, knowing how often to wash, how close to walk to others on crowded and empty pavements (sidewalks), and so forth. Without ubiquitous expertise one cannot engage in the social interaction needed to acquire more specialist abilities and every competent member of any society will possess huge amounts of ubiquitous expertise.[9] Using these three categories, we now turn to the analysis of textual data generated during Imitation Games.

Quantitative Analysis of Textual Data

Columns 2 and 3 in Table 1 indicate two kinds of quantitative content analysis. In column 2, the analysis focuses on the form of the dialogues and requires only a minimal understanding of their content. For example, we have coded questions according to the type of information Interrogators ask for, leading to the following categorization of question types which are applicable across all Imitation Game topics:

1) **Biographical**: The respondent is asked to recount a story or detail from their own life, or discuss how they have handled a particular situation or experience.
2) **Preference**: Distinguished primarily by reference to likes and dislikes.
3) **Opinion**: Questions which ask the respondents what they think about a particular issue, person or situation.
4) **Knowledge**: Either asks directly what the respondents know about a subject or requires respondents to understand specialized terms or jargon that are used without explanation in the question
5) **Situational**: Hypothetical questions such as 'If you had X what would you do?' or 'What advice would you give to someone if…'

The analysis compares the distribution of question types asked by different classes of player in different locations or on different topics. To illustrate this kind of analysis we turn to Project 2: Imitation Games on the topic of male sexuality played in the Cardiff (UK) and Wroclaw (Poland) in 2013 (see Table 2 for details). In each case, one round of Step 1 games was played with individual judges. These were followed by an asymmetrical version of Step 2 in which only straight participants were recruited. At Step 4, a new sample of gay Judges was recruited to judge the transcripts.

Initial analysis shows that the pass rates for Pretenders in Cardiff and Wroclaw are more similar than expected, at 65% and 58% respectively (($t(389)=-0.848$, $p=0.40$)) but analyzing the distribution of question types generated in each location suggests a possible explanation. As shown in Table 4, the proportion of questions coded as either 'preference' or 'knowledge' questions within the UK (Cardiff) data is approximately twice that found in games played in Poland (Wroclaw). In contrast, Polish sexuality games contain almost twice the proportion of

‘biographical’ questions than are found in UK games. This would suggest that Polish gay men consider personal history and life experiences as more significant in defining their sexuality, while in the UK players are more focused on the choices they are able to make as gay men. In each case, however, the ability of the local straight male population to provide plausible answers appears comparable, showing that apparently similar pass rates can be built on very different foundations.

Table 4 (Question Types) about here

Column 3 introduces analysis of the substantive content of the textual data, something that requires more in the way of cultural understanding. For example, thematic coding can be used to measure the prevalence of particular subjects in Imitation Games. Unlike coding question types, where the coding scheme can apply across different topics, thematic codes are derived directly from the data and can vary across topics. Once question themes have been coded, the relative proportions of each theme, or group of cognate themes, can be analyzed in the same way as question types. For example, in the sexuality Imitation Games in Cardiff and Wroclaw described above the topics ‘sex’ and ‘coming out/being out’ are two of the most commonly applied codes (which may be unsurprising) but the way in which they are applied suggests that the experience of homosexuality in the two locations is quite different, as shown in Table 5.

Table 5 (Question topic) about here

In UK Imitation Games there were very few questions about the experience of coming out or being out. In contrast, over 20 per cent of coding for the Wroclaw questions related to

questions addressing one or more of these topics. A similar difference occurs for questions about sex, where Interrogators in Cardiff asked approximately triple the proportion than did Interrogators in Poland. It is also possible to make connections between thematic coding and question types, as 'coming out/being out' questions seem likely to fit the 'biographical' question type, which was more common in Poland, while 'sex' questions are more typically about favored practices and are typically coded as 'preference'. This, in turn, sheds more light on the ways in which similar pass rates are constructed in different places from different kinds of knowledge and highlights how the construction of social identities like is flexible and dependent on locality and temporality. One of the strengths of the method is its ability to capture local, contemporary experiences and generate data to compare such cross-cultural differences.

