Evaluating the Imitation Game as a method for comparative research: a replication study using Imitation Games about religion

To be published in:
International Journal of Social Research Methodology

Abstract

The Imitation Game is a new method and, as such, it is important to show that its results are plausible and replicable. We tested this by conducting Imitation Games on religion in a range of European countries, returning approximately 12 months later to repeat the research. The idea was that non-Christian members of strongly Christian countries would find it easy to pass as members of the practicing majority because Christian beliefs and practices would be ubiquitous. In more secular countries, the expectation was that non-Christians would find it harder to pass as Christian because religious practices are less visible. We show that, despite some anomalous results, the data are consistent with expectations derived from survey data and that the claim to have replicated the results can be supported. We also suggest that our experiences show that questions of replication in the social sciences cannot be resolved by statistical meta-analysis alone.

Authors

Robert Evans, School of Social Sciences, Cardiff University (EvansRJ1@Cardiff.ac.uk)
Harry Collins, School of Social Sciences, Cardiff University
Martin Weinel, School of Social Sciences, Cardiff University
Hannah O'Mahoney, Cardiff University
Jennifer Lyttleton-Smith, School of Education and Social Policy, Cardiff Metropolitan University
Rik Wehrens, School of Health Policy and Management, Erasmus University Rotterdam

Publication History

Original version submitted to journal	28 November 2020
Revised version submitted to journal	6 July 2021
Accepted for publication:	21 September 2021

Evaluating the Imitation Game as a method for comparative research: a replication study using Imitation Games about religion

The Imitation Game is a new method and, as such, it is important to show that its results are plausible and replicable. We tested this by conducting Imitation Games on religion in a range of European countries, returning approximately 12 months later to repeat the research. The idea was that non-Christian members of strongly Christian countries would find it easy to pass as members of the practicing majority because Christian beliefs and practices would be ubiquitous. In more secular countries, the expectation was that non-Christians would find it harder to pass as Christian because religious practices are less visible. We show that, despite some anomalous results, the data are consistent with expectations derived from survey data and that the claim to have replicated the results can be supported. We also suggest that our experiences show that questions of replication in the social sciences cannot be resolved by statistical meta-analysis alone.

Keywords: Imitation Game, interactional expertise, replication, comparative research

Evaluating the Imitation Game as a method for comparative research: a replication study using Imitation Games about religion

Introduction

The replication crisis sparked by John Ioannidis's infamous claim that 'most published research findings are false' (Ioannidis, 2005) typically turns on the meaning of p-values and significance testing and associated assumptions about sampling and measurement (Gorard, 2016, 2019). Here we take a different approach and examine, in a reflexive way, how we determined if a result had been replicated. The problem is doubly difficult in the case reported here as the novel methodology means there is no readily available comparator data against which to judge results. It is, therefore, an example of the 'experimenter's regress' (Collins, 1992) and the paper explores the linked problems of (a) determining, a priori, what the 'right' answer might look like and (b) deciding whether results were close enough to this to count as 'the same' (c.f. Kuhn, 1961).

Our new research method – the Imitation Game – starts from the sociological axiom that everything we know is a result of our socialisation. We distinguish between socialisation through direct participation and socialisation that is acquired indirectly through linguistic interactions, and use the Imitation Game to explore what kinds of knowledge can be gained through language alone (Collins et al., 2006, 2017, 2019; for more on the Imitation Game see Collins & Evans, 2014; Evans, Collins, Weinel, et al., 2019; Evans & Crocker, 2013). The novel features of the Imitation Game are twofold:

(1) it maps the distribution of knowledge rather attitudes, examining what people know about a social group without being a member of that group or taking part in its practices.

(2) the method is designed to be 'bottom-up', putting participants at the centre of the research and enabling them to generate data that reflects local knowledge, traditions and priorities.

The aim of the research described in this paper was to conduct the first large-scale tests of the Imitation Game as a comparative method by first calibrating results against existing survey data and then seeking to replicate these results through repeated fieldwork visits.

The remainder of the paper is structured as follows. First, we explain the Imitation Game method in more detail, setting out the theory that informs its design and the way it is used in practice. Next, we describe a series of studies in which we explored knowledge about religion (specifically, the locally dominant form of Christianity) across a number of European countries, ranging from those traditionally seen as very religious (e.g. Italy and Poland) to those seen as more secular (e.g. Norway and Finland). We also explain how we classified each country as 'Christian' or 'Secular' and how we derived the hypotheses that framed our work. Finally, we turn to the data generated. We explore how we determined if a replication had been successful and, if not, what this meant for the Imitation Game and the challenge of replication more generally.

Imitation Game: theory and practice

The Imitation Game was originally developed to test the idea of interactional expertise, which is explained in more detail below, but it can be used to explore the nature of groups and group membership more generally. For example, it can be used to explore the uniformity or diversity within a group or how widely knowledge of a particular group's experience is shared, thus shedding light on both the character of that group and

the wider society of which it is a part. Used longitudinally, the Imitation Game can be used to track changes in both the content and distribution of this knowledge, again reflecting changes in both the experiences of the target group and their relationship with the wider society.