What kind of cultural understanding is needed to perform this analysis? Our claim is that ubiquitous expertises are sufficient for researchers to recognize a question type so long as questions are presented in the researcher's native language. Where the native language in which the game is played is not the researchers', then either the questions must be translated or a native speaker must do the coding. We are, therefore, assuming there is a 'hyper-ubiquitous expertise' that crosses the cultures in which we have conducted our research such that the same classifications would be produced whether they were translated or coded directly from the native languages.

In the same way we take mathematics and statistics to be uniform across these cultures, though here the argument is different: mathematical culture is uniform because all those who engage in mathematical practices are socialized in the same way wherever their schools and

universities are located.[10] In contrast, the uniformity of question types has its roots in the cultural mixing of different societies whose languages share common ancestry.

The second and third rows of columns 2 and 3 show the application of these kinds of content analysis to other aspects of the dialogues, with particular attention paid to the relationship between Judge/Interrogator and Non-Pretender. The next row applies the analysis to the mini-focus groups. Here the mixing of methods is unusual and striking. Focus groups are most often thought of as generating qualitative data but because a single large scale experiment will give rise to at least 20, independent, mini-focus groups, we should be able to count how often certain remarks or themes occur in each group and make comparisons across nations and with our own analyses of pass rates.[11] The levels of cultural understanding required to make the two types of quantitative sense out of the focus groups is deeper than that required for the typed dialogues because it is harder to understand spoken conversation than written conversation. Finally, the dialogues and transcripts can be analyzed in the same ways as other textual and linguistic data, revealing further insights such as the linguistic patterns favored by different groups of participants.

Qualitative Analysis of Textual Data

Column 4 in Table 1 takes us to the deepest level of cultural understanding – participant comprehension (Collins, 1984) – which is achieved along with the acquisition of interactional expertise in the most specialized aspects of the cultures being explored. Only by immersing oneself deeply into the native culture could one hope to accurately predict the strength of the taboos that would cause certain questions not to be asked at all or to understand the cultural work performed as participants mobilize particular aspects of their identity within an Imitation Game. One cannot gain this level of expertise from playing Imitation Games or

reading the dialogues or even listening to the discussions of the mini-focus groups; one must either be a member of the community in question or spend years gaining an understanding of them through immersion in their linguistic discourse. If one does gain this level of expertise, however, then one can understand the vocabulary of intention pertaining to that social group and why the Game is being played one way rather than another.

The Imitation Game transcripts, and those created from mini-focus groups, can also be analyzed in any of the other ways that textual data is analyzed throughout the social sciences – more or less contextual, more or less interpretative, more or less theoretical, and more or less political and so on. As with interpretative participatory research more generally, the great problem is promulgating the results of the participatory work done under the heading of column 4: how can deep experiences of social pressures be conveyed to those who have not themselves experienced life and discourse in the sub-groups in question. The answer is that it is in principle impossible, though in practice, the techniques of the 'writer' can convey something of the meaning of these lives, *illustrated* by the kinds of questions that are or are not asked and from phrases used in the Imitation Game dialogues or the mini-focus group discussions.

Future Directions

The primary aim of the research described here is to develop the Imitation Game method. As a consequence, all the large-scale games reported have been played by undergraduates in university settings. This has minimized cost and the practical problems of recruitment and information literacy but does mean that data is biased toward a relatively educated and wealthy demographic. This is not a fundamental problem for the Imitation Game method, only for the particular samples we have been able to recruit, and the extent to which the

results generalize. The solution is to repeat the research with more representative samples and we are confident that the protocols we have developed would make this possible.

What does generalize is the analysis of cultural competences and here the Imitation Game reveals the challenge faced by all mixed methods research: what is the correct trade-off between 'breadth' (pragmatism) and 'depth' (specialization)? In the Imitation Game, the dilemma arises later in the process than with other methods as the players supply the local cultural competences needed to generate the data, with the expertise of the researchers only being called upon as the analysis moves across Table 1.

Whatever the question, all analysis will require the researchers to have some understanding of whether unexpected results represent a failure or a novel finding. This dilemma is perfectly illustrated by the studies of sexuality in Wroclaw and Cardiff (Project 1), where the initial hypothesis was not supported but more detailed analysis suggested a range of possible explanations that are the subject of further, in-depth research including the extent to which Wroclaw is typical of Poland, students are typical of the general population, and whether 'importing' questions from the 'Cardiff' games produces a different pass rate. In these early stages the researcher is subject to the experimenter's regress as it is impossible to know if the pass rate has been measured correctly without some prior agreement on what the correct measurement should be.[12] In these cases, the role of the qualitative analysis, and the increasingly rich cultural expertise it demands, is to work alongside replication studies to find ways of assessing the credibility of the measured pass rate(s).

Where the initial quantitative analysis supports the research hypothesis – as happened with the gender Imitation Games in Granada (Project 2) – the analysis of question types and themes has a more straightforward role. Now fine-grained analyses can shed light on how

participants define their gender identities, what kinds of questions Pretenders answer well and what kinds of questions discriminate between the two groups. Analyzing the range of question types and themes reveals something of the diversity of the target culture with respect to the Pretender population and identifying the kinds of knowledge Pretenders fail to produce indicates where the two groups remain distinct and separate from each other. The results can be compared and contrasted with existing studies using other methods and, depending on the topic being researched, may also be used to design interventions that reduce the 'gap' between the two groups.

Looking to the future, we see three main areas where Imitation Game research could be developed and extended. These are:

1. Methodological research to improve the protocols and explore the effects of different variants. For example, should Judges at Step 4 give decisions on each question independently; what data should be collected from Judges to record their reasons; where participants are difficult to recruit, is it better to play Step 1 as 8 groups of 3 players or as 24 individuals?
2. Substantive research to replicate existing studies or extend the research by recruiting more representative samples. One by-product of this work would be a corpus of data that will provide a benchmark for new studies and a longitudinal, cross-cultural resource for secondary analysis.
3. Applied research in which Imitation Games are used to explore the extent to which service providers understand the needs and experiences of their clients and, if necessary, develop training interventions to improve their knowledge. The intuition is that the active engagement required by the Imitation Game will produce a deeper form

of learning that may have particular value where empathy with clients or service users is required.

In terms of mixed methods, we see the Imitation Game as contributing to development of both theory and practice. Here, the insight, reflected in typology of approaches set out in Table 1, is that although different methods make different demands on the researcher, all methods require at least some context-specific cultural knowledge.

Summary

In this paper we have described a new research method for measuring and researching cultural competence and examined how the data generated can be analyzed. Like other basic methods – the questionnaire, the focus group, mass observation – the Imitation Game can be used in many ways and for many purposes. By way of illustration, we have described its use to test the competence of individual ethnographers, of small groups with different degrees of exposure to the linguistic discourse of other groups, and of large samples of students pretending to be members of different social groups in their society.

We have also used the Imitation Game to discuss wider questions about the relationship between the quantitative and the qualitative and to distinguish between four kinds of analysis. The first kind is the most quantitative but even this depends on a deep cultural understanding that is, uniquely in the case of the Imitation Game, supplied by the players. The other three kinds of analysis deal with the textual data more directly, with the cultural knowledge of the researcher becoming ever more crucial. Thus, the content analysis described in Column 3 of Table 1 rests on a deeper understanding of culture than that of Column 2 because the substance of the questions, including the meaning of culturally specific words and slang, has to be understood, not just the type of information being asked for. The interpretive analysis

described in Column 4 requires a still deeper understanding as it refers not only to what was made explicit during the Game, but what could have been and was not.

These cultural competences are both society-wide and locally-specific. Though columns 2, 3 and 4 appear distinct in Table 1, there is inevitably some interaction between them. Sometimes question types can only be recognized with more cultural understanding than is intimated at the foot of Column 2 and sometimes what we need for column 3 requires something of the competences described at the foot of column 4. This is possible because when we do our work we are not completely isolated from native members of the cultures we are working in, and during casual discussions in the laboratory and the coffee bar, some vestiges of local native competence are acquired. Similar problems no doubt arise in other mixed methods studies.