Other uses of the method include examining the strategies used by different players to create questions, to generate answers to these questions and to evaluate these answers. It is also possible to supplement Imitation Game data with demographic and/or psychometric data to explore whether specific player characteristics affect the outcome. Alternatively, the analysis might focus on the whole corpus of data, mapping the presence and absence of different themes or examining the use of language. In other words, much like the survey or interview, the Imitation Game is a generic method that can be adapted to a wide variety of research designs and questions, with qualitative or quantitative data foregrounded as appropriate. In what follows, we focus on the use of quantitative data to test theory that informed the development of the method. The rationale for this is to demonstrate that the range of potential uses hinted at above is built on a firm foundation.

Theoretical foundations: interactional and contributory expertise

Members of a social group or culture who have been successfully socialised share what the philosopher Ludwig Wittgenstein called a form of life (Winch, 1958; Wittgenstein, 1953). Sharing a form of life means acquiring the set of tacit and explicit knowledge used by members of the group to coordinate and moderate their actions. The inclusion of tacit knowledge is crucial as this can only be gained through social interaction and is, therefore, peculiar to that group (Collins, 2010; Evans, 2008).

There are two ways in which this socialisation can take place. The first is via full and active participation in the group's activities, such that tacit knowledge becomes embodied in the person (Dreyfus, 2004). The problem with this view is that it is too restrictive. If direct, personal experience were the only route to expertise, it would be impossible for ethnography and anthropology to succeed as researchers would have to experience every aspect of a culture for themselves in order to understand it.

As an alternative to this way of thinking, we distinguish between contributory and interactional expertise (Collins et al., 2016; Collins & Evans, 2002, 2007, 2015). Contributory expertise corresponds to the embodied form of expertise described above and describes the abilities and knowledge of an individual who has been fully socialised into a particular group. Interactional expertise, which is the new concept, refers to expertise in the language that contributory experts use to describe their practices. Interactional expertise is similar to contributory expertise in that it can only be gained through interaction with contributory experts but differs in that it does not require any practical experience (Collins, 2004, 2011). Ethnography and anthropology are thus possible as researchers can gain interactional expertise through their conversations with participants but do not have to engage in the associated practices: criminologists do not have to commit crimes; sociologists of religion do not have to be devout believers; and it does not matter that sociologists of childhood are no longer children.

Returning to the idea of the Imitation Game, by examining both the content and distribution of interactional expertise, the Imitation Game provides a new way of mapping the interactions between members of social groups. The more the actions and beliefs of one social group are visible to and engaged with by members of a different group, the more likely it is that the second group will develop the interactional expertise needed to understand the experiences of the first. The extent to which this occurs

reveals something about the relationship between the groups, whilst the content of that interactional expertise provides an insight into the life of the target group. It is this argument – that interactional expertise (a) has a distribution and (b) that this distribution can be measured by the Imitation Game – that the research described below was intended to test.

Playing the Imitation Game: roles and data

The Imitation Game is based on the Turing Test (Turing, 1950), in which a human judge asks questions of a human and a computer and must decide which answers come from the computer and which from the human. Turing's claim was that, if the Judge cannot distinguish between the two sets of answers, then the computer should be classed as intelligent. Crucially, the Turing Test does not require that the computer have a body or do practical things in the way a human does; it is based solely on the convincing and contextual use of language (Collins, 1990, 2018). Re-framed in terms of contributory and interactional expertise, we say that, in the Turing Test, both the human players have contributory expertise (i.e. practical and linguistic fluency) whilst the computer needs only interactional expertise (i.e. linguistic fluency).

In our research, we take the parlour game that inspired Turing and develop a more formal set of protocols so that it can be used for social research. The basic setup, what we now call the 'Classic Imitation Game' and consists of three players:

- **Interrogator/Judge**. This player must be a contributory expert – that is a member of the target group – and plays two roles: an **Interrogator** who asks questions and a **Judge** who decides which answer comes from which player.
- **Non-Pretender**. This player is also a contributory expert and answers the Judge's questions by referring to their own experiences and knowledge.

- **Pretender**. This player attempts to answer the questions posed by the Interrogator as if they were a member of the target group (i.e. as if they were a contributory expert). If the Pretender has interactional expertise, then the Judge should find it difficult to work out which answers come from the Pretender which from the Non-Pretender. In contrast, if the Pretender does not have interactional expertise, then the Judge should find it relatively easy to identify the players.

Each Imitation Game proceeds with the Interrogator setting a question, the Pretendent and Non-Pretender providing answers and the Judge then attempting to determine which answer came from which player. This generates a set of qualitative and quantitative data consisting of:

- **Questions**: these indicate the topics that Interrogators think differentiate their group from the Pretender group (i.e. the Non-Pretender will know the answer but the Pretender will not)
- **Answers from Non-Pretender**: these provide an indication of the range of experiences within the target group. Where the group's experiences are very homogeneous, Non-Pretender answers will be very similar. Where the group permits diversity, a wider range of answers are possible.
- **Answers from Pretenders**: these indicate the extent to which the Pretender population has the relevant interactional expertise. Where they do, Pretender and Non-Pretender answers will be equally plausible. Where they do not, Pretender answers will be deficient in some way.
- **Judgements**: these are available for individual questions or the set of questions as a whole and consist of an **identification** (e.g. Player 1 is the Pretender),

which might be right or wrong, an indication of the Judge's **confidence** in that identification on a scale of 1-4, and the **reason** for that decision.