In our research we are now trying to formalize the process of developing inter-cultural understanding between researchers by adopting an apprenticeship model in which a single expert in the nature of the characteristic develops codes in collaboration with those having the ubiquitous expertises of the foreign societies. The hope is that this will lead to mutual socialization into each others' expertises as the joint coding scheme is developed. More importantly, the apprenticeship model reveals the intimate relationship between qualitative and quantitative analysis: it is not that one can add a quantitative measure to a qualitative measure, or triangulate one with the other, it is that quantitative measures require careful, preparatory, qualitative work if they are to be reliable. The Imitation Game simply makes this more visible as 'quant' and 'qual' approaches are part of *the same* 'meshed' method. Should any reader of this journal wish to try it, we will be happy to give all the help we can from

discussions of experimental design, through invitations to training workshops and conferences and access to the specialist software needed to run the large scale games.[13]

Glossary of terms

Interactional Expertise	The ability to speak the language associated with a practice, though may be learnt without necessarily experiencing the practice; a key concept in SEE
SEE	Studies of Expertise and Experience, a sub-field in Science and Technology Studies (STS), that pays particular attention to the role of socialization and tacit knowledge
SSK	Sociology of Scientific Knowledge, a founding element of STS, in which it is shown that scientific knowledge can – and indeed should – be analyzed in the same way as any other belief system.
Target Expertise	The expertise possessed by the Judge and Non-Pretender in an Imitation Game; the expertise the Pretender attempts to reproduce.
Turing Test	A natural language test of machine intelligence proposed by Alan Turing; the idea is that if machine responses are indistinguishable from human ones, then the machine must be classed as intelligence.

References Cited

Bloor, D. (1973). Wittgenstein and Mannheim of the Sociology of Mathematics. *Studies in the History and Philosophy of Science*, *4*, 173-91.

Bryman, A. (2012). *Social Research Methods (4th edition)*. Oxford: Oxford University Press.

Bryman, A. (2006). Integrating Quantitative and Qualitative Research: How is it Done? *Qualitative Research*, *6*, 97-113.

Calderon, J.L., Baker, R.S., and Wolf, K.E. (1999). Focus Groups: A Qualitative Method Complementing Quantitative Research for Studying Culturally Diverse Groups. *Education for Health*, *13*, 91-95.

Collins, H.M. (2013). *Gravity's Ghost and Big Dog: Scientific Discovery and Social Analysis in the Twenty-First Century*. Chicago, IL: University of Chicago Press.

Collins, H.M. (2012). Performances and Arguments. *Metascience*, *21*(2), 409-418.

Collins, H.M. (2011). Language and Practice. *Social Studies of Science*, *41*(2), 271-300.

Collins, H.M. (2010). *Tacit and Explicit Knowledge*. Chicago, IL: The University of Chicago Press.

Collins, H.M. (2004a). Interactional Expertise as a Third Kind of Knowledge. *Phenomenology and the Cognitive Sciences*, *3*(2), 125–43.

Collins, H. M. (2004b). *Gravity's shadow: the search for gravitational waves*. Chicago: University of Chicago Press.

Collins, H. M. (1990). *Artificial Experts: Social Knowledge and Intelligent Machines*. Cambridge Mass: MIT Press

Collins, H. M. (1985/1992). *Changing Order: Replication and Induction in Scientific Practice*. Beverley Hills and London: Sage [2nd edition, 1992, Chicago: University of Chicago Press]

Collins, H.M. (1984). Concepts and Methods of Participatory Fieldwork. in C. Bell, & H. Roberts (Eds.), *Social Researching* (pp. 54-69). Henley-on-Thames: Routledge.

Collins, H. M., & Evans, R. (2015). Expertise revisited, Part I—Interactional expertise. *Studies in History and Philosophy of Science Part A*. [available online at http://doi.org/10.1016/j.shpsa.2015.07.004]

Collins, H.M., & Evans, R.J. (2014). Quantifying the Tacit: The Imitation Game and Social Fluency. *Sociology*, *48*, 3-19.

Collins, H.M., & Evans, R.J. (2007). *Rethinking Expertise*. Chicago, IL: University of Chicago Press.

Collins, H.M., & Evans, R.J. (2002). The Third Wave of Science Studies: Studies of Expertise and Experience. *Social Studies of Sciences, 32*, 235-296.

Collins, H.M., Evans, R.J., Ribeiro, R., & Hall, M. (2006). Experiments with Interactional Expertise. *Studies in History and Philosophy of Science, 37*, 656-74.

Collins, H.M., & Kusch, M. (1998). *The Shape of Actions: What Humans and Machines Can Do*. Cambridge, MA: MIT Press.

Creswell, J.W., & Plano Clark, V.L. (2011). *Designing and Conducting Mixed Methods Research* (2nd ed). Thousand Oaks, CA: Sage.

Evans, R.J., & Crocker, H. (2013). The Imitation Game as a Method for Exploring Knowledge(s) of Chronic Illness. *Methodological Innovations Online, 7*.

Forman, P. (1971). Weimar Culture, Causality and Quantum Theory, 1918-1927: Adaptation by German Physicists and Mathematicians to a Hostile Intellectual Environment. In R. McCormack (Ed.), *Historical Studies in the Physical Sciences, No 3* (pp 1-115). Philadelphia, VA: University of Pennsylvania Press.

Giles, J. (2006). Sociologist fools physics judges. *Nature 442*, 8.

Hartland, J. (1996). Automating Blood Pressure Measurements: The Division of Labour and the Transformation of Method. *Social Studies of Science, 26*. 71-94.

Hodges, A. (1985). *Alan Turing: The Enigma of Intelligence*. London: Unwin.

Johnson, R. B., Onwuegbuzie, A.J., & Turner, L.A. (2007). Toward a Definition of Mixed Methods Research, *Journal of Mixed Methods Research*, 1(April): 112-133.

Kitzinger, J. (1995). Qualitative Research: Introducing Focus Groups. *BMJ, 311*, 299-302.

Mackenzie, D. (1981). *Statistics in Britain 1865-1930*. Edinburgh: Edinburgh University Press.

Mackenzie, D. (2001). *Mechanizing Proof: Computing, Risk, and Trust*. Cambridge, MA & London: MIT Press.

Morgan, D.L. (1996). Focus Groups. *Annual Review of Sociology, 22*, 129-152.

Morgan, D.L. (1995). Why Things (Sometimes) Go Wrong in Focus Groups. *Qualitative Health Research, 5*, 516 – 523.

Turing, A.M. (1950). Computing Machinery and Intelligence. *Mind, LIX, 236*,433-460.

Wehrens, R. (2014). The Potential of the Imitation Game Method in Exploring Healthcare Professionals' Understanding of the Lived Experiences and Practical Challenges of Chronically Ill Patients. *Health Care Analysis*, 1–19.

Tables and Figures

NUMERICAL OUTPUTS	TEXTUAL OUTPUTS DIALOGUES, REASONS, MINI-FOCUS GROUPS		
QUANTITATIVE ANALYSIS	QUANTITATIVE ANALYSIS OF TEXTUAL DATA		QUALITATIVE ANALYSIS
1	**2**	**3**	**4**
Overall quantitative analyses Pass rate comparisons/ statistical confidence	Content analysis of proportions of question types	Content analysis of topics included (or not) in questions e.g. questions relating to sexual practices	Participant comprehension and interpretive analysis (e.g. how identities are constructed)
Consistency analysis Judge comparisons etc.	Content analysis of types of reasons	Analysis of how reasons relate to unwillingness to ask delicate questions	Discourse analysis of text produced by players in each role
Cultural experiments Experiments with non-expert Judges/interrogators	Do Non-Pretender answers reflect Judge question types?	Do Non-Pretender answers reflect delicacy of Judges' questions?	'Conversation analysis' of text produced by players in each role
Protocol experiments Dialogue-set comparisons (groups vs. individuals)	Mini-focus group discussions of Judges, Pretenders and Non-Pretenders	Do mini-focus groups exhibit the same delicacies?	Conversation analysis of naturally occurring talk in focus discussions
Internal single game analysis Guess reversals/ Confidence changes	Analysis of turn characteristics (e.g. word count by turn, concordance) by player characteristics or role	Analysis of turn characteristics (e.g. use of qualifiers, don't know etc.) by player characteristics or role	Analysis of turn characteristics (e.g. metaphors and cultural references) by player characteristics or role
DEPTH OF RESEARCHERS' CULTURAL UNDERSTANDING →			
Researchers need no cultural understanding of actors' world but universal numerical skills	**Researchers need hyper-ubiquitous language understanding**	**Researchers need local language understanding**	**Researchers need interactional expertise in target culture**
← PLAYERS SUPPLY DEEP/LOCAL CULTURAL UNDERSTANDING →			