This basic format can be adapted to suit different needs and resources. Of particular relevance here is the development of the method to use large samples needed for quantitative analysis (Collins et al., 2017, 2019). Other developments include: the use of small groups, rather than individuals, to play the three roles in the Classic version of the Game (Evans, Collins, Weinel, et al., 2019), using the data to explore how Judge decisions are made (Arminen et al., 2018; Collins, 2016; Segersven et al., 2020) and using the Imitation Game as an intervention to prompt dialog and reflection in a larger project (Wehrens, 2014, 2018).

When analysing the results, it is possible to focus on either the qualitative or quantitative elements or both (Collins et al., 2017). When looking at the quantitative results, the success of Judges is measured by the Identification Ratio (IR), which is calculated using the formula:

$$IR = (Right - Wrong) \div (Right + Wrong + Don't\ Know)$$

where:

- Right = Number of correct identifications with confidence rating of 3 or 4
- Wrong = Number of incorrect identifications with confidence rating of 3 or 4
- Don't know = Number of identifications with a confidence rating of 1 or 2

In what follows, however, we are more concerned with the success of Pretenders as that provides a more direct way of talking about the distribution of interactional expertise. The success of Pretenders is called the pass rate and is given by:

$$Pass\ Rate\ (\%) = 1 - Identification\ Ratio$$

A high pass rate indicates that Pretenders were largely successful, suggesting that they possess the relevant interactional expertise and have the kinds of interactions with the target group that are necessary for this to be developed. In contrast, a low pass rate suggests that Pretenders do not possess the relevant interactional expertise and that they are, therefore, either isolated from or unaware of the social world of the target group.

Research design

In exploring the use of the Imitation Game as a tool for comparative, cross-national research we were particularly interested in whether pass rates varied between countries in ways that reflected important cultural characteristics. The hypothesis was that, where the integration of, or interaction between, similar social groups differs between societies then so will the distribution of interactional expertise about those groups and that this difference should be visible in the pass rates of Pretenders. We chose the topic of religion, with our initial hypothesis summarised as follows:

- Where a country has a strong, national religious tradition or identity, the practices and beliefs of that religion should be highly visible such that even those who are not religious will develop interactional expertise in that tradition. This would be made visible as a relatively high pass rate for non-religious players pretending to be religious.
- Where a country has a more secular tradition, religious practices will be hidden from those who do not directly engage in them, meaning that knowledge about them will not be widely shared. This lack of interactional expertise would be made visible as a relatively low pass rate for non-religious players pretending to be religious.

This, in turn, leads to two definitional questions: first, what do we mean by 'religious' and 'non-religious' and, second, what does it mean to say pass rates are 'high' or 'low'?

Fieldwork sites

We collected data in seven European countries – Finland, Hungary, Italy, Netherlands, Norway, Poland and the United Kingdom – in which the dominant religion is Christianity. To categorise these countries as 'religious' or 'secular', and hence derive a ranking against with Imitation Game results could be compared, we used survey data, with countries classified as 'religious' if surveys suggested religion played a significant role in the everyday life of a substantial majority of the population and 'secular' if it did not. Whilst not every survey covered every country, there were some clear patterns: [1]

- **World Values Survey (2005-6, wave 5)**: in response to a question that asked respondents to rate how important religion was in their life, 85% of respondents in Poland said either very or rather important, with only 13% saying religion was either not very or not at all important. The figures for Italy were 75% and 24%, making these were the only two countries in our sample where the proportion saying religion was important was greater than the proportion saying it was not important. The comparable figures for the other countries were 45%

[1] Sources are:

- For Gallup and Eurobarometer poll: https://en.wikipedia.org/wiki/Religion_in_Europe
- For World Values Survey: http://www.worldvaluessurvey.org/WVSOnline.jsp

and 55% for Finland, 40% and 58% for the United Kingdom, 38% and 62% for Hungary, 33% and 67% for Norway, and 30% and 66% for the Netherlands.[2]

- **World Values Survey (2005-6, wave 5)**: in response to a question that asked how often respondents attended a religious service, 75% of respondents from Poland said at least once a month, with only 11% saying they went no more once a year. The figures for Italy were 54% and 20%, making these the only two countries in the sample where more than half the population attends a religious service at least once a month. In all other cases, with the exception of Hungary, the majority of respondents attend a religious service no more than once a year. The comparable figures are 15% and 40% for Hungary, 15% and 62% for Finland, 18% and 65% for the Netherlands, 24% and 66% for the UK, and 11% and 74% for Norway.
- **Gallup (2009)**: 75% of respondents in Poland and 72% of respondents in Italy said religion was important in their daily life compared to 39% in Hungary, 33% in the Netherlands, 28% in Finland, 27% in the United Kingdom and 21% in Norway.
- **Eurobarometer (2012)**: Only 5% of respondents from Poland and 6% of respondents from Italy classified themselves as either atheists or agnostics. In contrast, 22% of Hungarian respondents, 32% of UK respondents, and 49% of Dutch respondents classified themselves in this way (Norway and Finland were not included in the survey)

Based on this data, we classified our fieldwork sites into two groups:

[2] Wave six of WVS is more recent but does not include all the countries on our list.