Table 1: Types of analysis of Imitation Games

Place	Topic	Step 1, 2 or 4	Sample size	Hypothesis / Innovation being tested
Granada (1)	Gender	1 (individual) 1 (group) 2 4	20 male, 20 female 40 male, 40 female 247 male, 290 female 65 male, 65 female	Group games have lower pass rate than individual games
Granada (2)	Gender	1 (group) 2 4	40 male, 40 female 135 male, 144 female 40 male, 40 female	Replication of group result obtained in Granada (1)
Cardiff	Sexuality	1 2 4	12 (gay); 12 (straight) 201 (straight) 40 (gay)	Pass rate for straight Pretenders will be higher than in Wroclaw.
Wroclaw	Sexuality	1 2 4	17 (gay); 17 (straight) 221 (straight) 40 (gay)	Pass rate for straight Pretenders will be lower than in Cardiff

Table 2: Characteristics of Imitation Games reported

		Female Judge	**Male Judge**
Individual	Wrong	30	49
	Don't Know	56	72
	Right	81	82
	Pass Rate (%)	**68**	**84**
Group (1)	Wrong	7	15
	Don't Know	15	19
	Right	59	51
	Pass Rate (%)	**36**	**57**
Group (2)	Wrong	12	19
	Don't Know	27	29
	Right	86	64
	Pass Rate (%)	**41**	**60**

Table 3: Distribution of Results in Granada Imitation Games on Gender

Question type	**Cardiff Judge**	**Wroclaw Judge**
Biographical	32%	59%
Knowledge	33%	16%
Preferences	35%	17%
Opinion	28%	26%
Situational	10%	6%
Total (n)	**92**	**122**

NB: Columns sum to more than 100 because questions can be coded to multiple question types

Table 4: Question Type Coding, Sexuality Imitation Games

Topic	**Cardiff**	**Wroclaw**
Being Out	0%	3%
Coming Out	6%	18%
Sex	23%	7%
Total Questions (n)	**92**	**122**