- Religious: Italy, Poland
- Secular: Finland, Hungary, Netherlands, Norway, United Kingdom

and refined our initial hypothesis to say that pass rates for participants who identify as non-religious and who are pretending to be religious, would be:

(1) Higher in the religious countries than in the secular ones.
(2) Similar within each of the two groups (e.g. within-group differences less then between-group differences)

In making these classifications, we recognise that there will be variations within each country. Nevertheless, some way of calibrating our new method by providing an independent rationale for the expected distribution of interactional expertise was needed. It should also be noted that, because the participants – principally Judges, Interrogators and Non-Pretenders – determine what is relevant, the 'religion' that forms the target expertise is the dominant religion in each country: Catholicism in Italy and Poland, Lutheranism in the Netherlands and Norway, and mixed denominations in all other locations.

Data collection

Fieldwork followed a similar pattern in each location. First, contact was established with a local university and a 'Local Organiser' recruited to assist with the research. Recruitment of participants took place via an online survey, with students from that university asked to self-identify as 'active Christians' or not, according to criteria including attendance at church services and the importance of religion in their

297 everyday life. [3] Next, a number of real-time Imitation Games were played in which
298 students who had identified as 'active Christians' played the role of Interrogator/Judge
299 and Non-Pretender and students who did not self-identify as religious played the role of
300 Pretender (Step 1 in Table 1).[4]

301 Next, each set of questions created during Step 1 was converted into an online
302 survey, and a new, much larger sample of non-Christian Pretenders recruited to provide
303 answers to these questions, with each new Pretender answering one set of questions
304 (Step 2 in Table 1). These new answers were then linked to the questions and Non-
305 Pretender answers created in Step 1 to produce a set of dialogs, one for each of the Step
306 2 Pretenders (called Step 3 but not shown in Table 1 as it is a database operation that
307 requires no participants).

308 These dialogs were then sent to a new sample of Judges (Step 4 in Table 1) who
309 were asked to work out which set of answers came from the Pretender and which from
310 the Non-Pretender. Step 4 Judges were always drawn from students who self-identified
311 as active Christians. As the total number of transcripts created is set by the number of
312 participants at Step 2, and each dialog was judged by two different Judges, each Judge
313 got between 6 and 8 dialogs. Pass rates were calculated as described above, with the
314 sample size given by the number of participants in Step 2.

315

---

[3] The use of students was for practical and logistical reasons. It would, of course, be desirable to repeat the research with more representative samples.

[4] The software that hosts the Game allows participants to play different roles in multiple games simultaneously. This means that equal numbers of each group are needed and not the 2:1 ratio required for a single 'Classic' Imitation Game.

316 [Table 1 about here]
317
318 Before discussing the results, there are some caveats that should be noted:

319 (1) The terminology used to recruit participants varied in response to advice
320 provided by our Local Organisers. For example, Pretenders were recruited as
321 'secular' in some cases and 'non-Christian' in others.
322 (2) Judges and Non-Pretenders may have been recruited as 'active' Christians in
323 some cases and 'practicing' in others, again depending on advice from our Local
324 Organisers
325 (3) The method and protocols evolved over the course of the project, as did the
326 software, as each fieldwork trip identified some problem or bug that needed to
327 be fixed for the next trip.

328 **Results**
329 The results of the Imitation Games are presented as follows:

330 (1) Pass rates for each of the fieldwork locations and visits
331 (2) Discussion of how and to what extent the differences hypothesised before the
332 research are represented and replicated within the data.

333 ***Pass rates by fieldwork locations***
334 There are two independent judgements for each transcript and hence two
335 complete sets of judgements. The pass rate can be calculated for each set and this
336 provides the first element of 'replication'. Assuming there is no statistically significant
337 difference between the two, the final pass rate is taken to be the average of the two pass

rates.[5] Each of these measures is reported in Table 2, which shows that in all cases, bar one (Helsinki, 2013) , there was no statistically significant difference between the two measures of the pass rate.

[Table 2 about here]

Table 2 also shows the ranking of the mean pass rates, which is consistent with expectations based on the survey data. For example, the mean pass rates in Palermo (Italy) and Wroclaw (Poland) are both very high (over 90%). The majority of the rest are much lower, typically below 70%, but there are some outliers at each end of this group. We now explore these results in more detail.

Measures of reliability

Table 2 reports the pass rate calculated using each of the two sets of judgements. Comparing the two provides a measure of the reliability of judgements, though what counts as a 'big' difference between the two is unclear. Given the concern about the use of significance tests, we developed a bootstrap method for estimating the probability of the observed data occurring randomly. This method takes the number of Right, Wrong and Don't Know answers used to calculate each pass rate as 'weights', simulates 10,000 iterations of the Game and uses these to calculate a 95% confidence interval for the difference between the two pass rates.

The outcome is also shown in Table 2. As noted above, apart from data collected in Helsinki in September 2013, there is no statistically significant difference between

[5] For a more detailed exposition of this and all other aspects of the Imitation Game method see (Evans, Collins, & Weinel, 2019)

the two estimates of the pass rate in any location. In this one case, therefore, a judgement is needed. On the one hand, the p-value is greater than 0.05 but, on the other, the results do not look particularly different to the previous year where there was no statistically significant difference. For example, the two pairs of values are relatively similar – 59 and 71 in 2012, 73 and 55 in 2013 – and so is the average – 65 in 2012 and 64 in 2013. Whilst this does suggest that there is something unusual about the Finnish data, we do not think there is a strong reason to exclude the mean pass rate from the analysis and so treat it as a successful replication.

Comparisons between fieldwork sites

We now turn to our principal hypothesis, that pass rates will be higher in those countries classed as religious than in those countries classed as secular. Initial inspection of the Table 2 suggests that the results can be split into three groups rather than the two we originally hypothesised:

(1) **High pass rate** (i.e. above 90 per cent): Palermo (May 2012); Wroclaw (Oct 2011).
(2) **Medium pass rate** (i.e. 50 to 75 per cent): Cardiff (Nov 2011); Helsinki (Nov 2012); Helsinki (Sept 2013); Cardiff (March 2012); Budapest (May 2013); Trondheim (Oct 2012).
(3) **Low pass rate** (i.e. 25 per cent or less): Rotterdam (Dec 2012); Rotterdam (Dec 2013).

There are also two results that sit in-between these categories – Budapest (April 2012) and Trondheim (Nov 2013) – for one fieldwork visit but lie within the medium category for the other visit.

Were statistical evidence needed to support this interpretation, the bootstrap method described above can also be used to make pairwise comparisons between each of the fieldwork sites. The results of this exercise confirm the initial interpretation:

- **High pass rate**: There is no statistically significant difference between Palermo (2012) and Wroclow (2011) but both of these are different to every other case except for the anomalous result from Budapest in April 2012
- **Medium pass rate**: There are no statistically significant differences between Cardiff (Nov 2011), Helsinki (Nov 2012), Helsinki (Sept 2013), Cardiff (March 2012), Budapest, (May 2013) and Trondheim (Oct 2012)
- **Low pass rate**: There is no statistically significant difference between the two results from Rotterdam but these are different to every other result, including the anomalous result from Trondheim in November 2013

Discussion

The aim of the research was to examine the extent to which data collected by a novel method would be (a) consistent with expectations derived from more traditional sources and (b) replicable over time. In what follows, we note the areas where the results of the Imitation Game research show good agreement with the expectations we derived from the existing data before looking in more detail at the three results that were more unexpected: the high pass rate recorded in Hungary in 2012, the low pass rate recorded in Trondheim in 2013 and the very low pass rate recorded in the Netherlands on both visits.

Conformity with survey-based expectations

The hypothesis that informed the research design was that there would be a measurable difference in pass rates in ‘religious countries’ when compared against more

secular countries. Broadly speaking, this was what we found. Pass rates in Palermo and Wroclaw were very high (over 90%) and these were two countries that were highly ranked in all measures of 'religiosity' found in cross-national surveys. In contrast, pass rates in Trondheim, Cardiff, Helsinki and Rotterdam were much lower and this is consistent with their rankings in the same surveys.

Putting these findings in the language of interactional and contributory expertise, we would say that contributory expertise in the nationally dominant religious tradition, in this case Roman Catholicism, is ubiquitous in countries such as Poland and Italy. This means that members of these societies who are not religious or who do not follow the Christian faith are routinely immersed in the language of that religion and that, as a result, acquire a relatively high degree of interactional expertise about it. This is evidenced by their ability to provide plausible answers in an Imitation Game.

In contrast, where religious practices are less mainstream, as in Scandinavian countries, the UK and the Netherlands, the contributory expertise associated with actively practising a faith is less visible – e.g. religion is less likely to be classed as important in everyday life, attendance at services is lower – and this reduces the opportunities for others to develop the related interactional expertise. This is not to say there is no public discourse about the dominant religion but, given the relative paucity of face-to-face social interactions with those who are actively living their faith, we would expect the pass rate to be lower.

Successful replication of results

We did not attempt to replicate results from Palermo or Wroclaw as the pass rate was close to the maximum of 100% and clearly consistent with expectations derived

from the survey data.[6] For other fieldwork sites, if successful replication is defined as a pass rate that appears in the same category on each occasion, we did successfully replicate results in Cardiff, Helsinki and Rotterdam.

Outliers and failures to replicate

In the case of Budapest and Trondheim, we did not replicate results: in each case, we had one result that fell within the 'medium' pass rate category and one that fell outside. In the case of Budapest fieldwork in 2012, the pass rate was higher than expected given the survey data so we initially wondered if this was due to some factor that was specific to Budapest. To check this, we recruited a new sample of Judges from Pécs, another city in Hungary but one that we expected to be more traditional. These Judges then rated the same transcripts as the Budapest Judges and returned a pass rate that was very similar to the one measured in Budapest.

This leaves two possibilities. One is that the Pretenders recruited in Budapest were genuinely knowledgeable about the beliefs and practices of the Christian faith and that this was reflected in authentic answers that Judges in both Budapest and Pécs found hard to distinguish from those provided by active/practicing Christians. In this case, the argument would be that the Imitation Game, by measuring knowledge rather than attitude or practice, has identified a degree of interaction between the two groups that is invisible to other methods.

The other possibility is that the results are an artefact. This would not be entirely surprising given that the research reported here was intended to develop the Imitation Game through using it, that the fieldwork in April 2012 was one of the earliest data

[6] There were also some practical reasons, namely that we also wanted to conduct Imitation Games on sexuality and gender and had a limited number of visits available.

collection visits, that protocols were changing and developing over time, and that, in each case, we were effectively working with a convenience sample.