Table 5: Question Topic Coding, Sexuality Imitation Games

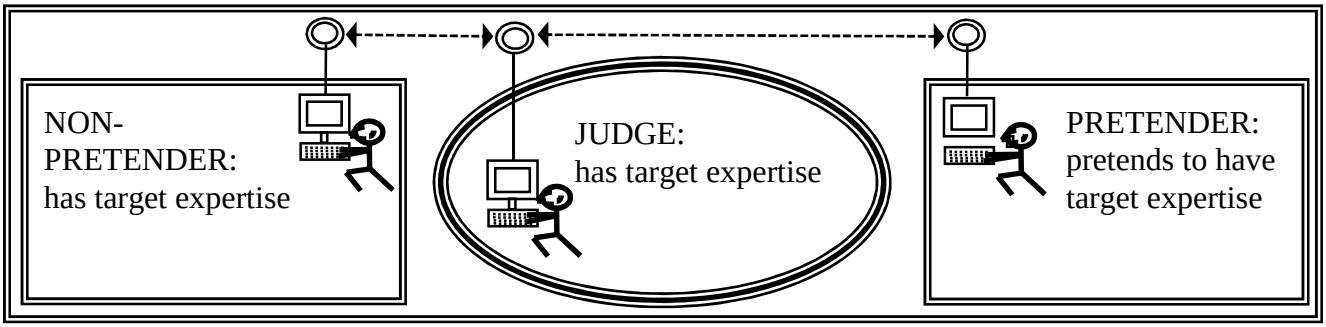

Figure 1: Schematic Representation of Basic Imitation Game

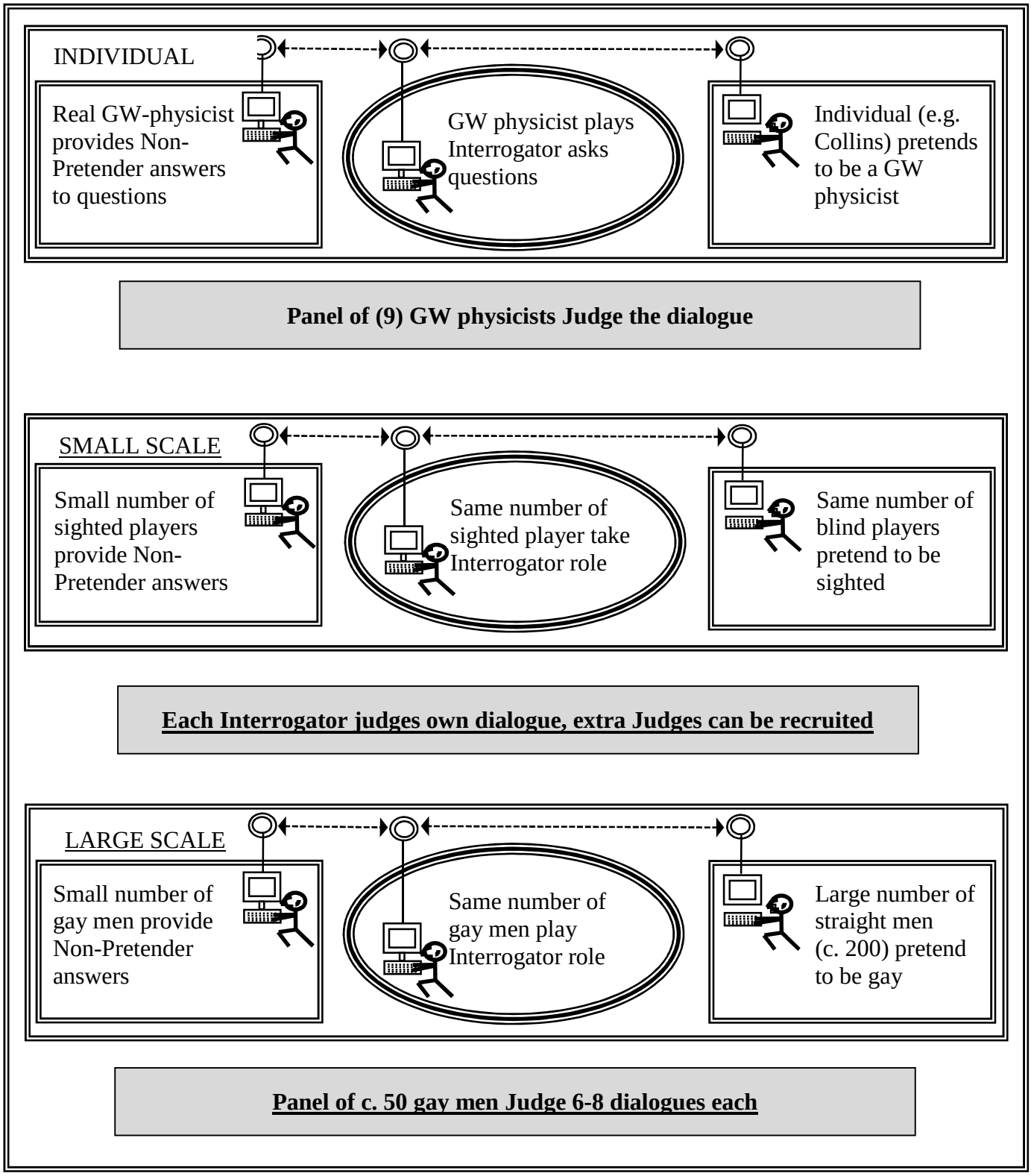

Figure 2: Versions of the Imitation Game

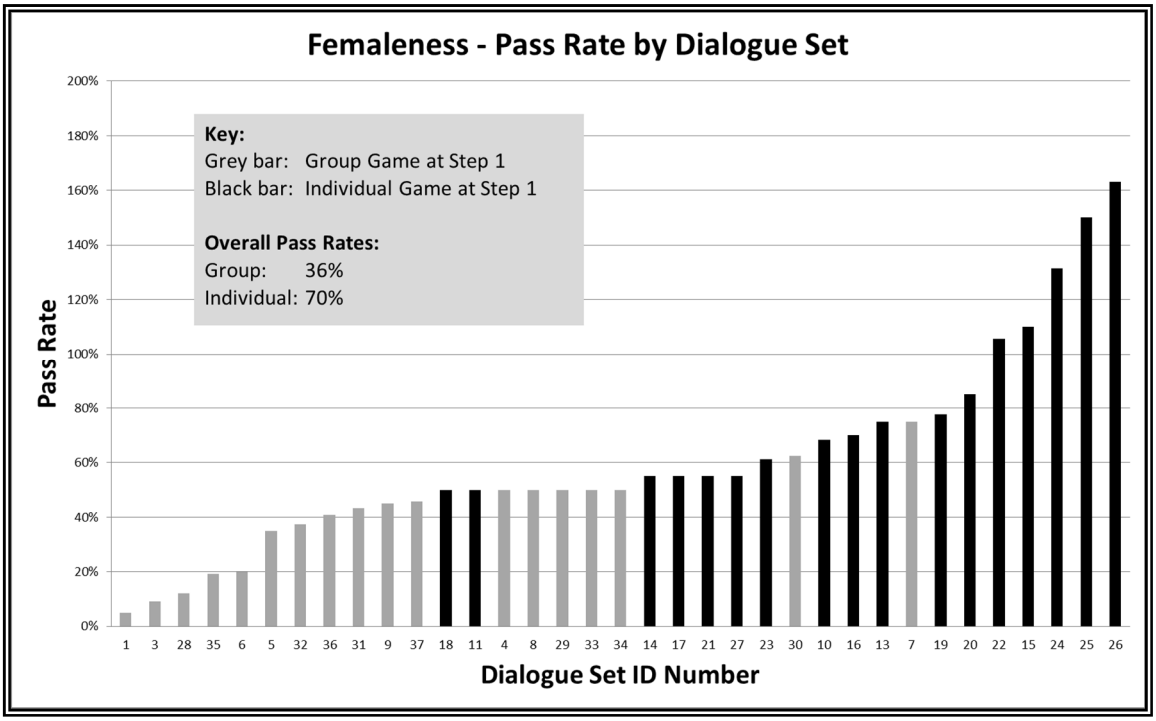

Figure 3: Pass Rates for Gender Imitation Games by Dialogue-Sets

Endnotes

---

[1] Collins started to develop the Imitation Game in 1996 with Hall writing the necessary computer software and continuing to be the programmer to this day; the small-number games described below first met with success in the mid-2000s and were run by Collins and Evans, while the large scale game has been under development with Weinel since 2011 with Lyttleton-Smith joining later and with contributions from Bartlett. Our large scale research has been made possible by European Research Council Advanced Grant (269463 IMGAME, 2011-2016, €2,260,083 *`A new method for cross-cultural and cross-temporal comparison of societies'*) and a European Research Council Proof of Concept Grant (297467 IMCOM, 2012-2013, €150,000 *`IMGAME Commercial'*).

[2] For this reason, there are no standardized instructions as each use needs to be tailored to the local context. That said, there is a well-developed body of knowledge about the principles the method embodies and an ever-increasing set of exemplars on which new research can build.

[3] We have also used the Imitation Game to investigate other topics including religious understanding and racial identity. These results will be reported in other publications.

[4] For details on how to access the software please contact Professor Harry Collins (CollinsHM@Cardiff.ac.uk) or Professor Robert Evans (EvansRJ1@Cardiff.ac.uk)