To investigate this scenario in more detail, we returned to Budapest in 2013 and ran another set of Imitation Games. In this case, the average pass rate came out as 59%, which is well within what we now call the 'medium' category and much closer to what we had initially expected. We also arranged for this second set of transcripts to be judged by a sample of Judges recruited in Pécs. Again, the results were much closer to our initial expectation, with a mean pass rate of 69%. Given the consistency between the 2013 pass rates and the data collected in other fieldwork sites, our view now is that the 2012 data represents an outlier, with sampling and the novelty of the method the most likely explanation for the difference.

Because the anomalous result in Trondheim occurred much latter in the fieldwork cycle, we have not been able return and conduct a third visit. As such, it is possible that either of the results could be the 'correct' one, though, given the other results and our increasing confidence in the Imitation Game method's reliability, we would give more weight to the data that matches the a priori expectations. Again, more research would be needed to determine what might account for the difference.

'New' finding

The other unexpected results were the surprisingly low pass rates recorded in Rotterdam. As with the outlying result from Budapest, we were able to return to Rotterdam to repeat the research. In this case, the initial result was not only replicated but the difference became even clearer, with the pass rate falling from 24% to 17%.

To explain this unexpectedly robust result, we worked with a colleague in the Netherlands to better understand the context in which the data had been generated. Of particular importance, we now believe, is the transformation the Netherlands during the

20th Century from very religious society, with strong Catholic and Protestant communities, to a more secular society. Whilst this process of secularisation may seem to have undermined the traditional pillars of Dutch society – Catholics, Protestants and liberals – it has been argued that the separation continues, particularly for those within the orthodox Protestant tradition. For example, according to Oomen, Guijt and Ploeg (2010) members of the orthodox reformed church have their own newspaper, attend reformed schools, vote for the SGP (an orthodox Calvinist political party), and structure the major part of their social life around these institutions. Indeed, it is possible that while there are fewer Christians in the Netherlands today, the saliency of their belief has been strengthened rather than weakened (Houtman, 2008; Vollaard, 2013)

More importantly, this orthodox part of the protestant population is geographically distinct: most of them live in a region called the 'Bible Belt', which runs close to Rotterdam and may be a significant source of students at the Erasmus University where we conducted our Imitation Games.. Given this, we now believe that the distinct and robust nature of the results are explained by the fact that, in recruiting from a protestant religious community, we have tapped into the increasing social isolation those holding more orthodox religious views. In other words, rather than being an artefact, the low pass rate in Rotterdam reveals something real – and, to us, unexpected – about the lives of those taking part in the research.

Calibration and Replication

Calibration and replication are two different ways of assessing the success of a new method of data collection. We have described each in detail in order to show that judgement is a crucial element of each. In the case of calibration, judgement is needed to determine the suitable proxy measurement against which the new data can be compared. For the Imitation Game research reported here, we made the judgement that

survey data on religious attitudes and practices provided a suitable proxy for religious knowledge. In most cases the ranking and absolute value of the pass rate did seem plausible given the survey data. Where there were outliers, however, determining how to treat the anomalous result the solution required further investigation of the specific case.

The question of replication raised similar concerns. Whilst we have included some quantitative information – e.g. pairwise comparisons of between-country pass rates – it would be incorrect to say that our decision about whether a result had been 'replicated' was, or could be, based purely on this. Instead, the quantitative analysis adds weight to an interpretation of the data that is based on our overall understanding of the fieldwork, something we have tried to convey in the detailed descriptions provided above. The more general point is, therefore, that for any statistical meta-analysis to be conducted, it would first be necessary to consider something like the analysis set out above – a meta-meta-analysis of the design, conduct and context of each study – in order to determine whether or not the data should be included (Collins, 2019, Chapter 9). Whilst this observation does not preclude the use of statistical meta-analysis it does, we hope, introduce a note of caution about the extension of meta-analysis from medical and biological sciences (e.g. Ioannidis, 2005) into social science more generally.

Conclusions

This paper has reported the results of an ambitious replication and calibration study in which a new method was used to collect data across Europe with the aim of (a) producing results that were consistent with existing national survey data and (b) demonstrating its reliability by replicating results from at least some of these fieldwork sites. Comparing results across twelve different fieldwork exercises, we have shown

that that Imitation Game method does work as advertised with more results replicated than not and with Imitation Game data generally matching that collected by larger and much more expensive cross-national surveys.

Where differences between expected and actual results occur, these fall into two groups. First, as with Budapest and Trondheim, it appears likely that the unexpected result is an outlier. More investigation is needed to establish whether methodological factors (e.g. sampling, time of year, phrasing of instructions etc.) contributed to the difference and hence to improving protocols. Second, and more importantly, the results in Rotterdam, suggest that Imitation Game is sensitive to local factors and variations, with the data picking up the importance of the local Protestant community, something which the research team had been unaware of prior to collecting data.

Finally, on the question of replication, we find that focussing purely and narrowly on statistical tests is unlikely to be productive given the complexity and variability of social science fieldwork. Instead, what is needed is a careful analysis of the context and conduct of each study that assesses its own unique strengths and weaknesses. That said, and as we have shown, this does not mean that replication in the social sciences in impossible. Rather the implication is that such conclusions need to be based on a holistic understanding of research data and not statistical testing alone.

References cited

Arminen, I., Segersven, O. E., & Simonen, M. (2018). Active and latent social groups and their interactional expertise. *Acta Sociologica*. https://doi.org/10.1177/0001699318786361

Collins, H. M. (1990). *Artificial experts: Social knowledge and intelligent machines*. MIT Press.

Collins, H. M. (1992). *Changing order: Replication and induction in scientific practice* (2nd edition (1st edition 1985)). University of Chicago Press.

Collins, H. M. (2004). Interactional expertise as a third kind of knowledge. *Phenomenology and the Cognitive Sciences*, *3*(2), 125–143. https://doi.org/10.1023/B:PHEN.0000040824.89221.1a

Collins, H. M. (2010). *Tacit and explicit knowledge*. The University of Chicago Press.

Collins, H. M. (2011). Language and practice. *Social Studies of Science*, *41*(2), 271–300. https://doi.org/10.1177/0306312711399665

Collins, H. M. (2016). An Imitation Game concerning gravitational wave physics. *ArXiv:1607.07373 [Physics]*. http://arxiv.org/abs/1607.07373

Collins, H. M. (2018). *Artifictional intelligence: Against humanity's surrender to computers*. Polity Press.

Collins, H. M. (2019). *Forms of life: The method and meaning of sociology*. The MIT Press.

Collins, H. M., & Evans, R. (2002). The Third Wave of science studies: Studies of expertise and experience. *Social Studies of Science*, *32*(2), 235–296. https://doi.org/10.1177/0306312702032002003

Collins, H. M., & Evans, R. (2007). *Rethinking expertise*. University of Chicago Press. http://dx.doi.org/10.7208/chicago/9780226113623.001.0001

Collins, H. M., & Evans, R. (2014). Quantifying the tacit: The Imitation Game and social fluency. *Sociology*, *48*(1), 3–19. https://doi.org/10.1177/0038038512455735

Collins, H. M., & Evans, R. (2015). Expertise revisited, Part I—Interactional expertise. *Studies in History and Philosophy of Science Part A*, *54*, 113–123. https://doi.org/10.1016/j.shpsa.2015.07.004

Collins, H. M., Evans, R., Hall, M., O'Mahoney, H., & Weinel, M. (2019). Bonfire Night and Burns Night: Using the Imitation Game to Research English and Scottish Identities. In D. Caudill, M. E. Gorman, S. N. Conley, & M. Weinel (Eds.), *The Third Wave in the Sociology of Science: Selected Studies in Expertise and Experience* (pp. 109–131). Palgrave Macmillan. http://dx.doi.org/10.1007/978-3-030-14335-0_7

Collins, H. M., Evans, R., Ribeiro, R., & Hall, M. (2006). Experiments with interactional expertise. *Studies in History and Philosophy of Science Part A, 37*(4), 656–674. https://doi.org/10.1016/j.shpsa.2006.09.005

Collins, H. M., Evans, R., & Weinel, M. (2016). Expertise revisited, Part II: Contributory expertise. *Studies in History and Philosophy of Science Part A, 56*, 103–110. https://doi.org/10.1016/j.shpsa.2015.07.003

Collins, H. M., Evans, R., Weinel, M., Lyttleton-Smith, J., Bartlett, A., & Hall, M. (2017). The Imitation Game and the Nature of Mixed Methods. *Journal of Mixed Methods Research, 11*(4), 510–527. https://doi.org/10.1177/1558689815619824

Dreyfus, S. E. (2004). The five-stage model of adult skill acquisition. *Bulletin of Science, Technology & Society, 24*(3), 177–181. https://doi.org/10.1177/0270467604264992

Evans, R. (2008). The Sociology of Expertise: The Distribution of Social Fluency. *Sociology Compass*, *2*(1), 281–298. https://doi.org/10.1111/j.1751-9020.2007.00062.x

Evans, R., Collins, H. M., & Weinel, M. (2019). Imitation Game. In P. Atkinson, S. Delamont, A. Cernat, J. W. Sakshaug, & R. A. Williams (Eds.), *SAGE Research*

599 *Methods Foundations*. SAGE Publications Ltd.
600 https://doi.org/10.4135/9781526421036838862
601 Evans, R., Collins, H., Weinel, M., Lyttleton-Smith, J., O'Mahoney, H., & Leonard-
602 Clarke, W. (2019). Groups and individuals: Conformity and diversity in the
603 performance of gendered identities. *The British Journal of Sociology*, *70*(4),
604 1561–1581. https://doi.org/10.1111/1468-4446.12507
605 Evans, R., & Crocker, H. (2013). The Imitation Game as a method for exploring
606 knowledge(s) of chronic illness. *Methodological Innovations Online*, *8*(1), 34–
607 52. https://doi.org/10.4256/mio.2013.003
608 Gorard, S. (2016). Damaging Real Lives through Obstinacy: Re-Emphasising Why
609 Significance Testing is Wrong. *Sociological Research Online*, *21*(1), 1–14.
610 https://doi.org/10.5153/sro.3857
611 Gorard, S. (2019). Do we really need confidence intervals in the new statistics?
612 *International Journal of Social Research Methodology*, *22*(3), 281–291.
613 https://doi.org/10.1080/13645579.2018.1525064
614 Houtman, D. (2008). God in Nederland 1996-2006: Enkele godsdienstsociologische
615 routines ter discussie. *Religie & Samenleving*, *3*(1), 17–35.
616 Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS*
617 *Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124
618 Kuhn, T. S. (1961). The Function of Measurement in Modern Physical Science. *Isis*,
619 *52*(2), 161–193. https://doi.org/10.2307/228678
620 Oomen, B. M., Guijt, J., & Ploeg, M. (2010). *CEDAW, the Bible and the State of the*
621 *Netherlands: The Struggle Over Orthodox Women's Political Participation and*
622 *Their Responses* (SSRN Scholarly Paper ID 1625682). Social Science Research
623 Network. https://papers.ssrn.com/abstract=1625682