[5] Forman (1971) indicates why it can be that in certain epochs populations some populations will exhibit more uncertainty than others – the pass rate formula eliminates this source of systematic error.

[6] That there are four Steps in the most elaborated form of the Imitation Game and four columns in Table 1is coincidental: the columns do not follow any temporal sequence and are merely an analytic device designed to highlight the demands of different approaches.

[7] For a related idea – the 'proxy stranger' – see Collins and Kusch (1998) and Hartland (1996).

[8] Analysis of video recordings of the mini-focus groups may also be interesting.

[9] See also Collins (2010) discussion of expertises in terms of '*Tacit and Explicit Knowledge'*.

[10] SSK shows that science, and even mathematics, is not as universal as is usually thought but this does not matter for the argument presented here. See e.g. Bloor (1973), Mackenzie (1981, 2001) and statistical practices – eg significance level acceptable for publication -- vary hugely across the sciences.

[11] Usually, where focus groups are concerned, quantitative analysis is either not mentioned or actively discouraged. See e.g. Kitzinger (1995), Calderon et al (1999), Bryman (2012). Morgan (1995,1996) is unusual in that he does not rule out quantitative analysis of focus group data despite the challenges it raises.

[12] Ironically, the experimenter's regress was first identified by one the authors of this paper – Collins (1985/1992) – and is one of his most well-known contributions to the STS literature.

[13] An app for Apple or Android devices is available and can be used to play informal 3-player games in the same location or remote locations. Sample instruction sheets, a video explaining the main features of the Game, and other information relating to research program are available from http://blogs.cardiff.ac..uk/imgame