Segersven, O., Arminen, I., & Simonen, M. (2020). Exploring Groupness—A Mixed Methods Imitation Game Inquiry. *International Journal of Multiple Research Approaches, 12(1)*, 96–109. https://doi.org/10.29034/ijmra.v12n1a3

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, *59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Vollaard, H. J. P. (2013). Re-emerging Christianity in West European Politics: The Case of the Netherlands. *Politics and Religion*, *6*(1), 74–100. https://doi.org/10.1017/S1755048312000776

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*, *23*(3), 253–271. https://doi.org/10.1007/s10728-014-0273-8

Wehrens, R. (2018). Experimentation in the sociology of science: Representational and generative registers in the imitation game. *Studies in History and Philosophy of Science Part A*. https://doi.org/10.1016/j.shpsa.2018.10.003

Winch, P. (1958). *The idea of a social science and its relation to philosophy*. Routledge & Kegan Paul.

Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trans.). Blackwell.

645 **Tables**

Fieldwork site	Date	Type of Christian	No of Step 1 Games	No of Step 1 players	No of dialogs used in Step 2	No of Step 2 participants	No of Step 4 Judges
Wroclaw Poland	Nov-11	Catholic	18	36	18	165	40
Cardiff UK	Nov-11	Mixed	18	36	16	198	36
Cardiff UK	Apr-12	Mixed	18	36	18	175	40
Budapest Hungary	Apr-12	Mixed	18	36	17	180	40
Palermo Italy	May-12	Catholic	27	54	6	189	72
Trondheim Norway	Oct-12	Lutheran	24	48	17	183	40
Helsinki Finland	Nov-12	Mixed	23	46	19	188	40
Rotterdam Netherlands	Dec-12	Lutheran	18	36	10	204	44
Budapest Hungary	May-13	Mixed	20	40	23	241	60
Helsinki Finland	Sep-13	Mixed	20	40	19	188	40
Trondheim Norway	Nov-13	Lutheran	23	46	21	211	55
Rotterdam Netherlands	Dec-13	Lutheran	25	50	23	184	40

646 Table1: Summary of fieldwork

647

Place (and date)	Pass Rate 'First' set of Judges	Pass Rate 'Second' set of Judges	Mean ('First' and 'Second' sets)	P value (diff. 'First' and 'Second' sets)
Palermo, Italy (May 2012)	97%	100%	99%	0.567
Wroclaw, Poland (Oct 2011)	92%	93%	93%	0.904
Budapest, Hungary (April 2012)	88%	80%	84%	0.394
Cardiff, UK (Nov 2011)	74%	74%	74%	0.986
Helsinki, Finland (Nov 2012)	59%	71%	65%	0.104
Helsinki, Finland (Sept 2013)	73%	55%	64%	**0.033**
Cardiff, UK (March 2012)	66%	57%	61%	0.211
Budapest, Hungary (May 2013)	56%	61%	59%	0.416
Trondheim, Norway (Oct 2012)	58%	57%	57%	0.811
Trondheim, Norway (Nov 2013)	40%	36%	38%	0.520
Rotterdam, Netherlands (Dec 2012)	24%	25%	24%	0.809
Rotterdam, Netherlands (Dec 2013)	18%	16%	17%	0.776

Table 2: Pass rates for non-Christian Pretenders in individual fieldwork trips

Acknowledgements

The research was funded by European Research Council Advanced Research Grant (269463 IMGAME) awarded to Collins. We also thank the Local Organisers who organised recruitment, logistics and IT support needed in each fieldwork site and without whom this research would not have been possible.

Author information

Robert Evans is a Professor in the School of Social Sciences at Cardiff University, where he specialises in science and technology studies. In addition to the Imitation Game, his research interests include the nature of expertise and its application to political decision making. He is currently working on a citizen science project examining community responses to a local incinerator.

Harry Collins is Distinguish Research Professor in the School of Social Sciences at Cardiff University and a Fellow of the British Academy. Widely acknowledged as a founder of the sociology of scientific knowledge, his work includes a 40 year study of gravitational wave physics, a sociological analysis of the limits and possibilities of artificial intelligence and, most recently, the role of science within democratic societies

Dr Martin Weinel is a research associate on the EU funded WaterWatt project that aims to improve energy efficiency in European Industries. His previous work includes a study of the controversy over the use of AZT to prevent Mother-to-Child transmission of HIV/AIDS in South Africa and the development of Imitation Game method.

Dr Jennifer Lyttleton-Smith is a Lecturer in Education at Cardiff Metropolitan University. Her research interests include identities and subjectivities, vulnerable and marginalised childhoods, gender and sexuality, well-being and co-production.

Dr Hannah O'Mahony currently works in the third sector. Her research interests include sociology of work, environmental sociology, and Arts and health research

Dr Rik Wehrens is Assistant Professor in the School of Health Policy and Management at the Erasmus University Rotterdam. With a background in science and technology studies, his previous research has examined the use of evidence and patient knowledge

680 in medical practice. His current research interests include the role and use of ‘big data’
681 and artificial intelligence in the context of public health policies and chronic illness.